

 **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 (s16)
Ida Riis Jensen (s161777)
Mikkel Laursen (s16)

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, March 1, 2019

A handwritten signature in black ink, consisting of a large, stylized 'J' followed by a series of loops and a final flourish.

Anton Stockmarr (s16)
Ida Riis Jensen (s161777)
Mikkel Laursen (s16)






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 Data	1
1.1 Original data	1
1.2 Cleaning and preparation	2
2 Exploratory Analysis	3
3 Vejledningsmøder	5
3.1 19. februar	5
3.2 26. februar	7
3.3 5. marts	9
3.4 Spørgsmål	9
4 Noter	11
4.1 Data	11
4.2 Exploratory Analysis	11

Todo list

	3.2 (1) Daily averages of consumption versus temperature differences	7
	3.2 (2) Læse artikler fra Peder	7
	3.3 (3) få styr på lorte parskip-pakken	9
	3.3 (4) Få aksefis af Grønning	9
	4.2 (5) Brug farverne frsa WATTS appen til plots!!!!	11

CHAPTER 1

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.

1.1 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 11 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 11: Attributes from the original house data..

The weather data set consists of the attributes seen in Table 12.

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 12: Attributes from the original weather data..

1.2 Cleaning and preparation

Loader en temporary data ind, som vi modificerer indtil vi putter den ind i vores endelige data. Vi sætter navnet på den første attribute til StartDateTIme. Vi ændrer formatet på de to første attributes til posix, som er `%d - %m - %Y%H : %M : %S`.

Så fjerner vi data fra 2017, fordi vi ikke har noget weather data der. 21 observationer.

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.

enddays og startdays sættes for hvert hus - hvornår starter målingerne og hvornår slutter målinger. Tidspunkterne for aller første og aller sidste måling.

StartDateTIme i weather formateres til rette format, så det passer med house data.

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.

CHAPTER 2

Exploratory Analysis

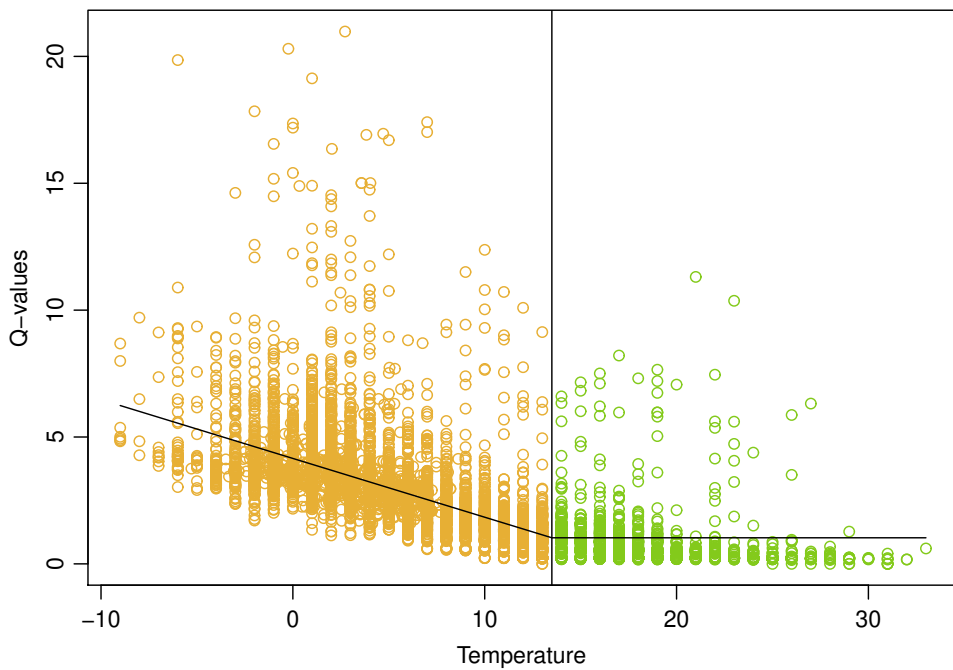


Figure 21: .

CHAPTER 3

Vejledningssmøder

3.1 19. februar

3.1.1 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.

3.1.2 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 varmemandsbeholder. 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 varmemandsforbrug 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.

3.1.3 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.

3.2 26. februar

3.2 (1) Daily averages of consumption versus temperature differences

3.2 (2) Læse artikler fra Peder

3.2.1 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?

3.2.2 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.

3.2.3 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?

3.3 5. marts

3.3 (3) få styr på lorte parskip-pakken

3.3 (4) Få aksefis af Grønning

3.4 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.

3.4.1 Hvad skal vi have lavet?

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

3.5 Noter

CHAPTER 4

Noter

Anton: "Man må jo forvente, at når der kommer spaghetti ind, kommer der også spaghetti ud."

4.1 Data

Alle csv-filerne sættes sammen i en liste(vektor), så hvert element indeholder en tabel over målingerne for én bygning.

Start- og sluttid for målingerne laves om til dato-format.

X-kolonnen fjernes, da den kun består af NAs.

4.2 Exploratory Analysis

4.2 (5) Brug farverne frsa WATTS appen til plots!!!!

pairs plot for hver bygning

Helt generelt kan vi se, at flowet generelt er lavere om sommeren.

Smartest at lave et nyt datasæt for vejrdata, hvor vi flipper sættet, så det seneste datapunkt skal være 29. januar kl. 07:00:00.

Q-plot for at undersøge hvad sammenhængen er mellem energiforbruget (consumption) og udendørstemperaturen. Tilføjer en abline til hvor vi vil sige der slukkes for varmen i huset. Der laves noget least squares på data på hver side af abline, og så fokuserer vi jo selv på det der ligger i den kolde periode. Dog kan vi undersøge hvordan huset performer ved at kigge på data i månederne uden varme.

