Advanced regression methods for energy signatures of buildings

Xie, Jinjing

Award date:
2016

Disclaimer
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
Advanced Regression Methods for Energy Signatures of Buildings

Master Thesis

Author:
Jinjing Xie

Supervisors:
Dr. Alessandro Di Bucchianico
Ing. Bert Elkhuizen
MSc Anna Keune

Eindhoven, November 2016
Abstract

This report is based on the results for the final project for the company ENGIE. For the company there is one main challenge is to recognize equipment and related climatisation processes in a building from energy signatures. My whole project will lead to direction to try to solve it. The project leads the way to figure out the tipping point in the energy signature and a start-up research on external weather information.

The sources of measured data sets are real meter data from clients and simulated data generated by the lab. The data exploration part introduces the structure of both measured data sets and then stores the reshaped real meter data set for further use. Then it follows by manual labeling on the simulated data set as the company did for real meter data sets. In this process, it is concluded that using specific time segments to cluster is not accurate enough. There can be several reasons: one is the winter/summer time changes, another is that the time segments may vary from building to building. Then by checking the winter/summer time changes back in the real meter data, a mistake of the labeling is figured out.

After introducing the methodologies that will be used in the project, the first analysis part is based on the regression model with simulated data set. Regardless of the weather effect here, the projects goes into research on the labels, sources of energy and outdoor temperatures. It shows that it is convenient to do MARS with labels to get separate models in one step. What is more, for the temperatures, taking the derivatives into account makes little difference to the tipping points and using one-hour-lag temperatures will mainly make the tipping points smaller.

The second analysis part is to experiment with an automatic clustering method. The results show there could be fewer number of clusters compared to manual clustering of four clusters using specific time segments. The outputs of clustering are also different with different sources of energy. It is for further research that whether we should do the automatic clustering based on the source of energy separately or together. Although using the combination of sources of energy in MARS suggests the tipping points far away from reality in the regression process, the automatic clustering suggests to consider the combination of sources of energy in the clustering process.

The final part is to start up a research on the weather information using principal component analysis. It shows that based on different data sets, the results could be totally different. Besides the results are dependent on the scale of original variables. Thus more research is needed for external weather information.
# Contents

## Contents

| List of Figures | vii |
| List of Tables  | ix |

## 1 Introduction
1.1 Background of the company ........................................... 1
1.2 Introduction to the project ........................................... 1
1.3 Problem introduction .................................................. 2

## 2 Data Exploration
2.1 Real Data .................................................................. 5
   2.1.1 Background ......................................................... 5
   2.1.2 Structure ........................................................ 6
   2.1.3 Reshaping the data ............................................. 6
   2.1.4 Store the reshaped data ...................................... 7
2.2 Simulated data set .................................................... 7
   2.2.1 Structure ........................................................ 7
   2.2.2 Labeling the data set ......................................... 8

## 3 Statistical Methodology
3.1 Multivariate Adaptive Regression Splines (MARS) ................. 15
   3.1.1 Description of MARS model ................................... 15
   3.1.2 Building process of the MARS model ....................... 15
3.2 Flexible Procedures for Clustering .................................. 17
   3.2.1 Introduction to automatic clustering ....................... 18
   3.2.2 Partitioning Around Medoids ................................ 18
   3.2.3 Description of Partitioning Around Medoids ............. 19
   3.2.4 Graphical output concerning each clustering (silhouette) .................................................. 21
   3.2.5 Alternative automatic clustering methods ............... 22
3.3 Principal Component Analysis (PCA) ................................ 23

## 4 Data Analysis
4.1 MARS analysis on separate source of energy as a whole ........ 25
4.2 Exploration on the MARS analysis with respect to labels ........ 27
   4.2.1 MARS analysis within each cluster ......................... 27
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Gas usages with day: 08:00:00 - 19:00:00 on weekdays</td>
<td>9</td>
</tr>
<tr>
<td>2.2</td>
<td>Electricity usages with day: 08:00:00 - 19:00:00 on weekdays</td>
<td>9</td>
</tr>
<tr>
<td>2.3</td>
<td>Gas usages with different day due to season on weekdays</td>
<td>10</td>
</tr>
<tr>
<td>2.4</td>
<td>Electricity usages with different day due to season on weekdays</td>
<td>11</td>
</tr>
<tr>
<td>2.5</td>
<td>Electricity usages with labels on weekdays</td>
<td>12</td>
</tr>
<tr>
<td>2.6</td>
<td>Gas usages with different “day” due to season on working days</td>
<td>13</td>
</tr>
<tr>
<td>2.7</td>
<td>Electricity usages with different “day” due to season on working days</td>
<td>13</td>
</tr>
<tr>
<td>2.8</td>
<td>Electricity usages with labels on working days</td>
<td>14</td>
</tr>
<tr>
<td>4.1</td>
<td>MARS on gas w.r.t temperatures</td>
<td>26</td>
</tr>
<tr>
<td>4.2</td>
<td>MARS analysis on electricity w.r.t temperatures</td>
<td>26</td>
</tr>
<tr>
<td>4.3</td>
<td>MARS analysis on gas with labels</td>
<td>28</td>
</tr>
<tr>
<td>4.4</td>
<td>MARS on electricity with labels</td>
<td>29</td>
</tr>
<tr>
<td>4.5</td>
<td>Gas usages with fpc gas labels</td>
<td>39</td>
</tr>
<tr>
<td>4.6</td>
<td>Electricity usages with fpc gas labels</td>
<td>39</td>
</tr>
<tr>
<td>4.7</td>
<td>Gas usages with 3 clusters</td>
<td>40</td>
</tr>
<tr>
<td>4.8</td>
<td>Electricity usages with 3 clusters</td>
<td>40</td>
</tr>
<tr>
<td>4.9</td>
<td>Electricity usages with fpc electricity labels</td>
<td>41</td>
</tr>
<tr>
<td>4.10</td>
<td>Gas usages with fpc electricity labels</td>
<td>42</td>
</tr>
<tr>
<td>4.11</td>
<td>Gas usages with 3 clusters with electricity labels</td>
<td>42</td>
</tr>
<tr>
<td>4.12</td>
<td>Gas usages with 4 clusters with electricity labels</td>
<td>42</td>
</tr>
<tr>
<td>4.13</td>
<td>Gas usages with fpc clustering</td>
<td>44</td>
</tr>
<tr>
<td>4.14</td>
<td>Electricity usages with fpc clustering</td>
<td>44</td>
</tr>
<tr>
<td>4.15</td>
<td>Gas usages with 2 clusters with combined labels</td>
<td>45</td>
</tr>
<tr>
<td>4.16</td>
<td>Electricity usages with 2 clusters with combined labels</td>
<td>45</td>
</tr>
<tr>
<td>4.17</td>
<td>Gas usages with 4 clusters with combined labels</td>
<td>45</td>
</tr>
<tr>
<td>4.18</td>
<td>Electricity usages with 4 clusters with combined labels</td>
<td>45</td>
</tr>
<tr>
<td>4.19</td>
<td>Variances for building 1</td>
<td>47</td>
</tr>
<tr>
<td>4.20</td>
<td>Cumulative Proportion of Variance Explained for building 1</td>
<td>47</td>
</tr>
<tr>
<td>4.21</td>
<td>Variances for building 1</td>
<td>48</td>
</tr>
<tr>
<td>4.22</td>
<td>Cumulative Proportion of Variance Explained for building 1</td>
<td>48</td>
</tr>
<tr>
<td>4.23</td>
<td>Variances for building 2</td>
<td>49</td>
</tr>
<tr>
<td>4.24</td>
<td>Cumulative Proportion of Variance Explained for building 2</td>
<td>49</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Structure of simulated data ........................................ 8
2.2 Outliers that should be holidays ................................. 11
2.3 Date time by modifying the summer/winter time ............. 14

3.1 Interpretation of the Silhouette coefficient (SC) .......... 22

4.1 Summary of tipping points for gas ............................. 36
4.2 Summary of tipping points for electricity ................... 37
4.3 Clusters with fpc gas label ................................. 38
4.4 Output of clusters with fpc gas label ................ 38
4.5 Clusters with fpc electricity label .......................... 40
4.6 Output of clusters with fpc electricity label ............ 40
4.7 Clusters with fpc clustering .............................. 43
4.8 Output of clusters with fpc clustering .................. 43
4.9 Available weather information ............................... 46
4.10 Composition of principal components for building 1 .... 47
4.11 Composition of principal components for building 1 .... 48
4.12 Composition of principal components for building 2 .... 50
Chapter 1

Introduction

1.1 Background of the company

The project is carried in the company of ENGIE (formerly Cofely GDF SUEZ) which leads the way in the energy and environmental efficiency. The company designs, implements and operates the most efficient and innovative energy solutions to help their customers (companies, institutions and business stakeholders). ENGIE develops its businesses on electricity, natural gas and energy services.

ENGIE operates on the entire chain of energy services and generation with its extensive knowledge and experience for technology: to make available renewable energy, to reduce the environmental impact (climate change) and to provide and ensure the responsible use of available resources.

1.2 Introduction to the project

The company provides an energy solution which can help the customers achieve their sustainable strategy with innovative renewable energy solutions for buildings and areas. My final project is a part of a smart energy analysis for buildings. It comes to a property owner or tenant to quickly understand the energy performance of the building. This may also provide cost-saving proposals on energy bills and reduction of carbon dioxide. It is suitable for road construction, schools and small-scale production environments.

Based on various input (such as general building attributes, hourly energy usages of the building and hourly weather information on the location of the building), ENGIE can calculate current energy performance of the building and define the operational gap with expected energy performance. The building attributes contain such as gross floor area, opening and closing times, etc. The hourly weather information is provided by the nearest weather station to the building from the Royal Netherlands Meteorological Institute (KNMI). The analysis from ENGIE gives a quick insight into the energy consumption without the need to enter or even see the building. ENGIE’s experience is that through intelligent adjustments to the building systems and user behavior, it is achievable of saving up to 20 to 40% on energy consumption. Another advantage is that the quality of life in the buildings for the users is improved by optimizing comfort. Within a few days the clients can already get an indication of the energy performance of the building and energy potential saving proposal.
1.3 Problem introduction

The main challenges for the company come from two aspects:

1. How to recognize equipment and related climatisation processes in a building from energy signatures (heating, cooling, lighting, domestic hot water, humidification, etc)?

2. How to detect malfunctioning of climatisation equipment in a building from energy values?

In order to answer these questions, we also have some associated questions:

1. To define performance indicators for the main challenges.

2. To define a catalogue of equipment configurations for main challenge 1.

3. To define a catalogue of malfunctioning types for main challenge 2.

4. To decide the size of data we need in order to answer the main challenges.

In my final project, I will mainly work on main challenge 1, i.e. to help recognize equipment and related climatisation processes in a building from energy signatures. For the internship which is the beginning of the final project, I will first look at the available data sets and try to figure out the data structure of the data set and turn the data set into the formal for the following analysis. These are concluded to the first three questions. Then I will go further in my final project, I propose the questions in detail for my whole project:

1. Which measured data sets are available?
   (a) Which variables have already been measured?
   (b) What is the sampling frequency of the measured data sets?
   (c) Which type of time stamps are used in the measured data sets?
   (d) What units are used for the variables in the measured data sets?
   (e) What is the size of the measured data sets?

2. Which simulated data set is available?
   (a) Which variables have already been measured?
   (b) What is the sampling frequency of the simulated data set?
   (c) Which type of time stamps are used in the simulated data set?
   (d) What units are used for the variables in the simulated data set?
   (e) What is the size of the simulated data sets?

3. Which regression method is available?
   (a) What algorithm is appropriate in this project and why?
(b) Which variables should be taken into consideration?
(c) Shall we get the model separately by the labels (clusters) or to involve the labels to get an overall model to improve the efficiency?
(d) What extra variables could be added in the regression model to improve the accuracy?

4. Which automatic clustering methods are available?
   (a) What clustering algorithms for automatic labeling could be used in this project and why?
   (b) What is the difference between the automatic clustering methods and manual clustering?
   (c) What is the advantage to use the automatic clustering methods instead of manual clustering?
   (d) What is the accuracy of the automatic clustering algorithms?

5. Which method is available for the research into weather information?
   (a) Which weather variables can be measured?
   (b) What algorithm can be used in this project?
   (c) How does the algorithm perform in this project?
Chapter 2

Data Exploration

In this chapter, we will do some exploration on the data sets, this work summarizes the results coming from the internship. For this project, we will have two sources of data sets, one is real meter data and another one is the simulated data. We will explain them separately in the following sections. This chapter only introduces the structure of the available data sets and then deal with them a little bit for further analysis.

2.1 Real Data

In this section, we will first introduce the structure of the real meter data in brief. Real meter data is provided by the clients about the energy consumption inside the building. Then ENGIE does some previous preparation for the further analysis, such as changing the time zone(UTC), adjusting the units, reshaping the data sets and clustering based on time segments. Then ENGIE saves the dealt data into the data base for the analysis as lists by the corresponding values of the specific energy. However this structure has split the data set into pieces, thus we need to reshape the data before analyzing.

2.1.1 Background

Before the introduction of the structure of real meter data, it is necessary to explain some background knowledge. Normally the value of the accumulated usages of the specific energy is provided every one hour and there are four kinds of energy here in all: gas, electricity, district heating and district cooling. But for a specific building, it may not use all these four sources of energy, usually two of them are provided and occasionally three are present.

ENGIE first divides the data set into two groups working-days and off-days based on the dates. Working days are normally weekdays and off days are weekends and public holidays. Then for each day, another label is used to divide the data set: day and night. This label is usually given based on a specific time segment to distinguish working hours and off hours for employees inside the building.

Now combined the above two ways of classification, the data set is divided into four clusters named “day”, “night”, “dayWeekend” and “nightWeekend”. They represent the working hours on working-days, off hours on working-days, working hours on off-days and
CHAPTER 2. DATA EXPLORATION

off hours on off-days. Then for every source of energy, the data set is divided by these four clusters.

2.1.2 Structure

In this section, a brief description of the structure of the real meter data set is given. According to the background, ENGIE gives labels to the data using specific time segments and then clusters the data corresponding to the labels. Every time we load the data from the data base and get the dealt data set. The data set is in structure of list by list with all the information, including the energy consumption and client identification as well as properties such as the opening and closing hours for labeling. For the real meter data, normally each data set contains 8670 data points which stands for the energy values recorded in a whole year hourly, however the provided time period for analysis also varies from buildings and clients.

For every source of energy, it contains the information with energy consumption as well as the corresponding time, temperatures and weather information. External weather information is provided by the nearest automatic weather stations from the Royal Netherlands Meteorological Institute (KNMI). We download the weather data from the website\(^1\).

Since ENGIE stores data by the source of energy, it is impossible to do a combined analysis on the different sources of energy. We now reshape the data set first for further use.

2.1.3 Reshaping the data

The company now analyzes the relationship between the usage and outdoor temperature for each specific energy. In case there is some relationship between the different energy sources, we need to reshape the data set to combine all the energy values into one big data set.

First we need to make clear that the labels to cluster the data sets into four parts may not be accurate because they are given manually according to common sense. This part will also be discussed with the simulated data in the following literature. However here when we reshape the data set, we still use the given labels first. ENGIE saves the energy values by the source of energy, thus for a specific source of energy, the data is divided by the labels. We then need to combine all the information according to one specific label. As introduced in the background, there are four sources of energy that the company may use, however normally there are only two which are used. Then we need to consider the circumstances by the number of the sources of energy that the company use. We create a function to combine the information as following.

Step 1: To pick a specific kind of energy and check whether the values are available, if they are **NULL** in \(R\), set such energy value as **NA**. If not, then combine the “date”, “temperatures”, energy usages and weather information within each label into data frame with setting the corresponding label. Then with the \(R\) command `rbind.data.frame` we can combine all the information into one data frame within the specific source of energy.

\(^1\)http://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi

6 Advanced Regression Methods for Energy Signatures of Buildings
Step 2: We merge all the four data frames into one data frame, the concrete way to deal is based on the number of available energy values. If there only one or two data frame is unavailable, then we could merge the data directly and rename the columns. If there is only one data frame available, then this must be electricity, and we directly add three \texttt{NA} columns with the names of other three sources of energy. If all the energy values are not available, then the whole data set is nothing, which is set to be \texttt{NULL} in R.

Step 3: We add the client identification as an extra column.

Now the data set is transferred to data frame in R, it is not the format of list by list any longer. This reshaping makes all the needed information in one list which is compact.

2.1.4 Store the reshaped data

In order for future use, we store the completed reshaped data set in \texttt{xlsx} files with the corresponding client id as the name of the file.

2.2 Simulated data set

In this section, we will give the introduction to simulated data which gives an example of a calculation with a dynamic building simulation program. This program calculates the heat balance of the building every hour. In this simulation “everything is working perfect” so it is the ideal reference signature. Analyzing simulated data could help us better understand the working systems in the building and by comparing the real meter data and energy signatures within the simulated data we may detect malfunctioning.

First as what we have done in the real meter data part, we introduce the structure of the simulated data set. We will introduce what variables are available in this simulated data set and tell something about the sampling frequency and sample size. Then we also introduce the unit to every variable.

2.2.1 Structure

In this section, the description of the structure of the simulated data is given. In this dynamic building, there are several systems working together, each of them makes use of energy from sources of gas or electricity. Note that there are two different heating systems working together in the building, and the main heating system uses gas. Note that gas is also only used for this heating system. The electricity is used for several systems which is recorded separately, the other heating system also uses a very small amount of energy.

Different from the real meter data, the simulated data is stored directly in \texttt{xlsx} file with all the information in the building, noticing that this simulation ignores the outside weather effect.

We would like to analyze the heat balance on two different system levels. The so called need of the building are with columns “Zone Sensible Heating” and “Zone Sensible Cooling” and the energy consumption of the building are with columns “Total Electricity” and “Total gas”. Then after loading the original data set, we only leave needed columns...
for analysis. We will show these variables in the Table 2.1. The whole data set contains 8760 data points, which is exactly the hourly data for a whole year.

Table 2.1: Structure of simulated data

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Mode</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date/Time</td>
<td>8760</td>
<td>numeric</td>
<td>Date and hourly time</td>
</tr>
<tr>
<td>Outside Dry-Bulb Temperature</td>
<td>8760</td>
<td>numeric</td>
<td>Average temperatures within one hour - °C</td>
</tr>
<tr>
<td>Total Electricity</td>
<td>8760</td>
<td>numeric</td>
<td>Accumulated electricity usages within one hour - W/m²</td>
</tr>
<tr>
<td>Total Gas</td>
<td>8760</td>
<td>numeric</td>
<td>Accumulated gas usages within one hour - W/m²</td>
</tr>
<tr>
<td>Zone Sensible Heating</td>
<td>8760</td>
<td>numeric</td>
<td>Need of heating in the building - W/m²</td>
</tr>
<tr>
<td>Zone Sensible Cooling</td>
<td>8760</td>
<td>numeric</td>
<td>Need of cooling in the building - W/m²</td>
</tr>
</tbody>
</table>

2.2.2 Labeling the data set

Noticing that the simulated data is not labeled yet, in order to analyze on the data set according to labels, we first label the data. This is a manual process using the same method the company does to the real meter data. ENGIE uses specific time segments to cluster working and off hours with label “day” and “night” and then uses date to distinguish working days and off days. The holidays in the Netherlands can be checked on the web site \(^2\). Although the simulated data is based on the date of year 2002, and we use the calendar date to distinguish the weekdays and weekends; however, the public holidays and winter/summer changing time may not be accurate according to year 2002. This is because that the dates of holiday as well as winter/summer time changes vary from year to year. Now we just suppose that we do not the exact date for holidays and winter/summer changing time, but they should be considered as existence in this data set. In fact during the following labeling process, we will figure out the corresponding dates. This whole process also indicates that in real meter data, only using specific time segments to label the data set may not be accurate.

Step 1: Labeling with specific time segment

For most buildings, ENGIE chooses the time segment as 08:00:00 – 19:00:00 as working hours to label with “day”, the rest are with label “night”. Now we first apply this time segment with the corresponding dates to get a first version of labeling the simulated data. After labeling the data set, it is necessary to have a look at the plot of data points with their corresponding labels, it may give us a first impression on the whole data set shown in Figure 2.1 and Figure 2.2.

\(^2\)https://www.timeanddate.com/holidays/netherlands/
CHAPTER 2. DATA EXPLORATION

Due to the property of simulated data, we know that the cooling systems could only work on working hour which is also the cause of the big jump shown in Figure 2.2. Then for all the data points with label “day” should only lie in the upper part in the Figure 2.2. We will also use this property to check whether the clustering is done in the following analysis.
By looking at the data points with label “day”, it is clear there are some data points in the lower part. By looking at these points we may find that data points on 08:00:00, 18:00:00 and 19:00:00 may cause the problem. First all data points with time 19:00:00 should all lie in the “night”. There is also a seasonal effect that 08:00:00 should not appear in the “winter” and 18:00:00 should not appear in the “summer”. Then we try to distinguish the season and label the data separately.

Step 2: Labeling with different time segment by seasons

By looking back at the data set, we find that the changes of winter/summer to distinguish the “day” and “night” happen on March 31 and October 31. For the period of March 31 to October 31, the working hours on weekdays should be 08:00:00 − 17:00:00 while for the rest dates the working hours should be 09:00:00 − 18:00:00. Then we recognize the date between March 31 to October 31 as “summer” and the rest date as “winter”.

Looking at Figure 2.3 and Figure 2.4, we could get an overview plot of the whole data set with the color standing for each cluster.

![Figure 2.3: Gas usages with different day due to season on weekdays](image)

However some points may be hidden to see from the picture. Thus we plot all the points of electricity separately by its corresponding label in Figure 2.5.

By looking at the “day” after this way of labeling in Figure 2.5, we could see there are only a few points left in the lower part. These 80 points are shown in the following Table 2.2. Then we find that these 80 points are exactly 8 days and actually they should be holidays and be labeled with weekend.
CHAPTER 2. DATA EXPLORATION

Figure 2.4: Electricity usages with different day due to season on weekdays

Table 2.2: Outliers that should be holidays

<table>
<thead>
<tr>
<th>date</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2002-01-01</td>
<td>09:00:00-18:00:00</td>
</tr>
<tr>
<td>2 2002-04-02</td>
<td>08:00:00-17:00:00</td>
</tr>
<tr>
<td>3 2002-04-10</td>
<td>08:00:00-17:00:00</td>
</tr>
<tr>
<td>4 2002-05-03</td>
<td>08:00:00-17:00:00</td>
</tr>
<tr>
<td>5 2002-05-31</td>
<td>08:00:00-17:00:00</td>
</tr>
<tr>
<td>6 2002-06-03</td>
<td>08:00:00-17:00:00</td>
</tr>
<tr>
<td>7 2002-12-25</td>
<td>09:00:00-18:00:00</td>
</tr>
<tr>
<td>8 2002-12-26</td>
<td>09:00:00-18:00:00</td>
</tr>
</tbody>
</table>

Step 3: Removing the outliers by holidays

Now we relabel the above 8 days as weekend, and according to the seasons and time segments, we now modify the labels for the final version. We could get the result of labeling and show the data points with labels in the following Figure 2.6 and Figure 2.7.

After checking all the clusters, especially the usages in electricity we could say now the label is more accurate than using a specific time segment, the results are shown in Figure 2.8. Then we have the doubt whether there is any proposed reason to make the time segments so different by seasons. Related to the reality we may doubt maybe this is because of the winter/summer time changes. Next we will have a look at the winter/summer time changes.
Step 4: Changing the summer/winter time

From the previous analysis, we find that the summer/winter time change on March 31 and October 31. Then we need change the summer/winter time as it should be in the reality. The original time and modified time are shown in the Table 2.3: this could be done by shifting the time segment from row 2139 to row 7273 by adding 1 in hour.

We now complete labeling the simulated data set before analyzing. Next chapter we will follow with the methodology that will be used in this project.
Figure 2.6: Gas usages with different “day” due to season on working days

Figure 2.7: Electricity usages with different “day” due to season on working days
Figure 2.8: Electricity usages with labels on working days

Table 2.3: Date time by modifying the summer/winter time

<table>
<thead>
<tr>
<th></th>
<th>date</th>
<th>original time</th>
<th>modified time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2136</td>
<td>2002-03-31</td>
<td>00:00:00</td>
<td>00:00:00</td>
</tr>
<tr>
<td>2137</td>
<td>2002-03-31</td>
<td>01:00:00</td>
<td>01:00:00</td>
</tr>
<tr>
<td>2138</td>
<td>2002-03-31</td>
<td>03:00:00</td>
<td>03:00:00</td>
</tr>
<tr>
<td>2139</td>
<td>2002-03-31</td>
<td>03:00:00</td>
<td>04:00:00</td>
</tr>
<tr>
<td>2140</td>
<td>2002-03-31</td>
<td>04:00:00</td>
<td>05:00:00</td>
</tr>
<tr>
<td>7271</td>
<td>2002-10-30</td>
<td>23:00:00</td>
<td>00:00:00</td>
</tr>
<tr>
<td>7272</td>
<td>2002-10-31</td>
<td>00:00:00</td>
<td>01:00:00</td>
</tr>
<tr>
<td>7273</td>
<td>2002-10-31</td>
<td>01:00:00</td>
<td>02:00:00</td>
</tr>
<tr>
<td>7274</td>
<td>2002-10-31</td>
<td>02:00:00</td>
<td>02:00:00</td>
</tr>
<tr>
<td>7275</td>
<td>2002-10-31</td>
<td>03:00:00</td>
<td>03:00:00</td>
</tr>
</tbody>
</table>
Chapter 3

Statistical Methodology

In this chapter, we will introduce a regression algorithm called Multivariate Adaptive Regression Splines (MARS), a clustering method from flexible procedures for clustering (FPC) and Principal Component Analysis (PCA). These are the main algorithms we may use in the analysis.

3.1 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) is introduced by Friedman (1991), this is the key algorithm in this case, since we would like to find the breaking points. For a brief introduction, one could search Friedman et al. (2009, pp. 321–329). For the usages in R, one can refer to Milborrow (2016a) and Milborrow (2016b).

3.1.1 Description of MARS model

MARS builds the model in the following form

\[ f(x) = \sum_{i=1}^{k} c_i B_i(x). \]  

The model is a weighted sum of basis functions \( B_i(x) \). Each \( c_i \) is a constant coefficient. For each basis function \( B_i(x) \), it takes one of the following three forms:

- Constant 1. Then this is the intercept.
- A hinge function. A hinge function has the form \( \max(0, x - \text{const}) \) or \( \max(0, \text{const} - x) \). MARS automatically selects variables and values of those variables for knots of the hinge functions. Knots are the constant in the hinge functions.
- A product of two or more hinge functions. These basis functions can model interaction between two or more variables.

3.1.2 Building process of the MARS model

MARS builds a model in two steps: the forward and the backward stepping. This is the similar approach used in stepwise regression, the details can be referred to (Jennrich, 1960, Chapter 4) and (Draper and Smith, 1998, Chapter 15).
CHAPTER 3. STATISTICAL METHODOLOGY

The forward stepping

MARS starts a model which consists of just the intercept term, this is determined by the mean of the available data points.

MARS then repeatedly adds basis function in pairs to the model. At each step it finds the pair of basis functions by certain statistical criteria. In the stepwise regression, we have the similar approach to add the predictors, here are some examples for stepwise regression from (Jenrich, 1960, pp 58–59):

1. Add the predictor whose partial correlation with $y$ is maximum, given the previously selected predictors.

2. Add the predictor that increases the multiple correlation between $y$ and the selected predictors to the maximum.

3. Add the predictor that decreases the residual sum of squares (RSS) to the maximum.

4. Add the predictor whose F-statistic has the largest value.

These four criteria are mathematically equivalent.

Here in MARS to determine the basis functions, it chooses the third criterion that reduces the residual sum of squares to the maximum. The two basis functions in the pair are identical except that a different side of a mirrored hinge function is used for each function. At each step, a new basis function is an existing term from the current model multiplies by a new hinge function. Since hinge functions is defined by a variable and a knot, to add a new basis function, MARS must search for all the following combinations:

- existing terms
- all variables to select one for the new basis function
- all values of each variable for the knot of the new hinge function

To calculate the coefficient of each term, MARS applies a linear regression over the terms. This process of adding terms will continue until the change in residual error is small enough or until the maximum number of terms is reached. The number of terms is usually specified by the user before building the model.

The backward stepping

The forward stepping usually builds an overfitted model. To prevent this, the backward stepping is used to remove the terms one by one. In stepwise regression, the backward stepping starts with the complete regression with all predictors and then removes one predictor at each time. To determine which predictor should be removed, certain criteria using F-statistic is offered by (Jenrich, 1960, pp 60) and (Seber and Lee, 2003, Chapter 12). In MARS it also deletes the least effective term until it finds the best model, however this is measured using the generalized cross validation as criterion. Generalized cross validation was first introduced by Craven and Wahba (1978) and then extended by Friedman (1991).
The backward stepping has an advantage over the forward stepping: at any step it can choose any term to delete since it starts from the complete regression model; however for the forward stepping it can only see the next pair of terms at each step.

Note that the forward stepping adds terms in pair while the backward stepping typically discards one side of the pair, so the terms are normally not seen in pairs in the final model.

**Generalized cross validation**

The backward stepping use generalized cross validation (GCV) to choose the best subset of model: it is better with lower values of generalized cross validation.

The reason why MARS chooses generalized cross validation instead of residual sum of squares is easy to explain. This is because we would use new data to estimate how the model performs; and such new data is not available when the model is building. Then the raw residual sum of squares is not adequate to compare the models since it will always increase when MARS terms drop. Then if we use RSS to compare models, the backward stepping will always choose the largest model.

The formula for the GCV is

\[ GCV = \frac{\text{RSS}}{(n * (1 - m))^2}, \]  

where \( m \) stands for the effective number of parameters. The effective number of parameters in MARS is defined as

\[ m = k + p * (k - 1)/2, \]

where \( k \) is the number of MARS terms and \( p \) is the penalty term which is 2 or 3 normally.

Note that \((k - 1)/2\) is the number of hinge function knots, so the formula penalizes the addition of knots. Thus the GCV formula increases the training RSS to make the model more flexible.

### 3.2 Flexible Procedures for Clustering

In this section, we will introduce an automatic clustering method named Partitioning Around Medoids (PAM). This is because we would like to do the clusterwise regression in the project. Clusterwise linear clustering is a combination of finding clusters and linear regression. It is first introduced by Späth (1977) and Späth (1979). The concrete steps involved in this algorithm are introduced by Qian and Wu (2011), DeSarbo and Cron (1988) and Hennig (1999). Hennig also researched into clusters and outliers for linear regression in Hennig (2002) and Hennig (2003).

FPC is an R package for flexible procedures for clustering for which the manual is accessed by Hennig (2015). Inside the package there is also the main algorithm we use in the analysis for automatic clustering. We use the R command `pamk()` from this package, this function is as the same as R command `pam()` from package `cluster`. For the following literature, we will introduce the automatic clustering algorithm related to function `pam()` referring to (Kaufman and Rousseeuw, 1990, Chapter 2).
3.2.1 Introduction to automatic clustering

Clustering is used in both statistics and machine learning technique for dividing the data into groups based on their similarities. The majority of all the clustering algorithms are from two main kinds: hierarchical and partitioning clustering.

Hierarchical clustering divides the data set by constructing the hierarchy. There are two kinds of hierarchical technologies: the agglomerative and the divisive. They construct the hierarchy in opposite directions, which will also suggest a different result. Partitioning clustering constructs exact number of partitions, and this number should be specified by the user. The representative method of partitioning clustering is known as k-means algorithm.

The choice of the clustering method depends on the type of the available data set as well as the purpose we do clustering. Sometimes several algorithms can be applicable, the recommendation is to run several algorithms and compare the outputs. The interpretation of the algorithms must then be based on the meaning of the original data set, however this is also the hardest part in the clustering process.

I have tried both typical hierarchy and partitioning clustering methods on the simulated data set. As for the hierarchy clustering on the energy usages, no matter we do the clustering based on the separate energy or the combination, it suggests to use the level of energy usages to cluster the data set. If we plot the data set with colors representing their labels, we can find that the data set is simply divided by horizontal lines. This clustering result is far from the expectation towards the energy behaviors. Then we may turn to the partitioning method which suggests a better result of clustering.

3.2.2 Partitioning Around Medoids

For the partitioning clustering, the most basic algorithm is k-means clustering, however R has provided an alternative algorithm called k-medoid clustering. It is very similar to k-means, and the major difference between them is that: while a cluster is represented with its mean value in the k-means algorithm, it is represented with the object closest to the center of the cluster in the k-medoid clustering; these representing objects of the clusters are called medoids referring to Kaufman and Rousseeuw (1987). Thus the k-medoid clustering is more robust than k-means in presence of outliers. The example that clustering with both the two methods on the same data set iris is given in Zhao (2012, pp 49–53).

Both functions pam() and pamk() represent the algorithm of PAM (Partitioning Around Medoids) which is classic for k-medoid clustering. All the information concerning about this algorithm is then summarized in Kaufman and Rousseeuw (1990, Chapter 2). Function pam() in package cluster referring to Maechler et al. (2016) is respectively implementation of PAM in R. Note that for function pam() the user has to specify k, the number of clusters to find. The advice to choose the best k is to run the algorithm several times with different values of k and then select the k with the “best” clustering, which will be explained in the output section. As an enhanced version of pam(), function pamk() in package fpc does not require the user to choose k. Instead, it calls the function pam() to perform with a given range of k and then the value of k is estimated by the maximum average silhouette width, which in detailed introduced in Rousseeuw (1987) and Kaufman and Rousseeuw (1990, pp 83–88).
3.2.3 Description of Partitioning Around Medoids

In this section, we will describe the algorithm, first we will give a brief introduction about the main idea of this algorithm and then give the mathematical definitions towards the algorithm. PAM is used to cluster the data set by the dissimilarities between the data points. When partitioning a given data set into $k$ clusters, the data points within the same cluster show a high degree of similarity while the data points belonging to different clusters should be dissimilar as much as possible. For every cluster we have a representative object which is called medoid of the clusters. PAM is used for searching for the $k$ medoids and then assigning each object of the data set to the nearest medoid.

There are two principles for the partitioning clustering:

- Each cluster must contain at least one object.
- Each object must belong to exactly one cluster.

They show that any two different clusters could not have any objects in common; and the $k$ clusters together add up to exactly the original data set.

First we will introduce the structure of the PAM algorithm referring to Struyf et al. (1997). We use $d(i, j)$ to measure the dissimilarity between the object $i$ and $j$. Partitioning around medoids minimizes the sum of the dissimilarities of all the objects to their nearest medoid. To have a mathematical expression is that we try to find a subset of $k$ objects $\{m_1, m_2, \cdots, m_k\}$ of $n$ objects $\{1, 2, \cdots, n\}$ which minimizes the objective function

$$\sum_{i=1}^{n} \min_{j=1, \cdots, k} d(i, m_j).$$

(3.3)

Then each object is assigned to the cluster corresponding to the nearest medoid. That is object $i$ is in cluster $p$ if

$$d(i, m_p) = \min_{j=1, \cdots, k} d(i, m_j)$$

or

$$d(i, m_p) \leq d(i, m_q), \quad \forall q = 1, 2, \cdots, k.$$

An alternative description of the model for this cluster analysis is introduced first by Vinod (1969) and also discussed by Rao (1971) and Mulvey and Crowder (1979). This way of description gets rid of the notations of medoids by introducing two indicator variables $y_i$ and $z_{ij}$:

1. If object $i$ is selected as a medoid, then $y_i$ equals to 1, otherwise it should be 0.
2. If object $i$ is a medoid of a cluster, and object $j$ is assigned to this cluster, then $z_{ij}$ equals to 1, otherwise it should be 0.

Now the $k$-medoid model can be rewritten as

$$\minimize \sum_{i=1}^{n} \sum_{j=1}^{n} d(i, j)z_{ij}$$

(3.4)
subject to

\[ \sum_{i=1}^{n} z_{ij} = 1, \quad j = 1, 2, \ldots, n \]  \hspace{1cm} (3.5) \\

\[ z_{ij} \leq y_i, \quad i, j = 1, 2, \ldots, n \]  \hspace{1cm} (3.6) \\

\[ \sum_{i=1}^{n} y_i = k, \]  \hspace{1cm} (3.7) \\

\[ y_i, z_{ij} \in \{0, 1\}, \quad i, j = 1, 2, \ldots, n \]  \hspace{1cm} (3.8)

These constraints imply the two main principles of this algorithm. Formula (3.5) shows that for a specific object \( j \), it should belong to one cluster of medoid \( i \) and only belongs to this one cluster. Formula (3.6) indicates that only if object \( i \) is a medoid, then object \( j \) could be assigned to its cluster. This is connected with the fact that each cluster must contain at least one object. Formula (3.7) shows that there is exactly \( k \) objects to be chosen as medoids. We can also rewrite the dissimilarity between the object \( j \) and the medoid \( i \) of which cluster that \( j \) belongs to as Formula (3.9)

\[ \sum_{i=1}^{n} d(i,j)z_{ij} \]  \hspace{1cm} (3.9)

Then the algorithm consists of two steps referring to Kaufman and Rousseeuw (1990, pp 102–104) and Struyf et al. (1997): the first step is called BUILD and the following one is called SWAP.

In step BUILD, an initial clustering is obtained by successive selection of medoids until all \( k \) medoids are found. The following steps are carried out:

1. The first medoid is the one of which the sum of the dissimilarities to all other points is as small as possible. This one can be regarded as the center of the data set.

\[ m_1 = \arg \min_j \sum_{i=1}^{n} \min_{j=1,\ldots,n} d(i,j) \]

2. The next medoid is selected by decreasing the objective function (3.3) as much as possible.

\[ m_2 = \arg \min_j \sum_{i=1}^{n} \min_{j\neq m_1} \{d(i,m_1), d(i,j)\} \]

3. We repeat step 2 until all \( k \) medoids are selected.

In step SWAP, the algorithm attempts to improve the medoids. The following steps are carried out:

1. Consider the pair of objects with

\[ i \in \{m_1, \ldots, m_k\}, \quad j \not\in \{m_1, \ldots, m_k\} \]

and make the \( i \leftrightarrow j \) swap (if any) which decreases the objective function 3.3 as much as possible.
2. We repeat the above step until convergence.

Since the objective function (3.3) only depends on the dissimilarities between the object, the order of the objects will make no difference on the result.

### 3.2.4 Graphical output concerning each clustering (silhouette)

When the clustering is done by the algorithm, a graphical representation of each clustering is provided to display the silhouette which is introduced in detail in Rousseeuw (1987) and Kaufman and Rousseeuw (1990, pp 83–88). Each cluster is represented by one silhouette, showing which objects lie well within the cluster and which objects merely hold an intermediate position. The entire clustering is displayed by plotting all silhouettes in one figure.

Let us first explain the silhouette plot. For each object \( i \) it belongs to a cluster \( A \), and we use \( |A| \) to denote the number of objects belonging to cluster \( A \). Then we compute the average dissimilarities of \( i \) to all other objects of \( A \):

\[
a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j) \quad (3.10)
\]

The value of \( a(i) \) measures how dissimilar the object \( i \) is to the medoid of the cluster it belongs to, thus a small value means that object \( i \) is close to the medoid and should lie in this cluster.

Now consider any cluster \( C \) different from \( A \) and define the average dissimilarity of \( i \) to all objects of \( C \).

\[
d(i, C) = \frac{1}{|C|} \sum_{j \in C} d(i, j) \quad (3.11)
\]

After computing \( d(i, C) \) for all clusters \( C \neq A \) we take the smallest of those:

\[
b(i) := \min_{C \neq A} d(i, C) \quad (3.12)
\]

The cluster \( B \) which attains this minimum, i.e. \( d(i, B) = b(i) \) is called the neighbor of object \( i \). This is like the second best cluster for object \( i \), this means that regardless of cluster \( A \), cluster \( B \) is closest to object \( i \). Therefore a large \( b(i) \) implies that object \( i \) is far from its neighboring cluster.

The silhouette value \( s(i) \) is then defined as

\[
s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3.13)
\]

We could rewrite this formula as

\[
s(i) = \begin{cases} 
1 - \frac{a(i)}{b(i)} & a(i) < b(i) \\
0 & a(i) = b(i) \\
\frac{b(i)}{a(i)} - 1 & a(i) > b(i)
\end{cases} \quad (3.14)
\]

When cluster \( A \) contains only one object, \( a(i) \) is not defined and we then simply set \( s(i) = 0 \). From Formula (3.13) it is not hard to see that for each object

\[-1 \leq s(i) \leq 1\]

We will introduce the three extreme situations to help us understand \( s(i) \).
1. \( s(i) \approx 1 \): \( a(i) \) is much smaller than \( b(i) \). Then the object is well classified in \( A \).

2. \( s(i) \approx 0 \): \( a(i) \) and \( b(i) \) are approximately equal and hence it is not clear whether object \( i \) should be assigned to which cluster. Then the object lies intermediate between cluster \( A \) and \( B \).

3. \( s(i) \approx -1 \): \( a(i) \) is much larger than \( b(i) \). Then the object lies much closer to cluster \( B \) than to \( A \). Thus it is more accurate that object \( i \) should lie in cluster \( B \), and we can say that object \( i \) is misclassified.

The silhouette of cluster \( A \) is a plot of all its \( s(i) \), ranking in decreasing order. The silhouette plot shows the silhouette of all clusters below each other. This could also help the measure how well a cluster is classified: a wide silhouette is better than a narrow one since the silhouette width is often from 0 to 1. Besides, there are two more summaries for the results of clustering. One is called the average silhouette width of the cluster, this is the average of \( s(i) \) for all objects within one cluster. Normally this shows beside each cluster. Another one is called average silhouette width for the entire data set which is denoted as \( \bar{s}(i) \), this is the average of \( s(i) \) for all objects. This measure how appropriately the data set has been clustered. Normally this shows in the bottom of the silhouette plot.

In general, we propose to run \texttt{pam()} several times, each time with a different \( k \), and to compare the resulting silhouette plots. Then we can then select that value of \( k \) with the maximum average silhouette width over all \( k \), which we call the silhouette coefficient:

\[
SC = \max_k \bar{s}(k), \quad k = 2, 3, \cdots, n - 1
\]  \hspace{1cm} (3.15)

For the function \texttt{pamk()}, you use set a range for the \( k \), and it automatically returns the result of \( k \) with the maximum average silhouette width.

\( SC \) is useful to measure the number of the clustering given by the algorithm. The silhouette coefficient is dimensionless which is at most equal to 1. Experience with the \( SC \) could give us a subjective interpretation shown in Table 3.1 referring to Kaufman and Rousseeuw (1990, pp 88). The silhouette and \( SC \) are not restricted to a particular partitioning algorithm. If \( SC \) is close to 1, then the data set is well clustered and a low

\begin{center}
\begin{tabular}{|c|c|}
\hline
\textbf{SC} & \textbf{Interpretation} \\
\hline
0.71 - 1.00 & A strong clustering structure \\
0.51 - 0.70 & A reasonable clustering structure \\
0.26 - 0.50 & A weak clustering structure and could be artificial; please try additional methods on the data set \\
\leq 0.25 & No substantial structure \\
\hline
\end{tabular}
\end{center}

\textbf{Table 3.1: Interpretation of the Silhouette coefficient (SC)}

SC may indicate that we could try another clustering method to get a better result.

\subsection*{3.2.5 Alternative automatic clustering methods}

Partitioning Around Medoids is not the only algorithm we can use; there are several other typical partitioning clustering methods that could be considered. We can have a try with these algorithms, but so far I will just have a quick explanation towards the algorithms.
Clustering Large Applicants

For large data sets, PAM will be not practical for the time and memory requirements. This method is also based on k-medoid and it is adapted to data set with large amount (so called applicants). In R package cluster, the related function is clara() (abbreviated from Clustering LARge Applicants). In brief, this is achieved by considering subset of data sets with fixed sample size. Each subset is partitioned into k clusters using the algorithm from pam(). For more information one could refer to Kaufman and Rousseeuw (1990, Chapter 3). This algorithm is more efficient when we are dealing with big data, it saves both time and memory.

Fuzzy Analysis

In partitioning clustering methods, each object of the data set should be assigned to one and only one cluster. However sometimes there are some objects lie in the center of two or even more clusters, then the choice of which cluster it should belong to will be kind of arbitrary. Then a fuzzy clustering is applied here. In R package cluster, the related function is fanny(). In a fuzzy clustering, each observation is “spread out” over the various clusters. The degree of belonging is denoted by the membership coefficient which ranges from 0 to 1. The sum of the membership coefficients for each object equals to 1. For more information one could refer to Kaufman and Rousseeuw (1990, Chapter 4).

Actually in our project, some data points with same value may come from the different energy behaviors, thus there could be an overlapping for some clusters. Then this fuzzy analysis may be useful for the overlapping part. For each data point, the algorithm will not assign it to one cluster with “yes” or “no”, but with the possibility to which clusters it could belong to.

3.3 Principal Component Analysis (PCA)

Principal component analysis (PCA) is to use an orthogonal transformation on a number of correlated variables to a smaller (hopefully) set of uncorrelated variables, which are called principal components. These principal components are a normalized linear combination of the original variables. The transformation is defined in such a way that the first principal component has the largest possible variance, which means that the first principal component accounts for the largest of the variability in the data set. The succeeding components has the smaller variance and they are orthogonal to the preceding components. This technique relies on the correlation or covariance matrix, thus the data set should be numeric. Also note that PCA is sensitive to the relative scaling of the original variables, thus principal components are supplied with normalized version of original predictors. PCA was introduced by Pearson (1901) and developed by Hotelling (1933). A detailed summary for principal component analysis and its applications are in Jolliffe (2002). For the different fileds of application, PCA sometimes has different names, such as discrete Kosambi-KarhunenLove transform (KLT) in signal processing or proper orthogonal decomposition (POD) in mechanical engineering. For a quick study of the technique one could refer to Jackson (2005).

For the following literature, we will give a brief introduction to the algorithm referring
CHAPTER 3. STATISTICAL METHODOLOGY

to Bro and Smilde (2014) and Hyv¨rinen (1970, pp. 82–104). They use eigenvalue decomposition (EVD) to calculate the principal components. Consider a data matrix $X \in \mathbb{R}^{n \times p}$, $n$ is the dimension for the objects and $p$ for the variables. Then we have a $p$ dimensional vector $w^{(i)} = (w_1, \ldots, w_p)^T$ with $||w^{(i)}|| = 1$, represents the weight or loading of each original variable for the $i$-th principal component. Thus we have a new vector of principal components $t$ with dimension $k, k < p$, is given by

$$t = Xw.$$ 

Since matrix $X$ contains variability to the problem, $t$ is the linear transformation of $X$, it is reasonable to consider the variance of $t$ instead. We know that the first component contributes most of the variability to the data set, to calculate the first component equals to maximizes the variance of $t$. Thus it is given by the formula:

$$w^{(1)} = \arg \max_{||w^{(i)}||=1} (t^T t) = \arg \max_{||w^{(i)}||=1} ||Xw||^2 = \arg \max_{||w^{(i)}||=1} (w^T X^T X w) = \arg \max \left\{ \frac{w^T X^T X w}{w^T w} \right\}.$$ 

It turns out that this equals to calculate the largest eigenvalue for the symmetric matrix $X^T X$ and $w^{(1)}$ is the corresponding eigenvector. With $w^{(1)}$ calculated, the first principal component could be given as

$$t^{(1)} = Xw^{(1)}.$$ 

For the further components, the $k$th component can be calculating the first $k-1$ principal components from $X$:

$$\hat{X}_k = X - \sum_{i=1}^{k-1} X w^{(i)} (w^{(i)})^T,$$

$$w^{(k)} = \arg \max_{||w^{(i)}||=1} ||\hat{X}_k w||^2 = \arg \max_{||w^{(i)}||=1} (w^T \hat{X}_k^T \hat{X}_k w) = \arg \max \left\{ \frac{w^T \hat{X}_k^T \hat{X}_k w}{w^T w} \right\}.$$ 

As given the first principal component, this gives the remaining eigenvectors of $\hat{X}_1^T \hat{X}_1$. Thus the weight or loading vector $w^{(k)}$ are eigenvectors of $\hat{X}_k^T \hat{X}_k$.

There are also other approaches to get the principal components, such as singular value decomposition (SVD) of $X$ which can refer to Richardson (2009).
Chapter 4
Data Analysis

This chapter will focus on the analysis we have done towards the available data sets. First based on the simulated data, we will do multivariate adaptive regression splines to get the tipping points. We will research on how the temperatures (independent variables), sources of energy (dependent variables) and labels can affect the tipping points. Then we will use an automatic clustering method to cluster the data set and compare the result with the manual labels. In the final part, we will go back to the real meter data to consider about the external weather information. By using principal component analysis, we can determine the most important weather variables and maybe use them later to implement the MARS model.

4.1 MARS analysis on separate source of energy as a whole

In this section we will begin to analyze on the simulated data which could help us to better understand the energy signatures. In MARS we focus on the relationship between the energy usages and temperatures. From Figures 4.13 and 4.14 we could find there is at least one tipping point so that linear regression is not applicable here, instead we use multivariate adaptive regression splines (MARS) which can deal with breaking points, there are open source implementations usually called “earth”.

According to the needs of the company, we restrict the MARS analysis to two pieces within the model, i.e. there is only one breaking point. This could be realized by the adding R command \texttt{nk = 4}. The results are showing in the following list and we also plot the regression line with 95% confidence level based on the data set in Figure 4.1 and 4.2.

First we do the analysis on the whole data set corresponding to gas. Restricting that there is one tipping point, we have the result shown with Formula (4.1):

\[
U_{\text{gas}} = \begin{cases} 
0.0144 + 0.8295(13.2375 - T) & T < 13.2375 \\
0.0144 & T \geq 13.2375
\end{cases} 
\] (4.1)

From the result we could get the tipping point at around $13^\circ C$ for gas.

Next we do the analysis based on electricity. As what we have done to gas, restricting to one tipping point, we have the result in Formula (4.2) and show the regression line...
with 95% confidence level based on each cluster in Figure 4.2.

\[
U_{electricity} = \begin{cases} 
5.379 - 0.131(13.175 - T) & T < 13.175 \\
5.379 + 1.073(T - 13.175) & T \geq 13.175 
\end{cases} 
\] (4.2)

From the result above we could get the tipping point at around 13°C for electricity.

However from the both figures, we can see there should be at least two clusters that the energy behaves differently from each other. The regression line is almost in the middle of the two clusters of data points. Then working on the whole data set is not accurate.
4.2 Exploration on the MARS analysis with respect to labels

4.2.1 MARS analysis within each cluster

Due to the inaccuracy of the results to do MARS on the whole data set, we do the MARS analysis within each cluster. Based on gas and restricting to only one tipping point to every cluster, we show the models in Formula (4.3) and show the regression line with 95% confidence level based on each cluster in Figure 4.3.

\[
U_{\text{gas}} = \begin{cases} 
  \text{day} : & 0.0437 + 2.686(12.7875 - T) & \text{if } T < 12.7875 \\
  & 0.0437 & \text{if } T \geq 12.7875 \\
  \text{night} : & 0.0694 + 0.4399(10.7 - T) & \text{if } T < 10.7 \\
  & 0.0694 & \text{if } T \geq 10.7 \\
  \text{dayWeekend} : & 0.0411 + 0.6405(8.375 - T) & \text{if } T < 8.375 \\
  & 0.0411 - 0.0394(T - 8.375) & \text{if } T \geq 8.375 \\
  \text{nightWeekend} : & 0.0547 + 0.5964(8.0125 - T) & \text{if } T < 8.0125 \\
  & 0.0547 - 0.6483(T - 8.0125) & \text{if } T \geq 8.0125 
\end{cases} \tag{4.3}
\]

We find that all the four tipping points are even lower than the tipping point for the whole set.

Then based on electricity and restricting to only one tipping point to every cluster, we show the models in Formula (4.4) and show the regression line with 95% confidence level based on each cluster in Figure 4.4:

\[
U_{\text{electricity}} = \begin{cases} 
  \text{day} : & 17.3784 & \text{if } T < 9.1125 \\
  & 17.3784 + 0.9599(T - 9.1125) & \text{if } T \geq 9.1125 \\
  \text{night} : & 0.1398 - 0.1245(13.95 - T) & \text{if } T < 13.95 \\
  & 0.1398 + 0.1064(T - 13.95) & \text{if } T \geq 13.95 \\
  \text{dayWeekend} : & 0.1133 - 0.0056(26.25 - T) & \text{if } T < 26.25 \\
  & 0.1133 + 0.5871(T - 26.25) & \text{if } T \geq 26.25 \\
  \text{nightWeekend} : & 0.0973 - 0.0055(22.9875 - T) & \text{if } T < 22.9875 \\
  & 0.0973 + 0.4647(T - 22.9875) & \text{if } T \geq 22.9875 
\end{cases} \tag{4.4}
\]

This result has shown a different output from what we get based on the whole data set. Only for the cluster with label “day” has a smaller tipping point while other clusters have a larger tipping point.

4.2.2 MARS analysis with different intercepts for each cluster

Now instead of doing MARS analysis within each cluster, we consider the labels into the model as an extra independent variable to see how this variable influence the final results. The new model will give us the relationship between the energy usages and temperatures. The results show that for all four clusters, they share the same slope and
breaking point, the labels only make sense in the intercepts for each cluster since it is in format as character. We show the result in Formulas (4.5) and (4.6).

\[
U_{gas} = \begin{cases} 
\text{day} : & \begin{align*} 
6.604 + 0.864(13.2375 - T) & \quad T < 13.2375 \\
6.604 - 0.237(T - 13.2375) & \quad T \geq 13.2375 
\end{align*} \\
\text{night} : & \begin{align*} 
-2.566 + 0.864(13.2375 - T) & \quad T < 13.2375 \\
-2.566 - 0.237(T - 13.2375) & \quad T \geq 13.2375 
\end{align*} \\
\text{dayWeekend} : & \begin{align*} 
-1.8 + 0.864(13.2375 - T) & \quad T < 13.2375 \\
-1.8 - 0.237(T - 13.2375) & \quad T \geq 13.2375 
\end{align*} \\
\text{nightWeekend} : & \begin{align*} 
-2.658 + 0.864(13.2375 - T) & \quad T < 13.2375 \\
-2.658 - 0.237(T - 13.2375) & \quad T \geq 13.2375 
\end{align*} 
\end{cases}
\]
From the results we can see that for the gas, the tipping points does not change from the result based on the whole data set. For the electricity, the tipping point is smaller than the one using the whole data set. However compared to Formula (4.4), this tipping point is in the middle level of four separated tipping points.
4.2.3 Using labels as an independent variable to get separated models

Now for a further step, we consider the interaction between the temperatures and labels into the model to see how labels influence the final results. This works by adding the interaction part into the model, and the new model will give us the relationship between the energy usages and temperatures for each cluster in one step. The result will give the four clusters their own slope, breaking point and intercept. The results may be different from what we get from Formulas (4.3) and 4.4, this is because that we still consider the data set as a whole despite we get separate results and the error is minimized for the whole set. We show the result in Formulas (4.7) and (4.8).

\[
U_{\text{gas}} = \begin{cases} 
  \text{day} : & 87.987 + 2.576(13.2375 - T) & T < 13.2375 \\
  & 87.987 & T \geq 13.2375 \\
  \text{night} : & 58.907 + 0.430(13.925 - T) & T < 13.925 \\
  & 58.907 - 0.103(T - 13.925) & T \geq 13.925 \\
  \text{dayWeekend} : & 59.190 + 0.4525(13.6625 - T) & T < 13.6625 \\
  & 59.190 & T \geq 13.6625 \\
  \text{nightWeekend} : & 59.072 + 0.460(13.875 - T) & T < 13.875 \\
  & 59.072 & T \geq 13.875 
\end{cases}
\]

\[
U_{\text{electricity}} = \begin{cases} 
  \text{day} : & 13.771 - 0.1477(10.7875 - T) & T < 10.7875 \\
  & 13.771 + 1.0286(T - 10.7875) & T \geq 10.7875 \\
  \text{night} : & -3.042 - 0.004(10.725 - T) & T < 10.725 \\
  & -3.042 + 0.081(T - 10.725) & T \geq 10.725 \\
  \text{dayWeekend} : & -3.053 - 0.035(10.525 - T) & T < 10.525 \\
  & -3.053 + 0.046(T - 10.525) & T \geq 10.525 \\
  \text{nightWeekend} : & -3.030 - 0.024(10.6625 - T) & T < 10.6625 \\
  & -3.030 + 0.045(T - 10.6625) & T \geq 10.6625 
\end{cases}
\]

The output shows that for both gas and electricity, although every cluster owns different tipping points, the tipping points are close to each other. Even compared to Formulas (4.5) and (4.6), the tipping points for the cluster “day” are the same for both gas and electricity. Note that in this model, the intercepts are somehow strange while the tipping points and the slope are close to what we got before.

4.3 MARS analysis on the combination of sources of energy

Since the gas is mainly used for heating, it depends on the current outdoor climate and even some hours before. This is caused by the mass effect of the building. Another effect is the interaction between the heating and the electricity for which ENGIE calls “heat
There are two (free) heat loads: one is the sun radiation through the window, and the second is from the electrical applications such as computers and lighting. The results show that about 65% of the electricity is affecting the heating. Due to the potential interaction within different energy sources, we may not do the automatic clustering on energy values separately. Now we consider the MARS analysis on the combination of sources of energy. This could be done in MARS directly. We would also notice that this result could be much different from what we got before. Compared to the reality we will make a choice whether using this way of analyzing the model. As we did before, we may consider the independent variable as temperature itself, adding the label and then adding the interaction part.

First we do the MARS analysis on both gas and electricity together with respect to temperatures based on the whole data set. The results are shown in Formulas (4.9) and (4.10).

\[
U_{gas} = \begin{cases} 
9.113 + 3.965(2.079 - T) & T < 2.079 \\
9.113 - 0.130(T - 2.079) & T \geq 2.079
\end{cases} \tag{4.9}
\]

\[
U_{electricity} = \begin{cases} 
9.113 - 0.189(27.433 - T) & T < 27.433 \\
9.113 + 1.593(T - 27.433) & T \geq 27.433
\end{cases} \tag{4.10}
\]

The results show a relatively slow tipping point for gas and high tipping point for electricity which may be far away from reality.

Then we follow by the MARS analysis within each cluster based on the combination of sources of energy. The results are shown in Formulas (4.11) and (4.12).

\[
U_{gas} = \begin{cases} 
day : \begin{cases} 
14.535 + 0.462(3.923 - T) & T < 3.923 \\
14.535 - 0.283(T - 3.923) & T \geq 3.923
\end{cases} \\
night : \begin{cases} 
12.312 + 10.178(0.773 - T) & T < 0.773 \\
12.312 - 0.266(T - 0.773) & T \geq 0.773
\end{cases}
\end{cases} \tag{4.11}
\]

\[
U_{electricity} = \begin{cases} 
day : \begin{cases} 
14.535 - 0.456(26.5376 - T) & T < 26.5376 \\
14.535 + 1.486(T - 26.5376) & T \geq 26.5376
\end{cases} \\
night : \begin{cases} 
12.312 - 1.616(4.818 - T) & T < 4.818 \\
12.312 & T \geq 4.818
\end{cases}
\end{cases} \tag{4.12}
\]

The results show a relatively slow tipping point for gas for all four clusters and tipping points for electricity for all clusters which may be far away from reality. Besides, the results
in source of energy show a contradiction to the reality that with outdoor temperatures going up, the consumption of electricity is getting stable.

Finally we would like to try to take the labels into account and see how the results change. However in this case, MARS could not work if there are more than two independent variables.

Comparing the results we get before and according to the reality, it is suggested not to do the MARS analysis based on the combination of sources of energy.

4.4 Exploration on the MARS analysis with respect to temperatures

Due to the heat load in the building, we know that the current energy usages may have some relationship with the past dynamics in the building. However we do not know the precise relationship, thus we now explore how the temperatures could influence the model. We will research on the derivatives of the temperatures and the one-hour-lag temperatures. By doing the MARS analysis with these variables as what we did before, and combine with the previous research results, we may choose the suitable way of using temperatures.

4.4.1 Taking the derivative of temperatures as an extra independent variable

Now first we use the derivative of the temperatures as an extra independent variable. According to the previous research, we may suggest to do the MARS analysis based on each source of energy instead of the combination. For all the results shown in this section, we will only give the formula concerned the tipping points for outdoor temperatures.

Based on the whole data set

First we analyze on the whole data set regardless of the labels. Based on gas we have the results shown in Formula (4.13), restricting that there is one tipping point on temperatures:

\[
U_{gas} = \begin{cases} 
1.802 + 0.836(13.2375 - T) & T < 13.2375 \\
1.802 & T \geq 13.2375 
\end{cases}
\quad (4.13)
\]

Next we do the analysis on the whole data set corresponding to electricity. Restricting that there is one tipping point on temperatures, we have the result in Formula (4.14).

\[
U_{electricity} = \begin{cases} 
2.627 - 0.019(13.125 - T) & T < 13.125 \\
2.627 + 0.988(T - 13.125) & T \geq 13.125 
\end{cases}
\quad (4.14)
\]

By comparing the results shown in Formulas (4.1) and (4.2), there is little change to the tipping points for both sources of energy.
Within each cluster

Then we follow by the MARS analysis with respect to temperatures within each cluster. The results are shown in Formulas (4.15) and (4.16).

\[
U_{\text{gas}} = \begin{cases} 
\text{day} : & 0.061 + 2.692(12.7625 - T) \quad T < 12.7625 \\
                & 0.061 \quad T \geq 12.7625 \\
\text{night} : & 0.069 + 0.440(10.7 - T) \quad T < 10.7 \\
                & 0.069 \quad T \geq 10.7 \\
\text{dayWeekend} : & 0.407 + 0.640(8.3875 - T) \quad T < 8.3875 \\
                    & 0.407 - 0.039(8.3875 - T) \quad T \geq 8.3875 \\
\text{nightWeekend} : & 0.576 + 0.600(7.925 - T) \quad T < 7.925 \\
                     & 0.576 - 0.068(7.925 - T) \quad T \geq 7.925 
\end{cases} 
\tag{4.15}
\]

By comparing the results shown in Formulas (4.3) and (4.4), still there is little change to the tipping points for all clusters regardless of source of energy.

Using labels to get separate models

Next we follow by the MARS analysis on each source of energy with respect to temperatures, using four labels to get models separately. The results are shown in Formulas (4.17) and (4.18).

\[
U_{\text{electricity}} = \begin{cases} 
\text{day} : & 17.378 \\
        & 17.378 + 0.960(T - 9.1125) \quad T \geq 9.1125 \\
\text{night} : & 0.140 - 0.012(13.95 - T) \quad T < 13.95 \\
                 & 0.140 + 0.106(T - 13.95) \quad T \geq 13.95 \\
\text{dayWeekend} : & 0.113 - 0.006(26.25 - T) \quad T < 26.25 \\
                       & 0.113 + 0.587(T - 26.25) \quad T \geq 26.25 \\
\text{nightWeekend} : & 0.095 - 0.005(22.9 - T) \quad T < 22.9 \\
                       & 0.095 + 0.456(T - 22.9) \quad T \geq 22.9 
\end{cases} 
\tag{4.16}
\]

\[
U_{\text{gas}} = \begin{cases} 
\text{day} : & 88.743 + 2.585(13.2375 - T) \quad T < 13.2375 \\
                & 88.743 + 0.047(T - 13.2375) \quad T \geq 13.2375 \\
\text{night} : & 59.552 + 0.423(13.925 - T) \quad T < 13.925 \\
                & 59.552 + 0.073(T - 13.925) \quad T \geq 13.925 \\
\text{dayWeekend} : & 59.942 + 0.432(13.6625 - T) \quad T < 13.6625 \\
                      & 59.942 \quad T \geq 13.6625 \\
\text{nightWeekend} : & 59.721 + 0.455(13.875 - T) \quad T < 13.875 \\
                     & 59.721 \quad T \geq 13.875 
\end{cases} 
\tag{4.17}
\]
CHAPTER 4. DATA ANALYSIS

\[ U_{electricity} = \begin{cases} 
\text{day} : & \begin{cases} 
13.771 - 0.1477(10.7875 - T) & T < 10.7875 \\
13.771 + 1.0286(T - 10.7875) & T \geq 10.7875 
\end{cases} \\
\text{night} : & \begin{cases} 
-3.042 - 0.004(10.725 - T) & T < 10.725 \\
-3.042 + 0.081(T - 10.725) & T \geq 10.725 
\end{cases} \\
\text{dayWeekend} : & \begin{cases} 
-3.053 - 0.005(10.525 - T) & T < 10.525 \\
-3.053 + 0.046(T - 10.525) & T \geq 10.525 
\end{cases} \\
\text{nightWeekend} : & \begin{cases} 
-3.030 + 0.001(10.6625 - T) & T < 10.6625 \\
-3.030 + 0.045(T - 10.6625) & T \geq 10.6625 
\end{cases} 
\end{cases} \] 

By comparing the results shown in Formulas (4.7) and (4.8), there is also little change to the tipping points for all clusters regardless of source of energy.

Comparing the results with derivative of temperatures as an extra independent variable and results with only temperatures, the tipping points has little difference for any model and cluster. Thus the derivative could be ignored from the model.

### 4.4.2 Using the temperatures one hour ago as the independent variable

Now we use the one-hour-lag temperatures as the independent variable. Here we still do the MARS analysis based on each source of energy value instead of the combination.

**Based on the whole data set**

First we analyze on the whole data set regardless of the labels. Based on gas we have the result shown in Formula (4.19), restricting that there is one one tipping point on temperatures:

\[ U_{gas} = \begin{cases} 
0.026 + 0.893(12.7375 - T) & T < 12.7375 \\
0.026 & T \geq 12.7375 
\end{cases} \] (4.19)

Next we do the analysis on the whole data set corresponding to electricity. Restricting that there is one one tipping point on temperatures, we have the result in Formula (4.20).

\[ U_{electricity} = \begin{cases} 
5.482 - 0.125(12.7 - T) & T < 12.7 \\
5.482 + 0.837(T - 12.7) & T \geq 12.7 
\end{cases} \] (4.20)

By comparing the results shown in Formulas (4.1) and (4.2), the tipping points are smaller than the original one from current temperatures for both sources of energy.
Within each cluster

Then we follow by the MARS analysis on each energy value with respect to temperatures within each cluster. The results are shown in Formulas (4.21) and (4.22).

\[ U_{\text{gas}} = \begin{cases} 
\text{day} : & \begin{cases} 
0.062 + 2.593(12.6875 - T) & T < 12.6875 \\
0.062 & T \geq 12.6875 
\end{cases} \\
\text{night} : & \begin{cases} 
0.089 + 0.464(10.4875 - T) & T < 10.4875 \\
0.089 & T \geq 10.4875 
\end{cases} \\
\text{dayWeekend} : & \begin{cases} 
0.376 + 0.586(8.5625 - T) & T < 8.5625 \\
0.376 - 0.038(T - 8.5625) & T \geq 8.5625 
\end{cases} \\
\text{nightWeekend} : & \begin{cases} 
0.050 + 0.555(9.525 - T) & T < 9.525 \\
0.050 & T \geq 9.525 
\end{cases} 
\right. \] (4.21)

\[ U_{\text{electricity}} = \begin{cases} 
\text{day} : & \begin{cases} 
17.379 & T < 8.65 \\
17.379 + 0.953(T - 8.65) & T \geq 8.65 
\end{cases} \\
\text{night} : & \begin{cases} 
0.160 - 0.014(14.175 - T) & T < 14.175 \\
0.160 + 0.082(T - 14.175) & T \geq 14.175 
\end{cases} \\
\text{dayWeekend} : & \begin{cases} 
0.125 - 0.006(26.775 - T) & T < 26.775 \\
0.125 + 0.895(T - 26.775) & T \geq 26.775 
\end{cases} \\
\text{nightWeekend} : & \begin{cases} 
0.084 - 0.005(23.7375 - T) & T < 23.7375 \\
0.084 + 0.45(T - 23.7375) & T \geq 23.7375 
\end{cases} 
\right. \] (4.22)

By comparing the results shown in Formula (4.3), we can see that for gas, the tipping points is not just getting smaller or larger to all labels from different sources of energy. The results oscillate within a small range.

Using labels to get separate models

Then we follow by the MARS analysis on each energy value with respect to one-hour-lag temperatures, using four labels to get models separately. The results are shown in Formulas (4.23) and (4.24).

\[ U_{\text{gas}} = \begin{cases} 
\text{day} : & \begin{cases} 
84.162 + 2.579(12.7375 - T) & T < 12.7375 \\
84.162 & T \geq 12.7375 
\end{cases} \\
\text{night} : & \begin{cases} 
56.206 + 0.449(13.3 - T) & T < 13.3 \\
56.206 - 0.070(T - 13.3) & T \geq 13.3 
\end{cases} \\
\text{dayWeekend} : & \begin{cases} 
56.188 + 0.41(13.125 - T) & T < 13.125 \\
56.188 & T \geq 13.125 
\end{cases} \\
\text{nightWeekend} : & \begin{cases} 
56.407 + 0.491(13.35 - T) & T < 13.35 \\
55.407 & T \geq 13.35 
\end{cases} 
\right. \] (4.23)
\[
U_{\text{electricity}} = \begin{cases}
\text{day} : & 14.804 - 0.117(10 - T) \quad T < 10 \\
& 14.804 + 1.010(T - 10) \quad T \geq 10 \\
\text{night} : & -2.197 - 0.004(9.975 - T) \quad T < 9.975 \\
& -2.197 + 0.145(T - 9.975) \quad T \geq 9.975 \\
\text{dayWeekend} : & -2.206 - 0.005(9.7375 - T) \quad T < 9.7375 \\
& -2.206 + 0.135(T - 9.7375) \quad T \geq 9.7375 \\
\text{nightWeekend} : & -2.194 + 0.001(9.8375 - T) \quad T < 9.8375 \\
& -2.194 + 0.040(T - 9.8375) \quad T \geq 9.8375 
\end{cases}
\] (4.24)

By comparing the results shown in Formula (4.7) and (4.8), the tipping points are smaller than the original one from current temperatures for both sources of energy.

Now we will conclude the tipping points from every model in the Tables 4.1 and 4.2. We will only keep the first two digits after the decimal.

<table>
<thead>
<tr>
<th>Source</th>
<th>Variables</th>
<th>Data Set</th>
<th>day</th>
<th>night</th>
<th>dayWeekend</th>
<th>nightWeekend</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Within each cluster</td>
<td>12.79</td>
<td>10.70</td>
<td>8.38</td>
<td>8.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>With labels</td>
<td>13.24</td>
<td>13.93</td>
<td>13.66</td>
<td>13.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within each cluster</td>
<td>12.76</td>
<td>10.70</td>
<td>8.39</td>
<td>7.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>With labels</td>
<td>13.24</td>
<td>13.93</td>
<td>13.66</td>
<td>13.88</td>
</tr>
<tr>
<td></td>
<td>One-hour-lag temperature</td>
<td>Whole data set</td>
<td>12.74</td>
<td>12.74</td>
<td>12.74</td>
<td>12.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within each cluster</td>
<td>12.69</td>
<td>10.49</td>
<td>8.56</td>
<td>9.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>With labels</td>
<td>12.74</td>
<td>13.30</td>
<td>13.13</td>
<td>13.35</td>
</tr>
</tbody>
</table>

Continued on next page
### Table 4.1 – continued from previous page

<table>
<thead>
<tr>
<th>Source</th>
<th>Variables</th>
<th>Data Set</th>
<th>day</th>
<th>night</th>
<th>dayWeekend</th>
<th>nightWeekend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combination</td>
<td>Temperature</td>
<td>Whole data set</td>
<td>2.08</td>
<td>2.08</td>
<td>2.08</td>
<td>2.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within each cluster</td>
<td>3.92</td>
<td>0.77</td>
<td>0.76</td>
<td>0.77</td>
</tr>
</tbody>
</table>

### Table 4.2: Summary of tipping points for electricity

<table>
<thead>
<tr>
<th>Source</th>
<th>Variables</th>
<th>Data Set</th>
<th>day</th>
<th>night</th>
<th>dayWeekend</th>
<th>nightWeekend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electricity</td>
<td>Temperature</td>
<td>Whole data set</td>
<td>13.18</td>
<td>13.18</td>
<td>13.18</td>
<td>13.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>With different intercept</td>
<td>10.79</td>
<td>10.79</td>
<td>10.79</td>
<td>10.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within each cluster</td>
<td>9.11</td>
<td>13.95</td>
<td>26.25</td>
<td>22.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>With labels</td>
<td>10.79</td>
<td>10.73</td>
<td>10.53</td>
<td>10.66</td>
</tr>
<tr>
<td>Temperature and derivative</td>
<td>Whole data set</td>
<td>13.13</td>
<td>13.13</td>
<td>13.13</td>
<td>13.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within each cluster</td>
<td>9.11</td>
<td>13.95</td>
<td>26.25</td>
<td>22.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>With labels</td>
<td>10.79</td>
<td>10.73</td>
<td>10.53</td>
<td>10.67</td>
</tr>
<tr>
<td>One-hour-lag temperature</td>
<td>Whole data set</td>
<td>12.70</td>
<td>12.70</td>
<td>12.70</td>
<td>12.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within each cluster</td>
<td>8.65</td>
<td>14.18</td>
<td>26.78</td>
<td>23.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>With labels</td>
<td>10.00</td>
<td>9.98</td>
<td>9.74</td>
<td>9.84</td>
</tr>
<tr>
<td>Combination</td>
<td>Temperature</td>
<td>Whole data set</td>
<td>27.43</td>
<td>27.43</td>
<td>27.43</td>
<td>27.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within each cluster</td>
<td>26.54</td>
<td>4.82</td>
<td>0.46</td>
<td>0.26</td>
</tr>
</tbody>
</table>
4.5 Automatic clustering of simulated data

We first try the automatic clustering method of PAM based on the simulated data. As introduced in the Chapter 3 we use R command `pamk()` in the package `fpc`. We will do the clustering separately on the specific source of energy and compare with the original given labels. Then we combine the sources of energy and do the clustering, and we compare the results with the clustering results from separate source of energy. Here since we use the function named `pamk` in R, the estimated number of clusters is directly given by the result of clustering.

4.5.1 Clustering based only on gas

First we cluster the data set based on the energy values of gas and give the output as in the following Table 4.3.

<table>
<thead>
<tr>
<th></th>
<th>day</th>
<th>dayWeekend</th>
<th>night</th>
<th>nightWeekend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1548</td>
<td>1109</td>
<td>3452</td>
<td>1550</td>
</tr>
<tr>
<td>2</td>
<td>982</td>
<td>11</td>
<td>90</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 4.3: Clusters with fpc gas label

If we look at the medoids of two clusters and the average silhouette widths for each cluster as well as the whole data set, we can get the following Table 4.4:

<table>
<thead>
<tr>
<th>Index of cluster</th>
<th>medoid</th>
<th>average silhouette width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>0.8944</td>
</tr>
<tr>
<td>2</td>
<td>20.475</td>
<td>0.5001</td>
</tr>
<tr>
<td>Average silhouette width</td>
<td>0.8450</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4: Output of clusters with fpc gas label

The average silhouette width shows that cluster 2 is not well classified, but it has already been the best result for the overall clustering if we take the number of clusters from 2 to 4. If we insist to do the automatic clustering based on only gas usages, we may try other methods to improve the accuracy.

From the table we can see that based on the energy values of gas, automatic clustering using PAM leads to two clusters. This number is much smaller than the original number of clusters. In order to see what is the clustering result for the original data set, we will use the index of automatic clustering results from Table 4.3 as a new label named with “fpcgaslabel”. Then “fpcgaslabel” contains “gas1” and “gas2” which stand for the first and second cluster. This is clear to see how all the data points are clustered under `fpc`, which is shown in Figures 4.5 and 4.6. Figure 4.5 shows that the data points are clustered into two parts basically based on a horizontal line at round \( y = 10 \). This means that the data points with gas usages more than \( 10W \) is considered to be one cluster, while the rest is another cluster. Then we apply this result to the data points with electricity usages shown in Figure 4.6. The figure shows that every cluster contains both data points from “day” and combination of rest three labels. Then this labeling method does not make sense on the data points with electricity usages.
CHAPTER 4. DATA ANALYSIS

Although \texttt{pamk()} has chosen the “best” clustering based on average silhouette width, we can still to cluster the data set using \texttt{pam()} from the package \texttt{cluster} to see the results if we would like more clusters. The output is the same as shown in Figure 4.5 if given 2 clusters in \texttt{pam()} . For giving more number of clusters, no matter 3 or 4 clusters we require, the data set is divided simply by the horizontal lines, as an example shown in Figure 4.7. This may also indicate that if we only based on the gas usages, the automatic clustering divides the clusters simply by the usages level, this is also similar.
to what hierarchy clustering does on gas usages. If we apply the generated labels to the electricity, the plot shows that it is more like a random clustering as shown in Figure 4.8.

![Figure 4.7: Gas usages with 3 clusters](image)

![Figure 4.8: Electricity usages with 3 clusters](image)

As what we have done based on the usages of gas, we will then use the usages of electricity to apply this automatic labeling method in the next section.

### 4.5.2 Clustering based only on electricity

Now we cluster the data set based on the energy values of electricity and give the output as in the following Table 4.5.

<table>
<thead>
<tr>
<th></th>
<th>day</th>
<th>dayWeekend</th>
<th>night</th>
<th>nightWeekend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1120</td>
<td>3541</td>
<td>1568</td>
</tr>
<tr>
<td>2</td>
<td>1625</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>905</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.5: Clusters with fpc electricity label

If we look at the medoids of three clusters and the average silhouette widths for each cluster as well as the whole data set, we can get the following Table 4.6:

<table>
<thead>
<tr>
<th>Index of cluster</th>
<th>medoid</th>
<th>average silhouette width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$2.1877 \times 10^{-4}$</td>
<td>0.9885</td>
</tr>
<tr>
<td>2</td>
<td>17.338</td>
<td>0.9205</td>
</tr>
<tr>
<td>3</td>
<td>27.385</td>
<td>0.7266</td>
</tr>
</tbody>
</table>

Table 4.6: Output of clusters with fpc electricity label

The average silhouette width shows that all clustered can be classified well. This indicated that it may be accurate if we only use electricity usages to do the automatic clustering and apply the labels to the gas usages.
From the table we can see that based on the energy values of electricity, the automatic clustering using PAM leads to three clusters. This number is still smaller than the original number of clusters. By looking at the results, we can see that the data points with label “day” and the rest points are almost separately perfectly. This result is easy to understand due to the gap between two main clusters which we could see from the Figure 4.9. This also shows that the property we used in last chapter to check whether the clustering is done is accurate, i.e. to use the big gap between the usages as a criteria to decide “day”. For the combination of data points with “night”, “dayWeekend” and “nightWeekend”, we can see some similarities in the energy values. For example for every Sunday 23 : 00 : 00 to Monday 00 : 00 : 00, the label changes from “nightWeekend” to “night” directly, however the working condition of the building should not change much in just one hour.

In order to see what the clustering result is for the original data set, we will use the index of automatic clustering results from Table 4.5 as a new label named with “fpceleclabel”. Then “fpceleclabel” contains “elec1”, “elec2” and “elec3” which stand for the three clusters. Then we take the clustering results from Table 4.5 as the label see how all the data points behave, which is shown in Figures 4.10 and 4.9.

The Figure 4.9 shows that the data points with “day” is clustered into two parts basically based on a horizontal line at round $y = 22.5$. Then we apply this result to the data points with gas usages shown in Figure 4.10. Except the data points with zero gas usages, the data points are well clustered into two parts, which is also what we expect to have: the green part represents the data points with “day” and the red part represents the rest of the data points.

Still here we can use `pam()` from the package `cluster` to see the results if we would like other number of clusters. In Figure 4.9 we can see that the data points with “day” is divided by a horizontal line which we can not easily explain. However if we compare the result with the output of other numbers of clusters, we may get some ideas of this
clustering method. Actually the output of 2 clusters suggests that the electricity usages can be divided due to the big gap. Then for more clusters, the algorithm simply gets more horizontal lines to divide the original data set, and mainly based on the level of the energy usages with label “day”.

If we apply the generated labels to the gas, the plot shows almost the same for the none zero data points with three different numbers of clusters as shown in Figures 4.11 and 4.12.
4.5.3 Clustering based on the combination of energy values

Now we combine gas and electricity, the result of clustering gives another result shown in Table 4.7. We compare this table with Table 4.5 and find that the first clusters from both output are almost the same. This cluster is actually made up with all the data points from the original “night”, “dayWeekend” and “nightWeekend”. However for the cluster with label “day”, it shows a different result.

<table>
<thead>
<tr>
<th></th>
<th>day</th>
<th>dayWeekend</th>
<th>night</th>
<th>nightWeekend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1120</td>
<td>3540</td>
<td>1568</td>
</tr>
<tr>
<td>2</td>
<td>1027</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1503</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.7: Clusters with fpc clustering

If we look at the medoids of three clusters and the average silhouette widths for each cluster as well as the whole data set, we can get the following Table 4.8:

<table>
<thead>
<tr>
<th>Index of cluster</th>
<th>medoid</th>
<th>average silhouette width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.5829 * 10^{-5}</td>
<td>0.8984</td>
</tr>
<tr>
<td>2</td>
<td>17.340</td>
<td>0.5696</td>
</tr>
<tr>
<td>3</td>
<td>26.062</td>
<td>0.7036</td>
</tr>
</tbody>
</table>

Table 4.8: Output of clusters with fpc clustering

The average silhouette width shows that cluster 2 is not well classified, but it has already been the best result for the overall clustering if we take the number of clusters from 2 to 4. Although the first cluster is almost the same, however due to the difference between the medoids of cluster 1 from Tables 4.6 and 4.8, the previous method shows a higher average silhouette width, which indicates that the medoid is more central. Then despite the medoids of cluster 2 and 3 are similar to each other from Tables 4.6 and 4.8, the average silhouette width for both clusters shows that doing the automatic clustering based on the combination of gas and electricity does not lead to a better clustering result. Then if we consist to do the automatic clustering based on the combination of gas and electricity usages, we may try other methods to improve the accuracy.

In order to see what the clustering result for the original data set is, we will use the index of automatic clustering results from Table 4.7 as a new label named with “fpclabel”. Then “fpclabel” contains “1”, “2” and “3” which stand for the separate three clusters. Now we add this label to both gas and electricity and plot the data points with color shown in Figures 4.13 and 4.14. By comparing the figures we have with three different fpc labels, we can find that for gas, the overlapping part may cause the trouble for us to do the clustering. For electricity, we may conclude for all the figures that the data points with original label “night”, “dayWeekend” and “nightWeekend” could be regarded as one cluster, the remaining problem is whether it is necessary to cluster the data points with original label “day”. This also indicates that for the processes happened in the time points with label ‘night”, “dayWeekend” and “nightWeekend” have a relatively big similarities.
Still here we can use `pam()` from the package `cluster` to see the results if we would like other number of clusters. If we specify the number of clusters to be 2, the results of the clustering in both energies are shown in Figures 4.15 and 4.16. It is almost the same as what we have got based only on electricity with 2 clusters. The differences lie only in the overlapping part for gas values. If we specify the number of clusters to be 4, the results are shown in Figures 4.17 and 4.18. It is quite different from any results what we have gotten so far. It seems to be a combination of results from 2 clusters and
horizontal lines in both energy values. This may indicate that the data set can be divided by 4 clusters based on the seasonal effect and working/off hours together.

Based on the different fpc labels and we have shown the clustering result with different number of clusters. No matter if we fix the label and compare with the results with different number of clusters or if we fix the number of clusters and compare the energy values we use for clustering, the differences indicate that there may be some interaction between the gas and electricity, thus we may consider the automatic clustering with the combination of sources of energy.

If we only look at the results from \texttt{pamk()}, we will conclude that no matter which fpc label (based on gas, electricity or the combination) is used, the number of clusters is smaller than the original number of clusters. This shows the truth that it may be redundant to cluster the original data set into four clusters.
4.6 Principal Component Analysis on the weather information

Now back to the real meter data, we have the external weather information from the weather stations from the Royal Netherlands Meteorological Institute (KNMI) which is shown in Table 4.9. This includes 11 variables that may affect the energy values. Here we would like to use the principal component analysis on these variables to reduce the dimension. They may also be involved in the MARS model later applying with the real meter data.

Table 4.9: Available weather information

<table>
<thead>
<tr>
<th>Mode</th>
<th>Remark - Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunRadiation</td>
<td>Sun radiation - MJ/m²</td>
</tr>
<tr>
<td>windAverage</td>
<td>Average wind speed per hour - m/s</td>
</tr>
<tr>
<td>bar</td>
<td>Atmospheric pressure - bar</td>
</tr>
<tr>
<td>heatDewpoint</td>
<td>Dew point temperature - °C</td>
</tr>
<tr>
<td>heatLowest</td>
<td>Lowest temperature in the past 6 hours - °C</td>
</tr>
<tr>
<td>rainDuration</td>
<td>Rain duration per hour - hours</td>
</tr>
<tr>
<td>rainTotal</td>
<td>Amount of rain - 0.1mm</td>
</tr>
<tr>
<td>sunDuration</td>
<td>Sunshine duration per hour - hours</td>
</tr>
<tr>
<td>windDirection</td>
<td>Wind direction - degrees</td>
</tr>
<tr>
<td>windGush</td>
<td>Maximum wind gush in the past hour - m/s</td>
</tr>
<tr>
<td>windSpeed</td>
<td>Average wind speed during the last 10 minutes of the past hour - m/s</td>
</tr>
</tbody>
</table>

Now the following analysis and results are based on the data from building 1 as an example. We know that the PCA is sensitive to the scale of variables, we will first show the results before normalizing the variables. Before doing the analysis, we simply take the cos of the “windDirection”. First we show the variance associated with the principal components in Figure 4.19. In the figure we could see that the first principal component could have explained most of the variability in the data set of all external weather information. Then we use the proportion of variance of each principal component to calculate the cumulative proportion of explained variance shown in Figure 4.20.

From the figure we can see that the first principal component could have reached more than 95% of the total variance (almost 98%), and this is enough to take only this component into consideration. Now we would like to know the composition of the principal components, especially the first one. The result is shown in Table 4.10. We will only keep the first three digits after the decimal.

---

1360=North, 90=East, 180=South, 270=West, 0=windless, 990=Variable direction
CHAPTER 4. DATA ANALYSIS

Figure 4.19: Variances for building 1

Figure 4.20: Cumulative Proportion of Variance Explained for building 1

Table 4.10: Composition of principal components for building 1

<table>
<thead>
<tr>
<th>Variables</th>
<th>Comp.1</th>
<th>Comp.2</th>
<th>Comp.3</th>
<th>Comp.4</th>
<th>Comp.5</th>
<th>Comp.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunRadiation</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>windAverage</td>
<td></td>
<td>0.327</td>
<td>-0.311</td>
<td>-0.752</td>
<td>0.106</td>
<td></td>
</tr>
<tr>
<td>bar</td>
<td>-0.601</td>
<td>-0.224</td>
<td>-0.102</td>
<td>-0.752</td>
<td>0.106</td>
<td></td>
</tr>
<tr>
<td>heatDewpoint</td>
<td>-0.757</td>
<td>-0.147</td>
<td>0.217</td>
<td>0.208</td>
<td>-0.120</td>
<td>-0.943</td>
</tr>
<tr>
<td>heatLowest</td>
<td></td>
<td></td>
<td></td>
<td>-0.120</td>
<td>-0.943</td>
<td></td>
</tr>
<tr>
<td>rainDuration</td>
<td>0.159</td>
<td>0.605</td>
<td>0.716</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rainTotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sunDuration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>windDirection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>windGush</td>
<td>-0.514</td>
<td>0.548</td>
<td>-0.477</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>windSpeed</td>
<td></td>
<td>0.327</td>
<td>-0.320</td>
<td></td>
<td></td>
<td>-0.101</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variables</th>
<th>Comp.7</th>
<th>Comp.8</th>
<th>Comp.9</th>
<th>Comp.10</th>
<th>Comp.11</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunRadiation</td>
<td>0.236</td>
<td></td>
<td>-0.853</td>
<td>-1.000</td>
<td></td>
</tr>
<tr>
<td>windAverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bar</td>
<td>-0.853</td>
<td></td>
<td></td>
<td>-1.000</td>
<td></td>
</tr>
<tr>
<td>heatDewpoint</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>heatLowest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rainDuration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rainTotal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sunDuration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>windDirection</td>
<td>-0.165</td>
<td>0.986</td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td>windGush</td>
<td>-0.614</td>
<td>-0.101</td>
<td>0.231</td>
<td></td>
<td></td>
</tr>
<tr>
<td>windSpeed</td>
<td>0.732</td>
<td>0.123</td>
<td>0.468</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 4. DATA ANALYSIS

From the table, it is clear to see that the first principal component is exactly “sun-Radiation”. Luckily this is not a combination of several variables, and it makes the most variability of the data set. However this is just because the scale of “sunRadiation” is much larger than the other variables, thus it absolutely contributes the most variances in the original data sets. In order to reduce the influence by the different scales of original variables, we scale the variables to have the unit variance.

We normalize the original variables to have the unit variance within the data set from building 1. We show the variance associated with the principal components in Figure 4.21. In the figure we could see that this time first seven principal components could have explained most of the variability in the data set of all external weather information. Then we use the proportion of variance of each principal component to calculate the cumulative proportion of explained variance shown in Figure 4.22.

![Figure 4.21: Variances for building 1](image)

![Figure 4.22: Cumulative Proportion of Variance Explained for building 1](image)

From the figure we can see that the first seven principal components could have reached over 95% of the total variance, this reduces the dimension of all available variables, however let us look at the composition of these principal components first. The result is shown in Table 4.11. We will only keep the first three digits after the decimal.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Comp.1</th>
<th>Comp.2</th>
<th>Comp.3</th>
<th>Comp.4</th>
<th>Comp.5</th>
<th>Comp.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunRadiation</td>
<td>0.087</td>
<td>-0.556</td>
<td>0.124</td>
<td>0.371</td>
<td>-0.0143</td>
<td>-0.117</td>
</tr>
<tr>
<td>windAverage</td>
<td>0.502</td>
<td>0.061</td>
<td>0.253</td>
<td>-0.065</td>
<td>-0.019</td>
<td>0.119</td>
</tr>
<tr>
<td>bar</td>
<td>-0.285</td>
<td>-0.089</td>
<td>0.238</td>
<td>0.125</td>
<td>0.2081</td>
<td>0.883</td>
</tr>
<tr>
<td>heatDewpoint</td>
<td>0.144</td>
<td>-0.323</td>
<td>-0.527</td>
<td>-0.232</td>
<td>-0.110</td>
<td>0.232</td>
</tr>
<tr>
<td>heatLowest</td>
<td>0.233</td>
<td>-0.372</td>
<td>-0.428</td>
<td>-0.228</td>
<td>-0.127</td>
<td>0.190</td>
</tr>
<tr>
<td>rainDuration</td>
<td>0.177</td>
<td>0.307</td>
<td>-0.337</td>
<td>0.484</td>
<td>-0.028</td>
<td>0.017</td>
</tr>
<tr>
<td>rainTotal</td>
<td>0.159</td>
<td>0.252</td>
<td>-0.330</td>
<td>0.554</td>
<td>-0.027</td>
<td>0.187</td>
</tr>
</tbody>
</table>

Continued on next page
We know that every principal component is a linear combination of the original variables. From the table we see that the first seven principal components actually are associated with all the original variables. Besides the parameter of the linear combination is not always the same for different data sets, for which we will show with the data set from building 2 as an example.

As what we have done to building 1, Figure 4.23 has shown the variance associated with the principal components. Figure 4.24 shows the cumulative proportion of explained variance by using the proportion of variance of each principal component.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Comp.1</th>
<th>Comp.2</th>
<th>Comp.3</th>
<th>Comp.4</th>
<th>Comp.5</th>
<th>Comp.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunDuration</td>
<td>0.030</td>
<td>-0.517</td>
<td>0.202</td>
<td>0.432</td>
<td>-0.0155</td>
<td>-0.152</td>
</tr>
<tr>
<td>windDirection</td>
<td>0.141</td>
<td>-0.064</td>
<td>-0.171</td>
<td>-0.044</td>
<td>0.962</td>
<td>-0.133</td>
</tr>
<tr>
<td>windGush</td>
<td>0.512</td>
<td>0.055</td>
<td>0.209</td>
<td>-0.041</td>
<td>0.007</td>
<td>0.109</td>
</tr>
<tr>
<td>windSpeed</td>
<td>0.495</td>
<td>0.055</td>
<td>0.259</td>
<td>-0.075</td>
<td>-0.028</td>
<td>0.119</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variables</th>
<th>Comp.7</th>
<th>Comp.8</th>
<th>Comp.9</th>
<th>Comp.10</th>
<th>Comp.11</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunRadiation</td>
<td>0.050</td>
<td>-0.712</td>
<td>-0.084</td>
<td>0.006</td>
<td>0.008</td>
</tr>
<tr>
<td>windAverage</td>
<td>0.012</td>
<td>0.018</td>
<td>-0.018</td>
<td>0.114</td>
<td>0.805</td>
</tr>
<tr>
<td>bar</td>
<td>0.121</td>
<td>-0.004</td>
<td>0.0315</td>
<td>0.005</td>
<td>-0.008</td>
</tr>
<tr>
<td>heatDewpoint</td>
<td>0.047</td>
<td>0.105</td>
<td>-0.677</td>
<td>0.080</td>
<td>0.015</td>
</tr>
<tr>
<td>heatLowest</td>
<td>-0.006</td>
<td>0.012</td>
<td>0.714</td>
<td>-0.101</td>
<td>-0.002</td>
</tr>
<tr>
<td>rainDuration</td>
<td>0.722</td>
<td>0.018</td>
<td>0.047</td>
<td>-0.029</td>
<td>0.002</td>
</tr>
<tr>
<td>rainTotal</td>
<td>-0.677</td>
<td>-0.024</td>
<td>-0.004</td>
<td>-0.025</td>
<td>0.002</td>
</tr>
<tr>
<td>sunDuration</td>
<td>0.003</td>
<td>0.693</td>
<td>0.005</td>
<td>-0.004</td>
<td>-0.001</td>
</tr>
<tr>
<td>windDirection</td>
<td>-0.019</td>
<td>0.015</td>
<td>0.006</td>
<td>-0.030</td>
<td>0.014</td>
</tr>
<tr>
<td>windGush</td>
<td>-0.001</td>
<td>0.011</td>
<td>0.037</td>
<td>0.652</td>
<td>-0.501</td>
</tr>
<tr>
<td>windSpeed</td>
<td>0.004</td>
<td>0.012</td>
<td>-0.143</td>
<td>-0.737</td>
<td>-0.317</td>
</tr>
</tbody>
</table>
From the figure we can see that in this data set, it is still the first seven principal components reaching about 95% of the total variance. Then we have the composition of the principal components shown in Table 4.12. We still only keep the first three digits after the decimal. From the table we find that although this number of the principal components is the same as the previous data set, the results of the parameters in the linear combination have changed.

Table 4.12: Composition of principal components for building 2

<table>
<thead>
<tr>
<th>Variables</th>
<th>Comp.1</th>
<th>Comp.2</th>
<th>Comp.3</th>
<th>Comp.4</th>
<th>Comp.5</th>
<th>Comp.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunRadiation</td>
<td>0.204</td>
<td>-0.423</td>
<td>0.338</td>
<td>-0.353</td>
<td>0.003</td>
<td>0.123</td>
</tr>
<tr>
<td>windAverage</td>
<td>0.505</td>
<td>-0.103</td>
<td>-0.201</td>
<td>0.148</td>
<td>-0.011</td>
<td>-0.117</td>
</tr>
<tr>
<td>bar</td>
<td>-0.235</td>
<td>-0.305</td>
<td>-0.084</td>
<td>-0.173</td>
<td>0.225</td>
<td>-0.872</td>
</tr>
<tr>
<td>heatDewpoint</td>
<td>0.113</td>
<td>0.292</td>
<td>0.576</td>
<td>0.101</td>
<td>-0.113</td>
<td>-0.230</td>
</tr>
<tr>
<td>heatLowest</td>
<td>0.217</td>
<td>0.195</td>
<td>0.559</td>
<td>0.145</td>
<td>-0.128</td>
<td>-0.245</td>
</tr>
<tr>
<td>rainDuration</td>
<td>0.171</td>
<td>0.427</td>
<td>-0.163</td>
<td>-0.480</td>
<td>-0.007</td>
<td>-0.036</td>
</tr>
<tr>
<td>rainTotal</td>
<td>0.145</td>
<td>0.388</td>
<td>-0.114</td>
<td>-0.582</td>
<td>-0.024</td>
<td>-0.103</td>
</tr>
<tr>
<td>sunDuration</td>
<td>0.128</td>
<td>-0.476</td>
<td>0.254</td>
<td>-0.432</td>
<td>-0.027</td>
<td>0.199</td>
</tr>
<tr>
<td>windDirection</td>
<td>0.103</td>
<td>0.132</td>
<td>0.164</td>
<td>0.043</td>
<td>0.958</td>
<td>0.147</td>
</tr>
<tr>
<td>windGush</td>
<td>0.515</td>
<td>-0.065</td>
<td>-0.154</td>
<td>0.108</td>
<td>0.031</td>
<td>-0.109</td>
</tr>
<tr>
<td>windSpeed</td>
<td>0.497</td>
<td>-0.114</td>
<td>-0.199</td>
<td>0.152</td>
<td>-0.017</td>
<td>-0.111</td>
</tr>
</tbody>
</table>

Concluding from the results, we have shown that the results can be totally different due to the differences from data set. Although the structure and values could be similar from data sets, the principal components could be totally different. What is more, although the dimension of principal components has been reduced compared to original available variables, the principal components may be the linear combination involving all the original variables, which is not what we expected. Thus it is suggested to use another method than PCA to research on the external weather information.

![Figure from the document]
Chapter 5

Results, Conclusions, Recommendations and Further Research

In this chapter, the analysis and results in relation to the research questions will be presented. Then based on the conclusions, some recommendations for ENGIE and the final part is the direction for further research are given.

5.1 Results and Conclusions

There is a seasonal effect when labeling the data sets using specific time segments.

By looking at the way the company uses in the real data set, the default time segment for “day” is 07:00:00 – 19:00:00. It turns out to be not accurate when applied in the simulated data set. The labels may be influenced by the seasons, which is reflected by the changes to the time zone here.

There is a wrong label to one data point due to the time shifting in real meter data.

By the conclusion of changing the winter/summer time in simulated data, the winter/summer time changing in the real meter data should be checked. In the real meter data, the winter/summer time changes could be checked by the calendar year\(^1\). ENGIE has already dealt with this issue, however there is an error with the label. For every accessible data set so far, the label for the hour from summer to winter should exactly be “nightWeekend”, however every label here is labeled as “night”. This should be corrected.

Adding the slope of temperatures into model does not make much difference.

For MARS analysis a further research has added the slope of temperature due to the effect of heat load as an extra independent variable. However, it does not make much difference compared to the original model with only current temperatures.

\(^1\)http://hemel.waarnemen.com/aarde/zomertijd.html
CHAPTER 5. RESULTS, CONCLUSIONS, RECOMMENDATIONS AND FURTHER RESEARCH

Using the one-hour-lag temperatures does not make much difference.

For MARS analysis, one-hour-lag temperature is considered as the independent variable as a new model. The results show that using this variable will give a smaller tipping point in most cases.

Automatic clustering turns out to have fewer clusters.

In the simulated data, an automatic clustering method is applied to find the labels. No matter for what sources of energy we use, the results show fewer clusters than four clusters, which is the current number of clusters we have.

5.2 Recommendations

Do not use fixed specific time segments as labels to all buildings.

Although in the real meter data, we have fixed the winter/summer time changes, it is still not accurate to use fixed specific time segments as labels to all buildings. This is because that the time segments may vary from building to building. Besides, the time segments may also vary from time to time in different seasons even after we modify the time zone.

To do the MARS analysis based on separate source of energy.

Although the interaction between gas and electricity should be taken into consideration, the results of MARS analysis based on different sources of energy suggest that we should do the MARS analysis on separate source of energy. Otherwise the tipping points from the results of MARS analysis based on the combination of sources of energy are away from the reality.

To choose MARS model within each cluster or with labels to get separate models.

Tables 4.1 and 4.2 have summarized the tipping points from models with different independent and dependent variables. It is suggested to consider the model within separate labels or using labels as an extra variable to get separate models in one step. The results of this two models differ from each other to some degree. We need to choose one close to reality after detailed research.

To use a scientific clustering method to label the data set.

It is suggested to use an automatic clustering method to label the data sets instead of manual labeling. `pamk()` in the package `fpc` and `pam()` in the package `cluster` are applicable.

Do not use fixed number of clusters to all energy sources.

When doing the automatic clustering based on different sources of energy, the optimal numbers of clusters are different. Thus it is not accurate to claim a fixed number of
clusters to all sources of energy as well as the combination of sources of energy. However
the optimal number of clusters from \textit{fpc} may be not the best related to reality, the exact
number of clusters should be chosen after a detailed analysis.

**To do automatic clustering on the combination of sources of energy.**

Even it requires the same number of clusters in the automatic clustering based on different
sources of energy, the results are different. Although MARS suggests to do the analysis
based on separate source of energy, the automatic clustering suggests to consider the
combination of sources of energy.

### 5.3 Further Research

**It requires more research on the heat load and to figure out the variables to
represent heat load in the model.**

Since adding the slope of temperatures or using the one-hour-lag temperatures does not
make much difference to the results, it should go to a further research on the effect of
heat load.

**To use another method than PCA to research on the external weather information.**

Two weather data set from different clients have been applied with PCA and the output
of principal components are totally different. More research is needed into the external
weather information and then consider which variables can be chosen into the MARS
model.


BIBLIOGRAPHY


56 Advanced Regression Methods for Energy Signatures of Buildings