Towards smart energy city

analysis of the energy usage in the services sector in the city of Eindhoven

Karimi, I.

Award date:
2012

Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

Towards Smart Energy City; Analysis of the Energy usage in services sector in the city of Eindhoven

By: Iman Karimi
Student Number: 0755771

In partial fulfillment of the requirements for the degree of
Master of Science in Construction Management and Engineering

Supervisors:
Prof.dr.ir. B. de Vries
Dr. Qi Han
Dr.ir. Erik Blokhuis

Date of final presentation:
3 July 2012

3-7-2012

Table of contents

Preface

This report is the outcome of the final graduation project of Construction Management and Engineering (CME) program in Eindhoven University of Technology. The subject of the project is related to the main theme of the current year of the CME group which is smart cities.

Following the CME program was a good opportunity for me to add values to my educational background. By working on this final project, I have also gained knowledge about various fields. The energy sector itself is an interesting area of research. Cluster analysis is the other field which I was working on it in this project. Furthermore, using the statistical analysis has provided me with the understandings about the useful software in this field.

I should respectfully thank my supervisors Prof.dr.ir. B. de Vries, Dr. Qi Han and Dr.ir. Erik Blokhuis. Without their supervisions, their valuable advices and their kindly supports I couldn't be able to finish this project.

I hope this research would be an interesting subject of study for the future students and provides a good insight about the considered issue for the readers.

Iman Karimi,

Eindhoven,
July 2012

Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

Towards Smart Energy City; Analysis of the Energy usage in services sector in the city of Eindhoven

By: Iman Karimi
Student Number: 0755771

In partial fulfillment of the requirements for the degree of
Master of Science in Construction Management and Engineering

Supervisors:
Prof.dr.ir. B. de Vries
Dr. Qi Han
Dr.ir. Erik Blokhuis

Date of final presentation:
3 July 2012

Table of contents

Preface

This report is the outcome of the final graduation project of Construction Management and Engineering (CME) program in Eindhoven University of Technology. The subject of the project is related to the main theme of the current year of the CME group which is smart cities.

Following the CME program was a good opportunity for me to add values to my educational background. By working on this final project, I have also gained knowledge about various fields. The energy sector itself is an interesting area of research. Cluster analysis is the other field which I was working on it in this project. Furthermore, using the statistical analysis has provided me with the understandings about the useful software in this field.

I should respectfully thank my supervisors Prof.dr.ir. B. de Vries, Dr. Qi Han and Dr.ir. Erik Blokhuis. Without their supervisions, their valuable advices and their kindly supports I couldn't be able to finish this project.

I hope this research would be an interesting subject of study for the future students and provides a good insight about the considered issue for the readers.

Iman Karimi,

Eindhoven,
July 2012

Summary

This research is about the energy usage in the services sector in the city of Eindhoven. The research after this issue is not satisfactory yet. In this research, through using different sources a dataset is created. This dataset contains the annual energy usage of different buildings separately for gas energy and electric energy. The energy usage attributes such as building characteristics and the user's size have been included in this dataset. Also each building belongs to one specifics sub-sector of the services sector on the basis of SBI 93 code.

After building the dataset two different kinds of analysis were performed on it. Firstly, by using the cluster analysis the buildings have been grouped in different clusters. In clustering of the buildings different variables have been considered for various types of clustering. Also two different methods of clustering have been used namely k-means clustering and two-step clustering method. The results of this step are various clusters of the buildings.

In addition, the multiple linear regression method has been used for statistical analysis on our dataset. The reason for using this method is that we would like to have the prediction equations for the energy usage of the buildings and find out which variables are more important in the prediction of the energy usage.

To reach these goals, the analysis is performed for each sub-sector and also then the buildings divided into different clusters. For each cluster an energy usage equation has been derived. The results of this part are the important energy usage predictors and the equations for energy usage, separately for the electric energy, gas energy and total energy.

Chapter 1: Introduction

The energy sector faces numerous problems, e.g. climate change, environmental and human accidents, reliability of energy supply and oil dependency. It is therefore time to launch a fundamental change with respect to our energy supply. The drivers for such a change originate in broad societal ambitions, and materialize in policy that is mostly formulated at national and cross-border levels. For instance, the European Commission has formulated major objectives for future energy systems, e.g. to reduce carbon emissions by 20%, to increase the share of renewable energies by 20%, and to increase the energy efficiency by 20% before 2020. This fundamental change implies transitions towards new sustainable energy systems, in which energy reduction ambitions play a major role.

On the other hand, city councils face a major task implementing these ambitions. However, they have limited power to steer developments that lead to energy reduction and energy system adjustments. For instance, development plans for new houses are very limited in scale and renovations plans can only be executed subsequently and will only yield a substantial effect in the long run. Energy network operators play a very special role in the development of energy reduction plans and system adjustments. Their main task is to provide and maintain enough network capacity such that the energy demand of all citizens (inhabitants of dwellings, offices, industrial buildings, etc.) can be satisfied safely. Energy demand (natural gas and electricity) has shown a gradual increase over the last decades. Recently, new energy sources are under development (WKO, thermal energy, solar cells, etc.) that will change energy distribution through the network significantly.

In the Netherlands, the national government aims to achieve ambitious climate targets to become one of the most efficient countries in the world in terms of energy. For example, the municipality of Eindhoven wants to become energy-neutral in 2035-2045. Indeed, the importance of having the insight into the user energy profiles is clear enough which could help authorities to lower the redundancy of the networks and use energy networks more optimal. This will lead to more sustainable energy networks and energy conservation.

One important sector in terms of energy usage is the services sector. Within the province of Noord-Brabant this sector uses around 20% of the total energy usage. In the table 1 the energy usage of different sectors in this province has been shown [1].

Eindgebruikers	Energie gebruik 2040 [PJ]	Extra potentiele besparing [%]	Extra potentiele besparing [PJ]
Huishoudens	63 - 98	40 - 60%	25 - 59
Diensten	73 - 108	40 - 60%	30 - 65
Landbouw	13 - 27	> 50%	6 - 14
Industrie	84 - 146	10 - 20%	8 - 29
Transport	73 - 126	20 - 65%	12 - 82
Totaal	***375 - 625***	***160 - 255%***	***83 - 249***

Table 1: Primary yearly energy use according to end users and potential energy savings in Noord_Brabant (Source: SOLET report, 2008)

As you can see in this table, the potential for energy saving in the sector of services is 40% to 60% or 30-65 Peta Joules (PJ). This provides the municipality with great opportunity for making policies to save energy. Indeed, it is obvious that study the energy usage in the services sector is really important. This research has focused specifically on this issue.

When we talk about the energy saving potentials in different sectors the building type and the activities performed in the building that prescribe the energy demand are important [2]. On the basis of this fact, the building type could have great effects on the energy usage. Also, the other possible variables which are vital in energy saving policies should be studied carefully.

In this research we are going to find out which variables play important roles in the energy usage in the services sector. To reach this goal an appropriate dataset about the energy usage of some buildings in the services sector in the city of Eindhoven is going to be built. The information of this dataset is extracted from different available sources. After creation of the dataset, the available variables are going to be tested through the statistics analysis and their importance in the energy usage could be found out. Also, the cluster analysis will be done on the basis of different variables. This will help the governmental agencies, designers and energy management engineers to have clear view from different perspectives about the effective variables in energy usage and propose the policies for energy reduction measures.

1.1. Problem description

On average, residential energy use covers a relative large part of the total energy use in cities (approximately 40%). Indeed, modeling the energy use of households by considering the occupant's behavior is really important. In this regard we could find many articles which have studied this issue.

However, the energy use of commercial and non-commercial sector (for example the services sector) accounts for almost the same amount of city energy. The research after the energy usage in the services sectors remains a shallow area of the research. Through the literature study we could not find a suitable article which describes the energy use patterns in this sector clearly. Also, the data about the energy usage in details is not enough and satisfactory in the services sector. Indeed, studying the energy usage variables in the services sector could be a vital initial step for beginning of the research in this area.

This research aims to study the energy usage in the services sectors by considering the important variables for the energy usage prediction and also clustering the different building types on the basis of these variables.

1.2. Research questions

The major aspect that this research aims to gain insight in is the energy use dynamics of services sector. The reason for this is the presumed large influence of characteristics of individual buildings and users on energy usage in this sector. By understanding the dynamics of energy use, these insights can be used to optimize several aspects of existing energy systems in a city.

The main research questions can be written down as follow:

- Is it possible to make a dataset for the energy usage and related variables in the services sector?
- What are the most important variables (predictors) for the energy usage of the buildings in the services sector?
- Could we define some equations for the energy usage of buildings in the services sector?
- How can we separate the buildings in the services sector to different clusters?

These main questions lead to some sub-questions as below:

- What variables could we find for energy usage in the services sector?
- Which methods should we use for the prediction of the energy in this sector?
- Which variables are suitable for performing cluster analysis for the buildings in this sector?

In the future sections of this report, we will answer to the above mentioned questions clearly.

1.3. Aims and objectives

The design of a 'smart energy city', in which energy use is minimized, usage peaks are flattened, and energy exchange is optimized, requires fundamental knowledge about the energy demand dynamics, energy use reduction options, and possible energy generation. Only when we are able to predict the energy demand and generation of any type of energy user at any given time with accuracy, we can redesign dwellings, offices, districts and their energy networks, together with the policies of the 'smart energy city'.

Some studies have tried to define the measures for the energy usage reduction options in the services sector [2]. However these studies are more about the individual characteristics of the buildings and variables. Studying the energy usage in the entire district for the services sector would be also very interesting and important. It is beneficial if we find out by the statistical analysis that which variables are more important in prediction of the energy usage of a building in the future. Also, clustering the buildings provides different perspectives for the authorities.

Therefore, the aim of this research is firstly generating a dataset which includes the energy usage amount of the buildings and also related variables of the buildings such as floor area or façade type. Then, by using the statistical analysis the important predictors of the energy usage could be recognized. Finally, the clustering of the buildings in this dataset will provide good insight into the buildings in the services sector.

1.4. State-of-the-art

The energy use of industrial companies has been researched extensively. Especially, the contribution of Ang (1995) [3] and Ang and Lee (1996) [4] are interesting; they focus on

decomposition of industrial energy composition, aiming to study the impacts of structural change and changes in sectoral energy efficiencies.

Also, efforts have been made in modeling the energy usage of the non residential sectors. Gaglia et al. (2007) [5] elaborated the approach used to determine the potential energy conservation in the Hellenic non residential building stock. Choudhary (2012) [6] presented a Bayesian approach for developing city-scale energy models of the built environment and demonstrates its application to non-domestic buildings in Greater London. Liu (2006) [7] had the study on the decomposition of the industry energy consumption. Pan et al (2008) [8] developed energy simulation models with Energy Plus for two office buildings in a R&D center in Shanghai, China to evaluate the energy cost savings of green building design options compared with the baseline building.

However, by looking through the available articles in this field, research after services energy use is often executed on national scale, not discussing individual cases. Also there is not any dataset available which includes information about the energy usage and the important variables in this regard. Indeed, creating a dataset which has the information about the companies in the services sector and buildings characteristics and employees numbers would be very beneficial for also future studies in this field.

The focus of this research is on the industrial areas of the city of Eindhoven. After creating the dataset through the statistical analysis, we could find the most important variables which are effective in energy usage. This issue has not been considered in the district level or city level previously and by considering the energy usage of the individual buildings.

Furthermore, cluster analysis of the buildings in the services sector in the city of Eindhoven is a new subject of study. The clustering has been done on the dwellings in the city of Eindhoven previously [9]. However, there is not any clustering for the services sector's buildings.

1.5. Previous studies

Through the literature study we could see the research about the energy issues in the industrial or services sector is not comparable with the research about the residential sectors. Most of the available articles are about the energy usage and reduction policies in the domestic sectors. Also the scale of the studies is mostly the national level.

One useful article about the energy usage in the services sector is ICARUS - 4 sector study for services sector in Netherlands [2]. In this study a vast research about the energy usage in the services sector has been done. Some energy saving measures have been offered in this report. Also the services sector has been divided to some sub-sectors. Table 2 shows the breakdown of the energy use in 1995 on the basis of these sub sectors.

Icarus sub sector	SBI-code	Fuel demand (PJ)	Electricity demand (PJ)	Relative fuel demand (%)	Relative electricity demand (%)
Commercial offices	65-74	12.3	9.4	8%	16%
Public offices	75	13.3	8.8	9%	15%
Non-office based services (including sports & culture)	90-99	17.6	6.9	12%	11%
Retail	50-52	44.5	20.1	30%	33%
Catering	55	17.1	6.6	11%	11%
Education	80	14.7	2.7	10%	4%
Intramural health	8511,8531.1-4	14.7	3.9	10%	6%
Other health (incl homes for elderly)	85 (ex. 8511,8531.1-4	16.2	2	11%	3%
Total		**150.4**	**60.4**		

Table 2: Breakdown of the energy use in services sector in 1995

Also in this report, the total energy has been divided to electricity and fuel. The potential of the savings of the proposed energy saving measures has been calculated. In addition, the implementation costs of these measures in different sub sectors have been estimated.

1.6. Research methods

This research consists of three blocks. After the literature study, firstly a dataset needs to be built for the energy usage in the services sector. Through the literature study one can find out that the available data in this sector is not satisfactory and enough for the analysis purposes. Many important issues for the energy usage of the buildings and important energy usage variables are missing in the services sector. Hence, the basic step to perform this research is creating the database by reaching as much as possible information. To achieve this goal, different sources should be investigated and different organizations should be asked for the information.

Furthermore, we would like to use the cluster analysis. This method is useful for grouping the buildings in to specific groups on the basis of different variables. We will use the variables which we have investigated before in the dataset in the first step. The energy usage should be the main focus of clustering and then the other variables will provide extra measures for clustering the buildings.

The next step of the research is assessing the important variables of the energy usage in the buildings of the services sector. These variables could be building related or the user related. Then, we could make some equations for the energy usage of the buildings. For this aim the statistical analysis is going to be used. For the statistical analysis we can use the linear regression model which is the most usable method for the prediction of the dependent variables (here the energy usage) by using the independent variables (here for example building characteristics). For the prediction equations we are going to separate the buildings into different clusters (groups). This issue will be done by using some effective variables which can lead to desired outcomes and minimized errors.

Figure 1 shows the integration of the different steps of this research which we have described above.

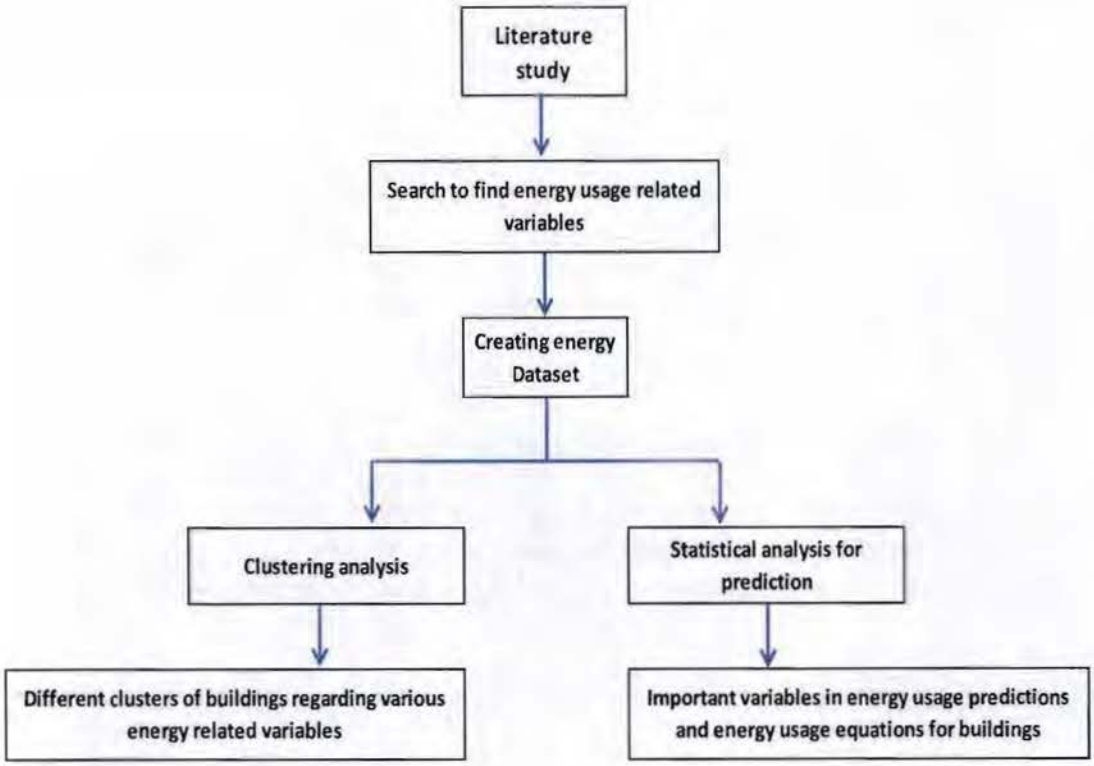

Figure 1: research design

1.7. Research relevance

1.7.1. Relevance TU/e

This research completely fits with the theme of the Construction Management and Engineering (CME) master program of Eindhoven University of Technology (TU/e) for the final graduation projects of the students which is under the subject of smart cities. Also, the aim for doing research in the field of sustainability, which is one aspect of this research, is set in the TU/e in different master programs.

1.7.2. Relevance MKB

The energy subject is very important for the province of Noord-Brabant. The aim of the municipality is to turn to the energy neutral city up to 2040. This research is relevant with the concept of energy reduction usage which is vital for reaching the above mentioned goal.

1.8. Expected results

This research will lead to having insight into the important variables in the energy usage of the buildings in the services sector. The final results of the current research can be mentioned as below:

- A dataset with the annual energy usage of the buildings, the addresses of the companies and buildings, building characteristics and number of employees of the

related company in different sub sectors of the services sector for the city of Eindhoven.

- Important variables in the energy usage prediction of the different buildings.
- Prediction equations for different clusters of the buildings in the services sector on the basis of the available variables.
- Different clusters of building regarding different variables of the energy usage on the services sector for the city of Eindhoven.

1.9. Structure of the thesis

The report starts with the "introduction" chapter which describes the main features of the project and the background about the subject which has gained through the literature study.

After that, the chapters are on the basis of the different steps in the implementation of the thesis. The "creating the dataset" chapter provides the reader with information about the procedure of making the dataset and the resources which have been used in this dataset.

The next chapter is "clustering analysis". This chapter offers the theory of the cluster analysis and also the case study of the clustering for the buildings in the services sector in the city of Eindhoven.

Then the "energy usage prediction" chapter contains the theoretical aspects of the prediction analysis for the energy usage of the buildings and also the case study about the services sector in the city of Eindhoven.

The next chapter which is "conclusions" includes the conclusions of the current research with recommendations for the future studies. Then the "references" chapter is written down. Finally the "appendixes" chapter consists of the related tables and information about the analysis in different sections.

Chapter 2: Creating the Dataset

Through the intensive research in the available data for the energy usage in the services sector in the Netherlands and in the city of Eindhoven, we could not find a comprehensive dataset. The available information does not include the variables such as building characteristics. Therefore, for starting the statistical analysis for the prediction or for clustering analysis, creating a dataset which includes the energy usage of specific buildings in a year and the related variables (such as the building characteristics or number of employees) is vital.

The dataset which we are looking to build should gain the information from different sources. Some websites, previous studies or even the Google Earth software are among these sources. To clarify the point, we are going to explain these sources and the related data which is extracted from them.

2.1. Dataset scope

As it is described in the introduction chapter, the main focus of this research is about the energy usage prediction of the buildings in the services sector and also clusters analysis of these buildings for the city of Eindhoven. So the main focus of the dataset is the buildings in the services sector in the city of Eindhoven. The services sector itself is divided to some sub sectors which can be seen in table below with the related SBI 93 code [2].

Services sub sectors	SBI 93 Code
Retail	20-52
Catering	55
Commercial offices	65-74
Public offices	75
Education	80
Health	85
Non-office based services (including sports & culture)	90-99

Table 3: Services sub sectors and SBI 93 codes for the dataset

We would like to use as much as possible energy related characteristics in our database. The energy usage for the buildings is the annual amount for a building.

2.2. Data sources

2.2.1. Endinet Database

The first data source that we have used is the Endinet dataset. The Endinet is a company which is active in providing energy in the area of Noord-Brabant [10]. The dataset includes the yearly energy usage of the companies of the city of Eindhoven with the complete addresses. This energy usage is divided to two parts; the electric energy consumption and gas energy consumption. This dataset has been created in year 2008. From this database we

could have the energy usage (in MJ) for each building which is divided to electric energy usage and gas energy usage.

2.2.2. Werkgelegenheid Dataset

The Werkgelegenheid dataset is used for extracting the information about the name of the companies, the work field and description of the work area, the SBI or CBS code number, the number of employees (in certain interval) and the addresses of the companies. This dataset has been created in 1 April 2008 by Stichting LISA te Enschede.

Hence, from the combination of this dataset with the Endinet dataset, we could have the yearly energy usage amount for each company and also we know the number of employees and the SBI 93 sector code of the company.

2.2.3. Leegstand Bedrijventerreinen Eindhoven Dataset

This dataset is created by A.P.J. Avontuur et al. for the purpose of vacancy checking in the industrial areas in the city of Eindhoven. The useful data in this dataset is the addresses of the companies with their floor area in square meter and also the number of the floors of the building. The dataset has been created in the Eindhoven University of Technology in spring of 2011.

Unfortunately, there was not possible to combine the above datasets together easily. So it was necessary to search for the buildings one by one in the datasets and try to match the desired data manually which it took too much time.

2.2.4. BAG-viewer website

This website is created by Ministerie van I&M and Kadaster [11]. Through this website you can type in the address of the building that you are interested in and you can find some information about that building such as floor area (Oppervlakte in Dutch), year of construction (Bouwjaar in Dutch) and .... From this website the construction year of the buildings is taken. Also, the floor areas of the buildings from the previous dataset were controlled again. The available information in this website has been updated in March 2012.

2.2.5. Google Earth

The last tool for getting the intended data for the final dataset was the Google Earth. Through this software, by entering the address of the building, we could reach some information about the building such as the outside view of the building and type of the facade, the surrounding situation, the location of the building, the height of the building, and approximately estimate the glass percentage of the outside view of the building. Hence, it is possible to assign this information to the dataset.

2.3. Energy related variables

The important variables that we have gained from the above mentioned sources are listed below with short description about them:

1- Building address, company name and sector:
Full address of the company and building (street, house number, postal code) can be used from the Werkgelegenheid Dataset. Also the name of the company, sector of the activities and the description of the working field of the company are included in the dataset.

2- Building age:
From the BAG-viewer website we can find the year of the construction of the buildings. Then the building age can be calculated. The energy data is from year 2008. So the building age should also be calculated up to this year.

3- Height of the building:
The height of the building can be taken from the Google earth software in feet. Then we can convert it to centimeter and use in our dataset.

4- Building location situations:
As one of the important variables for the energy usage, we could include the building location situation in our dataset. Three different types for the building locations situation have been defined in this research; between building location, single building location and side building location. Figure 2 shows these location types clearly. This information can be read from the BAG-viewer website.

Figure 2: Location situations for the buildings

5- Building surroundings situations:
Surroundings of the building also could be important in the energy usage of the building. There are three different situations for the surrounding in current research. The building might be located in the building blocks, in an open area or next to the river. Figures 3 to 5 visualize the different surrounding situations for the buildings.

Figure 3: Building block surrounding

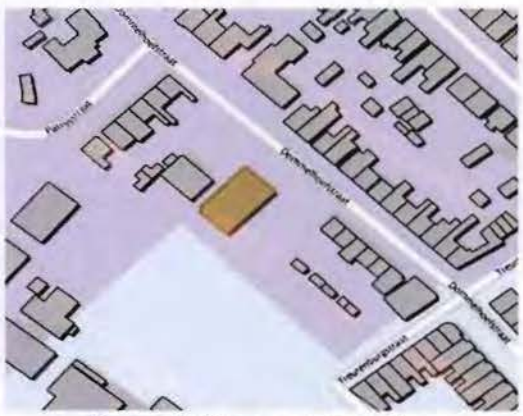

Figure 4: Open surrounding

Figure 5: River surrounding

6- Glass percentage:

The glass percentage of the outside interface of the building can be estimated approximately from the view of the building in the Google earth software. We use the percentage for showing this variable.

7- Façade type:

The façade type of the building can be seen from the Google earth software. This could be brick façade, curtain wall façade or mixed façade.

8- Floor number:
Number of the floors of the building can be taken from two sources; Google earth software or Leegstand Bedrijventerreinen Eindhoven Dataset.

9- Total area:
Floor area of the buildings (in square meter) can be taken from the Leegstand Bedrijventerreinen Eindhoven Dataset.

10- Electric energy usage:
Annual electric energy usage of the building (in MJ) can be taken from the Endinet dataset.

11- Gas energy usage:
Annual gas energy usage of the building (in MJ) can be taken from the Endinet dataset.

12- Total energy usage:
Annual total energy usage of the building (in MJ) can be derived from summation of gas energy usage and electric energy usage for each building.

13- Number of employees in the company:
Number of the employees of the company for each specific building can be taken from the Werkgelegenheid Dataset. In this dataset different groups for the employee's numbers have been defined which can be seen in table 4.

Group of employee	Number of employees
01	0 - 0 personen
02	1 - 1 persoon
03	2 - 4 personen
04	5 - 9 personen
05	10 - 19 personen
06	20 - 49 personen
07	50 - 99 personen
08	100 - 199 personen
09	200 - 499 personen
10	500 - 799 personen
11	800 - 999 personen
12	1000 en meer personen

Table 4: Number of employees for each group of employees in the dataset

2.4. Energy usage - Variables dataset

By using the above mentioned variables, the dataset has been built. In this dataset 387 buildings (companies) have been included in the city of Eindhoven. However, for all of these buildings all the variables are not available.

This dataset is the initial step for the future analysis. We are going to use the dataset for statistical analysis and cluster analysis in the next steps of our research. However, before

the analysis steps, the data preparation is essential. The next chapter will describe how we have done the data preparation on this dataset.

This dataset could be also the initial step for future studies about the energy usage in the services sector of the Eindhoven. The different sub-sectors have been divided clearly in the dataset and it provides the authorities and experts with a beneficial tool for assessing the energy usage and important variables.

2.5. Data preparation

The first step after creating the dataset is making the data ready for using it in the statistical analysis or clustering analysis. This step is called data preparation. In this part the method of data preparation is going to be described.

2.5.1. Missing Data

In creating the dataset different sources have been used as it is described in the previous parts. However for some of the variables we couldn't find the data. Table 5 shows the number of missing data of the buildings for different variables of our dataset.

Variable	Number of missing data
Building age	10
Height	49
Location situations	1
Surrounding situations	1
Façade type	49
Glass percentage	49
Gas energy	11

Table 5: Number of missing data for each variable in the dataset

2.5.2. Outliers detection

The next step in data preparation is the outlier detection. The outlier is a data which is numerically distant from the rest of data in a dataset. The definition of Grubbs [12] could help to clarify the point more:

"An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs."

For our dataset it is important to find the outliers and modify the data on the basis of the available rules. The outlier labeling rule [13 & 14] is going to be used to detect outliers in data that roughly correspond to a normal distribution. In this regard we can define some boundaries for the outliers. The software which is used for the statistical analysis and also clusters analysis is the SPSS which is the most useable software in the scientific purposes. The formulas which can be used for outlier detection are:

$$\text{Upper boundary} = Q3 + (1.5 * (Q3 - Q1)) \quad (1)$$
$$\text{Lower boundary} = Q1 - (1.5 * (Q3 - Q1)) \quad (2)$$

The different values for Q can be taken from the SPSS software.

For our dataset, six variables have outliers in their data. For these variables the outlier detections surveys have been done. The "exploring outlier" command in the software is used and then with the information from the outcomes the outliers have been modified in the dataset. Table 6 shows the variables, the upper and lower boundaries and also the number of modified cases.

Variable	Lower boundary	Upper boundary	Number of modified cases
Building Age	0	71	3
Height	3.21	9.29	16
Glass Percentage	0	80.00	4
Total Area (M2)	0	5,508.80	34
Electric Energy (KWH)	0	286,946.45	48
Gas Energy (M3)	0	37,902.25	44

Table 6: Outliers detection information

It is worthwhile to mention that the value of the outlier is replaced in the dataset with the upper or lower boundary. By this method the data can be saved and be used later on in the analysis. The full tables of the outlier detection are included in the appendixes (part A. data preparation: detecting outliers).

2.5.3. Dummy variables

Some variables in our dataset are categorical or nominal variables. These variables are location situations, surrounding situations and façade types. For using these variables in the statistical analysis with SPSS, we should convert them to the dummy variables. The dummy variables can be used in multiple regression analysis.

A dummy variable is a dichotomous variable which has been coded to represent a variable with a higher level of measurement. A dummy variable can have 0 or 1 value. These variables are qualitative variables in regression analysis.

In our dataset some numbers have been considered for each of the above mentioned categorical variables. Table 7 shows the categorical variables with their assigned qualitative numbers.

As you can see we have three levels for these variables. In general, a categorical variable with k levels will be transformed into k-1 variables each with two levels [15]. In our case k is 3 so we need 2 dummy variables for each categorical variable.

Assigned number in dataset	Categorical variables		
	Location situation	Surrounding situation	Façade type
1	Between buildings	Building block	Curtain wall
2	Single	Open	Brick
3	Side	River	Mixed

Table 7: Categorical variables in the dataset

For example, we can consider the location situation. The dummy coded variables for the location situation are shown in table 8. These dummy coded variables are going to be used in the analysis instead of the categorical variables. The same procedures have been done for surrounding situation and façade type.

Categorical variable	Assigned number to location situation	Dummy Coded Variables	
		Building block	Single
Between buildings	1	1	0
Single	2	0	1
Side	3	0	0

Table 8: Dummy coding the location situation types

Indeed, the dummy coded variables of between building location, single location, building block surrounding, open surrounding, curtain wall façade type and brick façade type are going to be used in the statistical analysis and clustering analysis in future steps.

2.6. Conclusions

The main outcome of the current chapter is the dataset which contains the information of the buildings in the service sector of city of Eindhoven in terms of the energy usage and the important variables related to the energy. This dataset is built based upon different sources. In total, 387 buildings have been included in this dataset. However for some of them some information is missing.

After creation of the dataset the data preparation method has been used to make the data ready for using in the analysis in future steps. The missing data has been recognized and the outliers have been modified. Finally, some categorical variables have been converted to dummy coded variables. By this way, we can use these variables in the statistical analysis and clustering analysis in following chapters.

The created dataset is useful for the analysis in the service sector regarding the energy usage. This dataset can be completed more by using the future information about the buildings of the city of Eindhoven.

Chapter 3: Cluster analysis

In this chapter we are going to perform the cluster analysis on the created dataset in previous steps. For the buildings in the services sector in city of Eindhoven, different methods of clustering are going to be used. Firstly, a short theory about the clustering analysis and the used methods are included and then the case study for the Eindhoven is depicted. Finally, the conclusions of this chapter are written down.

3.1. Aims

In order to optimize energy aspects like network usage, excess energy sharing, renewable energy generation, and energy reduction, specific energy user clusters should be developed. If insight is gained in specific individual energy usage profiles, such clusters can be designed, possibly ignoring readily established clusters in cities like neighborhoods or city districts.

Cluster analysis is often used in two adjoining and related research streams, which can be applied to the topic of smart energy cities: market research and housing submarkets. These two streams deal with clusters on (a) individual users and (b) building related characteristics. In the current research the second option is considered.

Therefore, the aim of this chapter is cluster analysis of the buildings in the services sector in city of Eindhoven by considering the energy usage and related variables to the energy usage. Building characteristics and other variables are going to be used in this regard and different cluster analyses are going to be done on the entire dataset.

3.2. Cluster analysis theoretical underpinnings

Cluster analysis or clustering can be defined as the unsupervised classification of patterns (e.g. observations) into groups [16]. With the other words, cluster analysis is the task of assigning a set of objects into groups (called clusters) while the objects in the same cluster are more similar to each other (in some sense or another) than to those in other groups (clusters). Before the actual clustering can take place, some choices have to be made concerning the representation of the pattern and the clustering procedure.

Different clustering techniques are available. In general, two different types of clustering can be distinguished, hard and soft (fuzzy) clustering. Hard clustering divides every object into respectively a single hard cluster or more clusters for a certain degree. For example, partitional clustering is a hard clustering method. In research for housing submarkets "k-means clustering for non hierarchical clustering" and "Ward's method for hierarchical clustering" are mostly used.

A distinction can be made between hierarchical clustering and partitional clustering [16]. In a hierarchical clustering procedure nested series of partitions are produced. Either every pattern starts as a separate cluster and patterns are combined step by step until one cluster remains (agglomerative clustering), or all patterns are initially combined into one cluster and then separated into smaller clusters until all patterns represent individual clusters (divisive clustering). The optimal number of clusters is determined afterwards, using a dendrogram, which visualizes the variation within the clusters for different steps of the clustering procedure [17]. On the other hand, unlike hierarchical clustering, partitional

clustering leads to only one partition, depending on the number of clusters that is chosen in advance.

For more information about the theoretical aspects about the cluster analysis the references books or articles can be read. Here, we are not going into the theoretical details about the cluster analysis anymore. We would like to show the analysis results for the case study of services sector in Eindhoven in the next sections of this chapter.

3.3. Clustering methods

The clusters analyses for the created dataset have been done in two different methods. The first method is the famous "k-means method" which is a method from the Centroid Based Clustering (CBC). The other used method is "two-step cluster analysis method". At first short descriptions about the theoretical aspects of these two methods are going to be described. Then the results of clustering analyses on the database are presented.

3.3.1. K-means clustering method theoretical underpinnings

K-means clustering method is the most useable method of the CBC clustering. In general, CBC is non-hierarchical method which is based on an iterative process. In the k-means method, we assign cases to a fixed number of groups (clusters) whose characteristics are not yet known but are based on a set of specified variables [18]. Hence, the number of "k" mostly assume upfront.

The basic of clustering with k-means method is on initially the construction of initial cluster centers. Then cases are assigned to clusters based on distance from the cluster centers. By the iterative process we can update the locations of cluster centers based on the mean values of cases in each cluster. These steps are repeated until any reassignment of cases would make the clusters more internally variable or externally similar.

In many cases the number of iterative steps needs to be specified upfront. The outliers should be put outside of the dataset because outliers would be selected as initial cluster centers, resulting in clusters with only a few members. As it was mentioned in the data preparation section of the chapter 2, we have done the outlier analysis in the previous steps so we use the modified dataset here.

The principal component analysis (factor analysis) can be used also to make the number of variables less for using in this cluster analysis method. We will back to this point in the upcoming sections in this report.

3.3.2. Two-steps clustering method theoretical underpinnings

This method is a useful method for clustering the objects. Some of the main features of the two-step clustering methods by SPSS software are the ability to create clusters based on both categorical and continuous variables, automatic selection of the number of clusters and the ability to analyze large data files efficiently [18].

In the two-step clustering method a likelihood distance measure is used which assumes that variables in the cluster model are independent [18]. Also, each continuous variable is

assumed to have a normal (Gaussian) distribution and each categorical variable is assumed to have a multinomial distribution.

The two steps of the two-step cluster analysis procedure's algorithm can be summarized as follow [18]:

- **Step 1:** The procedure begins with the construction of a Cluster Features (CF) Tree. The tree begins by placing the first case at the root of the tree in a leaf node that contains variable information about that case. Each successive case is then added to an existing node or forms a new node, based upon its similarity to existing nodes and using the distance measure as the similarity criterion. A node that contains multiple cases contains a summary of variable information about those cases. Indeed, the CF tree provides a capsule summary of the data file [18].

- **Step 2:** The leaf nodes of the CF tree are then grouped using an agglomerative clustering algorithm. The agglomerative clustering can be used to produce a range of solutions. To determine which number of clusters is "best", each of these cluster solutions is compared using Schwarz's Bayesian Criterion (BIC) or the Akaike Information Criterion (AIC) as the clustering criterion [18].

3.4. Clustering with k-means method

In this method, as we mentioned before, first we are going to implement the principal component analysis for our dataset. Then by using the results of this method the cluster analysis is done. Also, another k-means clustering is done by using the variables themselves.

By using these two different kinds of the k-means clustering implementation, two perspectives can be achieved for the buildings and their energy usage with this method.

3.4.1. Principal Component Analysis (PCA) Theory

Principal component analysis (PCA) is used to reduce the number of the variables in the dataset for the statistical analysis. This method is one of the many methods of factor analysis available where linear components are extracted out of the variables. We can see the usage of this method in the housing submarkets for examples in the articles from (Bates, 2006) [19] and (Wu & Rashi Sharma, 2011) [20].

In general, "Factor Analysis" is primarily used for data reduction or structure detection [18]. The purpose of data reduction is to remove redundant (highly correlated) variables from the data file, perhaps replacing the entire data file with a smaller number of uncorrelated variables.

The purpose of structure detection is to examine the underlying (or latent) relationships between the variables. In our case, it reduces the complexity of the dataset used in the cluster analysis by indicating the contribution of variables to a component accounting for the most variance in the dataset.

For more theoretical aspects of this method the references books or articles can be read. In the next section the results of PCA analysis on our dataset are presented.

3.4.2. Principal Component Analysis (PCA) execution

In this part the PCA is done on the database to find out about the components extractions. In the previous steps outlier analysis and correlation analysis have been done on the database. These two steps are really important before implementation of the factor analysis. Also, in the correlation analysis (the related tables are in the appendixes part B: Principal Component Analysis), we can see that there are correlations between some variables which makes the factor analysis worthwhile enough.

On the other hand, there are some tests and outcomes from the PCA that we should consider during the assessing the results. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy should be considered carefully. KMO is an index for comparing the magnitude of the observed correlation coefficients to the magnitude of the partial correlation coefficients. The closer the KMO measures to 1 indicate a sizeable sampling adequacy. When KMO is equal to 0.8 and higher it is great, 0.7 is acceptable, 0.6 is mediocre and less than 0.5 is unacceptable. Small KMO values show that a factor analysis of the variables may not be a good idea.

The descriptive statistics of the variables in our dataset is shown in table 9.

Descriptive Statistics

	Mean	Std. Deviation	Analysis N	Missing N
Building Age	23.15	15.688	377	10
Height (cm)	6.34	1.388	338	49
Bwteen Buildings Location	.33	.470	386	1
Single Building Location	.31	.461	386	1
Building Block Surrounding	.54	.499	386	1
Open surrounding	.37	.484	386	1
Curtain wall Façade	.14	.343	338	49
Brick Façade	.54	.499	338	49
Glass Percentage	38.92	15.343	338	49
Total Area (M2) scaled	17.76	16.255	387	0
Electirc Energy (scaled MJ)	31.08	33.999	387	0
Gas Energy (Scaled MJ)	49.90	50.039	376	11
Group of employees	3.98	1.656	387	0

Table 9: descriptive statistics table of the variables for PCA

Communalities are the other outcome of the analysis. They indicate the amount of variance in each variable that is accounted for. Extraction communalities are estimates of the

variance in each variable accounted for by the components [18]. The table 10 shows the communalities for our variables.

Communalities

	Initial	Extraction
Building Age	1.000	.756
Height (cm)	1.000	.680
Bwteen Buildings Location	1.000	.784
Single Building Location	1.000	.548
Building Block Surrounding	1.000	.891
Open surrounding	1.000	.914
Curtain wall Façade	1.000	.718
Brick Façade	1.000	.677
Glass Percentage	1.000	.443
Total Area (M2) scaled	1.000	.805
Electirc Energy (scaled MJ)	1.000	.783
Gas Energy (Scaled MJ)	1.000	.827
Group of employees	1.000	.479

Extraction Method: Principal Component Analysis.

Table 10: Communalities of the variables for PCA

As it can be seen in this table, almost all of the extraction communalities are higher than 0.6, except three of them; single buildings location, glass percentage and group of employees. Indeed, these variables should be put outside of the analysis. By using the trial and error procedure, finally the most acceptable variables have been chosen. The table 11 and 12 show the results of the descriptive statistics and communalities for these variables.

Descriptive Statistics

	Mean	Std. Deviation	Analysis N	Missing N
Building Age	23.15	15.688	377	10
Height (cm)	6.34	1.388	338	49
Building Block Surrounding	.54	.499	386	1
Open surrounding	.37	.484	386	1
Total Area (M2) scaled	17.76	16.255	387	0
Electirc Energy (scaled MJ)	31.08	33.999	387	0
Gas Energy (Scaled MJ)	49.90	50.039	376	11

Table 11: descriptive statistics table of the final variables for PCA

Communalities

	Initial	Extraction
Building Age	1.000	.769
Height (cm)	1.000	.690
Building Block Surrounding	1.000	.916
Open surrounding	1.000	.924
Total Area (M2) scaled	1.000	.837
Electirc Energy (scaled MJ)	1.000	.809
Gas Energy (Scaled MJ)	1.000	.859

Extraction Method: Principal Component Analysis.

Table 12: Communalities of the final variables for PCA

As it can be seen in table 12, all the communalities are higher than 0.6 and we can use them in the PCA.

In addition, the results for the KMO and Bartlett's Test can be seen in table 13. This table indicates that the KMO measure is 0.681 which can be described as acceptable. On the other hand, Bartlett's test of sphericity tests the null hypothesis that the correlation matrix is an identity matrix. An identity matrix is matrix in which all of the diagonal elements are 1 and all off diagonal elements are 0. Hence, the results of these tests are satisfactory for our dataset.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.681
Bartlett's Test of Sphericity	Approx. Chi-Square	1167.506
	df	21
	Sig.	.000

Table 13: KMO and Bartlett's Test for PCA

For the component extraction, the number of components to be extracted can be determined by using the Kaiser criterion, the point of inflection in the scree plot and the percentage of number of residuals above 0.05. The boundary for extraction communalities is 0.6 which is good enough for all of the variables. This indicates that we can use the Kaiser criterion in which components with a total eigenvalue above 1.0 are selected.

The next table that we are going to use is the total variance explained table which can be seen in table 14.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.986	42.651	42.651	2.986	42.651	42.651	2.770
2	1.512	21.593	64.244	1.512	21.593	64.244	2.046
3	1.307	18.671	82.915	1.307	18.671	82.915	1.347
4	.577	8.239	91.154				
5	.255	3.650	94.803				
6	.209	2.989	97.793				
7	.155	2.207	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

Table 14: Total variance explained for the components for PCA

As this table shows, three components have the eigenvalue more than 1. These three components represent a substantial amount of variance in the data. As it can be seen in this table, these components can represent about 83 percent of the dataset which is satisfactory.

On the other hand, the scree plot helps us to determine the optimal number of components. The eigenvalue of each component in the initial solution is plotted in figure 6.

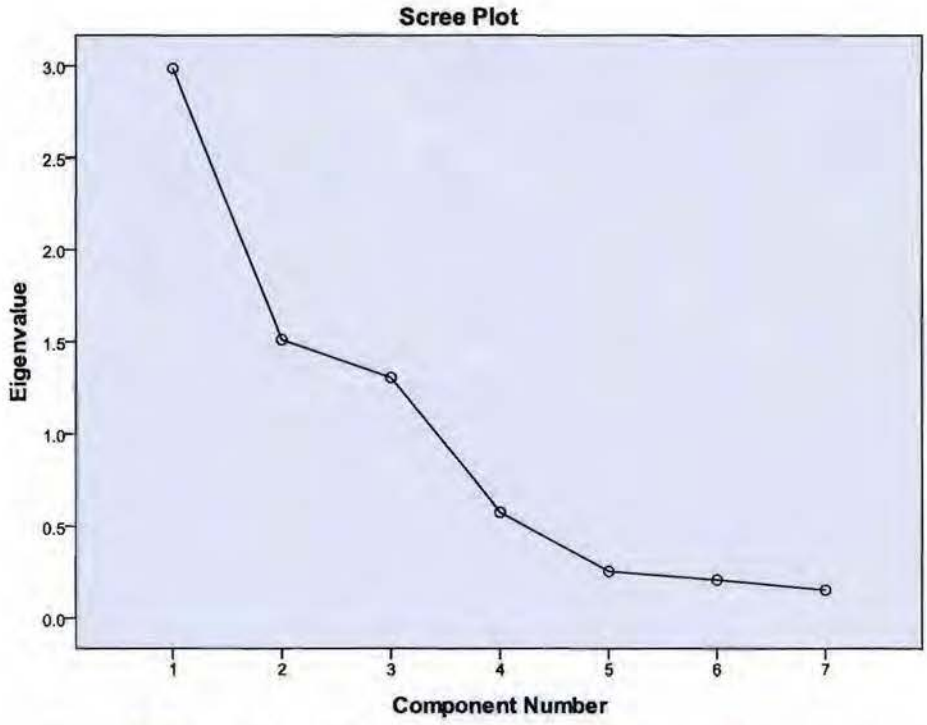

Figure 6: Scree plot of PCA, eigenvalues and components number

Generally, we want to extract the components on the steep slope because the components on the shallow slope contribute little to the solution. As the figure shows, the last big drop occurs between the third and forth components. Indeed, using the three components is the best choice.

In the analysis we have used the oblique rotation (Direct Oblimin in SPSS). Table 15 shows the component scores from pattern matrix and table 16 shows the component scores from structure matrix.

Pattern Matrix[a]

	Component		
	1	2	3
Building Age	.231	-.012	.863
Height (cm)	.293	-.012	-.754
Building Block Surrounding	-.059	.939	.009
Open surrounding	-.058	-.975	.017
Total Area (M2) scaled	.915	.005	-.012
Electirc Energy (scaled MJ)	.865	-.087	-.070
Gas Energy (Scaled MJ)	.938	.033	.082

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 4 iterations.

Table 15: Pattern matrix for PCA

Structure Matrix

	Component		
	1	2	3
Building Age	.168	-.033	.845
Height (cm)	.353	-.122	-.777
Building Block Surrounding	-.305	.955	.056
Open surrounding	.195	-.959	-.023
Total Area (M2) scaled	.915	-.234	-.082
Electirc Energy (scaled MJ)	.893	-.315	-.140
Gas Energy (Scaled MJ)	.923	-.208	.012

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Table 16: Structure matrix for PCA

The pattern matrix contains the regression coefficients between each variable and a component. On the other hand, the structure matrix contains the correlation coefficients between each variable and a component. These tables help us to determine what the components represent. As it is shown in these tables, each component has correlation with some specific variables of our dataset.

Component 1 is highly correlated to gas energy, total area and electric energy. Component 2 is highly correlated to open surrounding and building block surrounding. Component 3 is highly correlated to building age and height.

According to these correlations, we can choose some specific labels (names) for each of the components. Table 17 shows the chosen name for the three extracted components of our dataset.

Component No.	Chosen name
Component 1	Energy Usage and Scale
Component 2	Surrounding Type
Component 3	Building Age and Height

Table 17: Final components with the given names of PCA

The component 1 highly represents the energy usage and scale of the building. Component 2 represents the surrounding type and component 3 the building age and height.

The other important tables of the PCA are included in the appendixes part B. In the next section, these components are going to be used for the cluster analysis with the k-means clustering method.

3.4.3. K-means clustering results with extracted components

By using the above extracted components, we are going to perform the k-means cluster analysis on our dataset. Upfront specified number of clusters (K) is going to be used during the analysis. We changed the k number to check the best fit and most interpretable results for our cluster analysis.

To choose the best option for the k, we checked the final results and the derived clusters. We tried to interpret them first and then choose the best number for k. Also, the connectivity and coherence of the buildings in each cluster were assessed carefully. We tried the k number equal to 4, 6, 8 and 10. Finally, we have chosen k=10 which makes the results more interpretable.

For checking the homogeneity of the clusters, we have used the weighted average standard deviation (WASD) measure [20]. The Standard deviation per cluster regarding a characteristic (s or SD) is the square root of the variance (s2) per cluster regarding that characteristic. The equation below shows the way of calculation for WASD:

$$\mathrm{WASD}_{\text{per characteristic}} = \frac{\sum_{i=1}^{i=n}(N_i * SD_i)}{N} = \frac{\sum_{i=1}^{i=n}\left(N_i * \sqrt{\frac{\sum_{j=1}^{j=N_i}(x_j - \bar{x})^2}{N_i}}\right)}{N}$$

A low WASD indicates a high homogeneity within the cluster. We have run the analysis for different k numbers and for each of the clusters, the WASD have been calculated for the

energy use and scale component (because this component is the main focus of the clustering). The results of the calculation can be seen in table 18.

Number of K	WASD										
	Cluster 1	cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Average
4	21.12	29.69	26.06	33.76							27.6575
6	23.72	14.79	23	19.22	25.06	8.05					18.9733333
8	10.4	14.55	8.08	12.86	6.2	26.73	7.14	12.03			12.24875
10	5.21	13.35	5.74	23.72	8.56	11.1	10.08	5.82	5.13	7.07	9.578

Table 18: WASD for different k numbers in each cluster

As it is obvious in this table, the low amount for the WASD can be seen for 10 clusters. Indeed, we have chosen k=10 for gaining better results. The k number more than 10 does not make sense anymore because the difference between the WASD is not effective enough anymore and also the numbers of buildings in the clusters are not logical.

From the analysis results for the k-means clustering with k=10, first the ANOVA table is presented in table 19. The ANOVA table indicates which variables contribute the most to your cluster solution [18].

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Energy usage and Scale	27.803	9	.201	314	138.383	.000
Surrounding Types	34.691	9	.052	314	673.409	.000
Building age and Height	28.765	9	.213	314	134.857	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 19: ANOVA table of the k-means clustering with k=10

Variables with large F values provide the greatest separation between clusters. The table indicates that surrounding types, energy usage and scale and building age and height are more important, respectively. Furthermore, the sig equal to .000 shows the effectiveness of the analysis.

On the other hand, table 20 shows the number of the buildings in each cluster. As it is obvious from this table, 63 buildings are put outside of the analysis because for them the components were not derived in the previous steps. Indeed, in total 324 buildings have been divided in ten clusters. These buildings in all clusters are separated logically and the distributions of them in the clusters are acceptable.

Number of Cases in each Cluster

Cluster	1	17.000
	2	41.000
	3	22.000
	4	92.000
	5	31.000
	6	32.000
	7	27.000
	8	16.000
	9	12.000
	10	34.000
Valid		324.000
Missing		63.000

Table 20: number of cases in each cluster for the k-means clustering with k=10

Regarding the interpretation of clusters, we try to look at the final cluster centers and the separation of the buildings on the basis of their components. Hence, the final cluster centers and the descriptive interpretation are the basis for the interpretation of each cluster. Table 21 shows the final cluster centers.

Final Cluster Centers

	Cluster									
	1	2	3	4	5	6	7	8	9	10
Energy usage and Scale	1.83894	-.08893	-.41081	-.55709	-.57922	-.28977	1.70426	1.20336	1.49558	-.74043
Surrounding Types	.60541	.86388	-1.18660	.84731	-1.20493	-1.23861	-1.25146	-1.03421	.10452	.85297
Building age and Height	.33397	.68371	1.44591	-.60367	-1.00705	.08309	-.93661	1.26006	-1.19136	1.33605

Table 21: final cluster centers for k-means clustering with k=10

According to these centers and also the average of the important variables, the table 22 shows the final interpretation of each cluster.

Cluster number	Average Total Area (m2)	Average Electric energy (MJ)	Average Gas energy (MJ)	Total area level	Electric energy level	Gas energy level
1	4,606	695,010	1,521,963	High	High	High
2	1,551	202,020	510,547	Medium	Medium	High
3	1,051	182,451	335,154	Medium	Low	Medium
4	954	134,366	231,938	Low	Low	Medium
5	940	166,714	206,782	Low	Low	Medium
6	1,397	243,883	355,503	Medium	Medium	Medium
7	4,312	903,966	1,218,739	High	High	High
8	3,523	695,029	1,079,237	High	High	High
9	4,578	797,608	1,013,510	High	High	High
10	671	89,699	190,544	Low	Low	Low

Table 22: levels of energy usage and total area for each cluster

3.4.4. K-means clustering results with all the variables

In this part the k-means cluster analysis is done by using all the variables that we have in the dataset. In the previous section, by using the extracted components, some categorical variables are also entered in the analysis. Here we would like to consider the scale variables in the analysis only. By this way we can include all the buildings in our clustering results and there is not any missing value anymore.

For the analysis, firstly gas energy usage, electric energy usage, total area, height, building age and glass percentage were chosen to be used in the analysis. With these variables several tests have been run on the dataset with different numbers of clusters. However, the results of the analysis indicated that the gas energy, electric energy and total area of the building are the most important variables for clustering and the other variables are not really effective in the final results outcome. For example, table 23 clarifies this issue better.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Building Age	926.713	3	240.640	373	3.851	.010
Height (cm)	13.460	3	1.822	334	7.386	.000
Glass Percentage	312.581	3	234.710	334	1.332	.264
Electric Energy (MJ)	1.287E13	3	1.572E10	383	818.719	.000
Gas Energy (MJ)	2.703E13	3	3.440E10	372	785.860	.000
Total Area (m2)	1.991E8	3	1103045.823	383	180.532	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 23: ANOVA table of k-means clustering for choosing the variable for analysis

As it can be seen in this table, the gas energy, electric energy and total area have the high amount of F but the other variables are not really significant in terms of F number.

In this part again different numbers for the k tested. The results were assessed carefully to choose the best and interpretable clusters. Furthermore, the buildings should be divided in the clusters in good orders. The k number was assumed as 4, 6, 8 and 10. Finally, k=8 is chosen for the analysis. The ANOVA table for this analysis is shown in table 24.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Electric Energy (MJ)	5.962E12	7	7.619E9	379	782.484	.000
Gas Energy (MJ)	1.289E13	7	9.965E9	368	1293.491	.000
Total Area (m2)	8.963E7	7	1035593.428	379	86.545	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table 24: ANOVA table of k-means clustering with k=8 with variables

According to the F amount in this table, the gas energy usage has the most effects in making the clusters and then the electric energy and total area are effective. Also the final cluster centers are shown in table 25.

Final Cluster Centers

	Cluster							
	1	2	3	4	5	6	7	8
Electric Energy (MJ)	398463	993702	349916	72087	155235	414507	957467	979345
Gas Energy (MJ)	949513	982027	1476926	108390	390505	460182	300812	1523328
Total Area (m2)	2438	3533	3585	666	1421	2166	2959	4432

Table 25: Final cluster centers table of k-means clustering with k=8 with variables

On the basis of these cluster centers, we would like to define some levels for the three used variables in this analysis. The clusters can be interpreted like the table 26. In this table different levels for each of the variables have been mentioned. The buildings in each of these clusters follow the same trend in terms of the energy usage and the total area.

Cluster number	Electric energy level	Gas energy level	Total area level
Cluster 1	Medium	High	Medium
Cluster 2	High	High	High
Cluster 3	Medium	High	High
Cluster 4	Low	Low	Low
Cluster 5	Medium	Medium	Low
Cluster 6	Medium	Medium	Medium
Cluster 7	High	Medium	Medium
Cluster 8	High	high	High

Table 26: levels of energy usage and total area for each cluster

The number of buildings in each cluster can be seen from table 27. The acceptable separation between the buildings in the clusters is achieved through the analysis. Cluster 4 with 160 buildings has the highest share of the buildings. In this cluster the buildings have low amount of electric and gas energy usage and also very low total area compare to the other clusters.

Number of Cases in each Cluster

Cluster	1	22.000
	2	17.000
	3	17.000
	4	160.000
	5	84.000
	6	37.000
	7	10.000
	8	40.000
Valid		387.000
Missing		.000

Table 27: Number of buildings in each cluster

3.5. Clustering with two-step clustering method

In this part we would like to implement the cluster analysis by using the two-step clustering method. To reach this goal, we are going to use different variables from the dataset in different steps. In this way good overviews can be gained from different perspectives about the energy usage and the related attributes. In the following sections we have presented the clustering results by dividing the clustering regarding the variables that have been used in that specific clustering analysis.

3.5.1. Clustering regarding energy usage

The electric and gas energy usage are the variables that have been used for the clustering in this part. On the basis of these variables and by trial and error method for the best number for clustering, three clusters have been defined. The results can be seen in figure 7. As it can be seen in the figure, three different clusters are on the basis of the energy usage; low energy usage, medium energy usage and high energy usage.

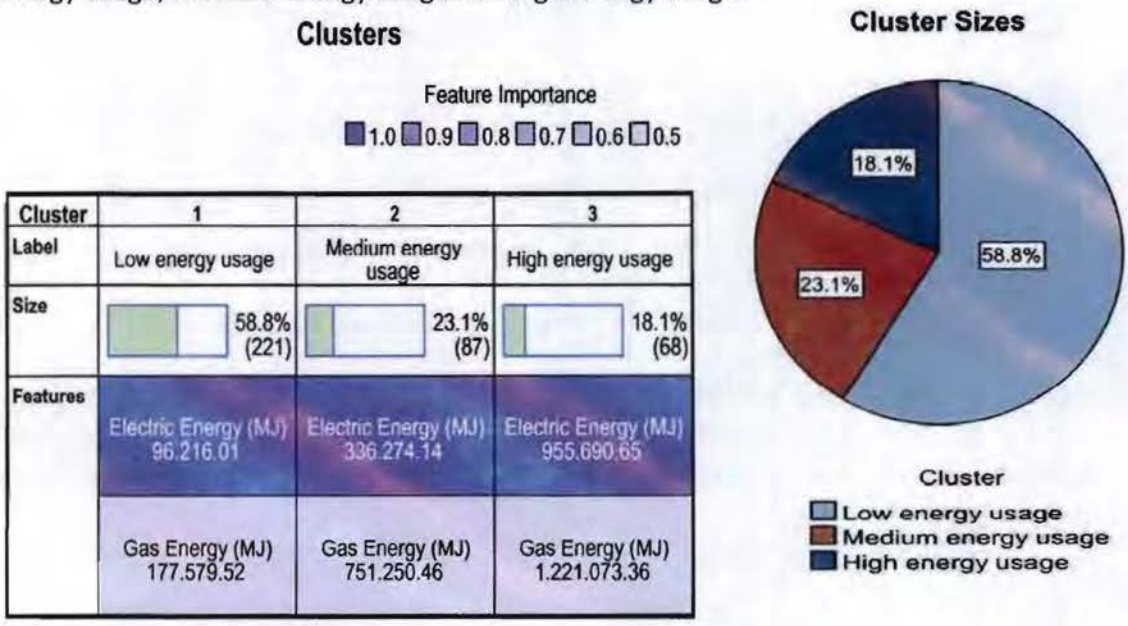

Figure 7: Two-step clustering results regarding energy usage

The electric energy is more important compared to the gas energy in clustering the buildings. The importance of electric energy usage is 1 and for gas energy usage this importance in clustering is 0.5.

The results show that most of the buildings (221) from the dataset are in the low energy usage group while 87 buildings are in the medium energy usage group and 68 buildings have high energy usage performance. The mean of each cluster is shown in the features part of the figure for each of the variables.

3.5.2. Clustering regarding energy usage and building characteristics

This clustering is on the basis of the energy usage of the buildings and the building characteristics including total area, height of the building, building age and glass percentage of the building.

The results of the clustering can be seen in figure 8. Four clusters have been defined. On the basis of the main features of these clusters, we can call these clusters as; low energy usage with old buildings cluster, medium energy usage cluster, low energy usage with new buildings cluster and high energy usage cluster. The gas energy usage is the most important variable in the clustering and glass percentage has the least importance.

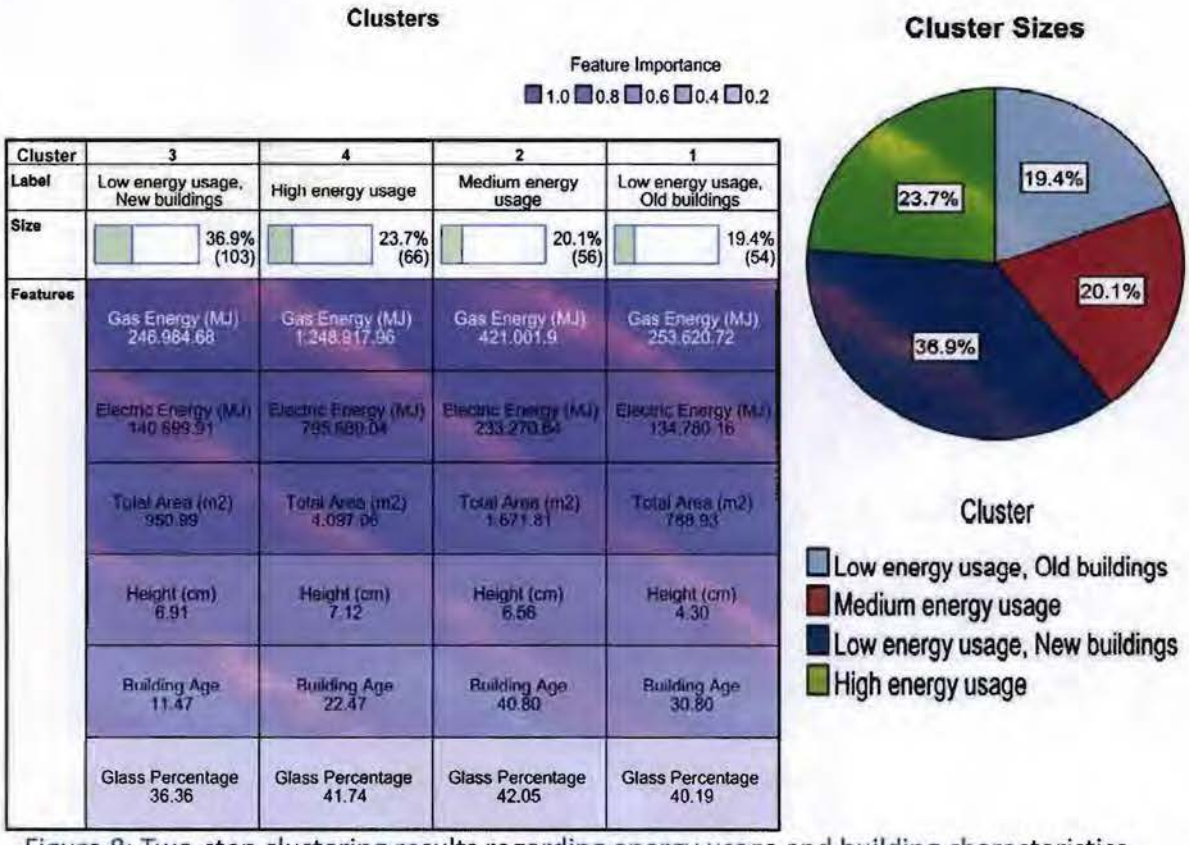

Figure 8: Two-step clustering results regarding energy usage and building characteristics

We can see in the cluster 1 with 54 buildings that the low energy usage is occurred with the old buildings and total area and height of the buildings are also low. However in cluster 3 with low energy usage we have new buildings, low total area and relatively high buildings.

In cluster 2 the medium energy usage comes with medium total area, medium height of the building and high building age. Cluster 4 shows the high energy usage with high total area, high height of the buildings and medium building ages. The glass percentages are more or less same in all the clusters.

One can conclude from this clustering that the total area of the buildings is more important compare to the building age or height of the buildings in energy usage. Hence, when the total area is higher we expect more energy usage for the buildings. Further assessment on these clusters in the future can derive more features of the energy usage performances of the buildings.

3.5.3. Clustering regarding energy usage and users number

The other clustering is by considering the energy usage and number of employees of the company. In this regard three clusters have been recognized; low energy usage and low employee numbers, medium energy usage and high employee numbers, high energy usage and high employee numbers.

As it can be seen, when the number of employees are high the energy usage could be medium and high and it should be investigated more to find out about the buildings of these clusters. However, when the number of employees is low, then the energy usage is also low.

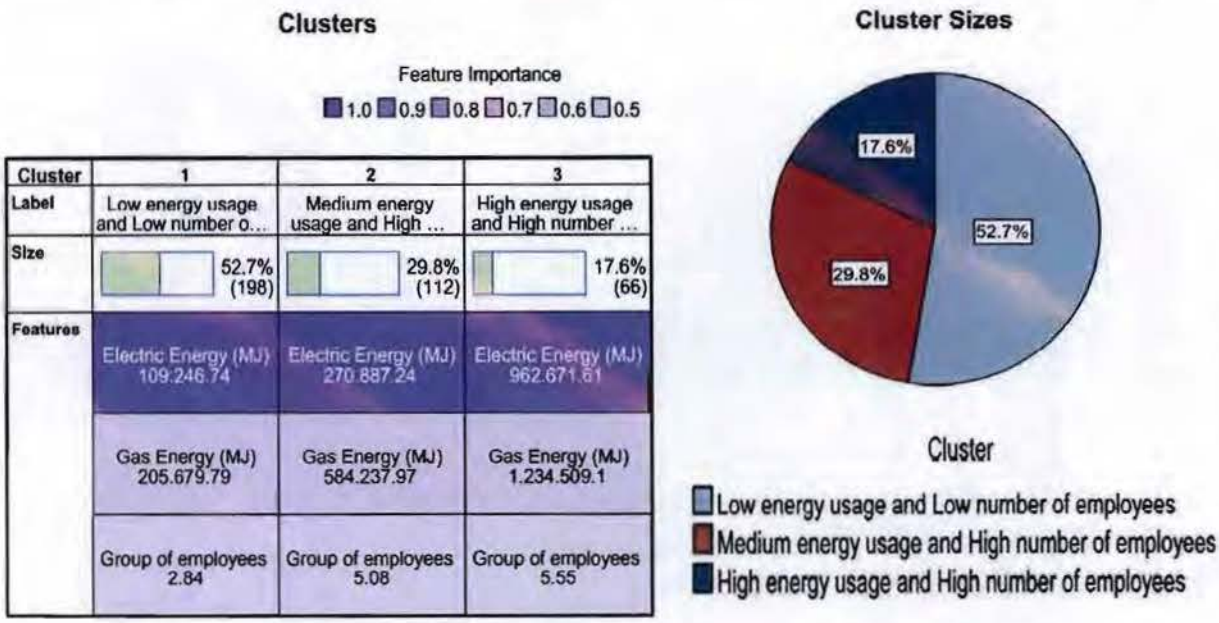

Figure 9: Two-step clustering results regarding energy usage and users number

3.5.4. Clustering regarding energy usage and location type

The location types are the other variables which we would like to perform the cluster analysis by using them in combination with the energy usage. Four clusters have been derived after analysis; High energy usage and single location cluster, low energy usage and single location cluster, medium energy usage and side location, medium energy usage and between buildings location.

As it can be seen in the figure 10, the buildings locations are important variables in the clustering of the buildings. The appropriate separation can be seen in the clusters regarding the building locations and buildings are divided in the groups on the basis of their location types.

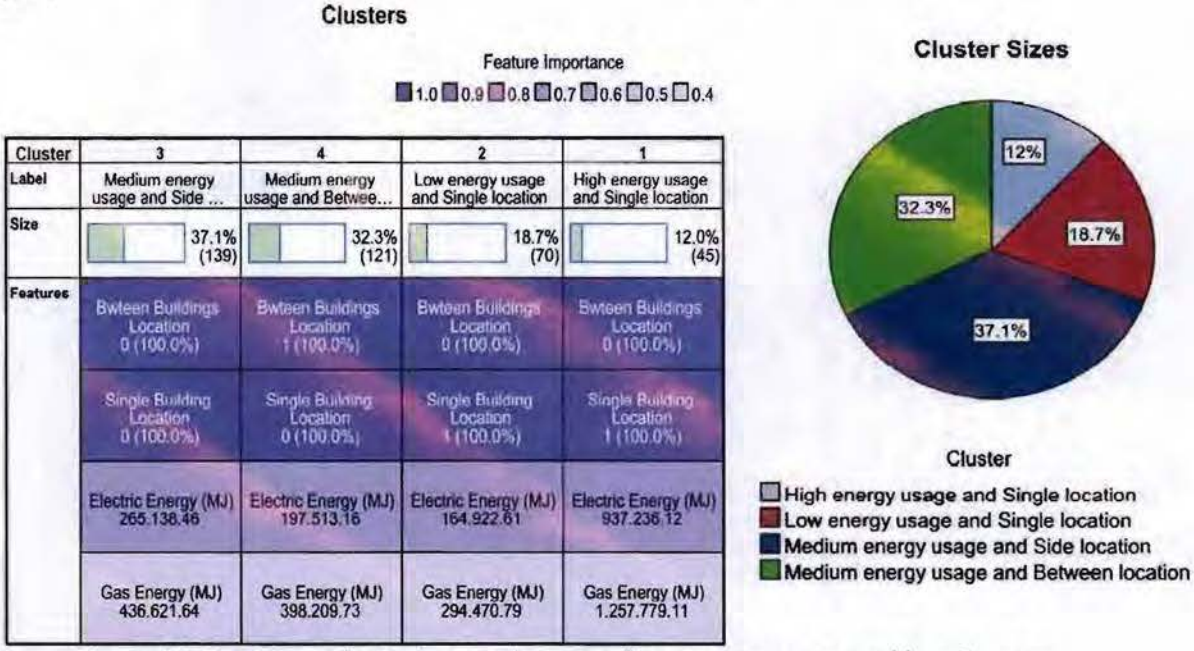

Figure 10: Two-step clustering results regarding energy usage and location type

Two categories of the energy usage is possible when the buildings are in the single locations and further studies are needed to find out about the reasons behind the low or high energy usage in these kinds of buildings. However, when the building locations are between buildings or side buildings the energy usage level is medium. This means that the locations cannot be good variables for prediction of the energy usage and further variables are needed next to these categorical variables for the energy usage prediction.

3.5.5. Clustering regarding energy usage and surrounding type

The surrounding type in combination with the energy usage is also used for the clustering the buildings. Three clusters have been defined on the basis of the surrounding types; low energy usage and building block surroundings, high energy usage and building block or river surroundings, medium energy usage and open surroundings. Most of the buildings are in the first cluster. Figure 11 shows this clustering in more details.

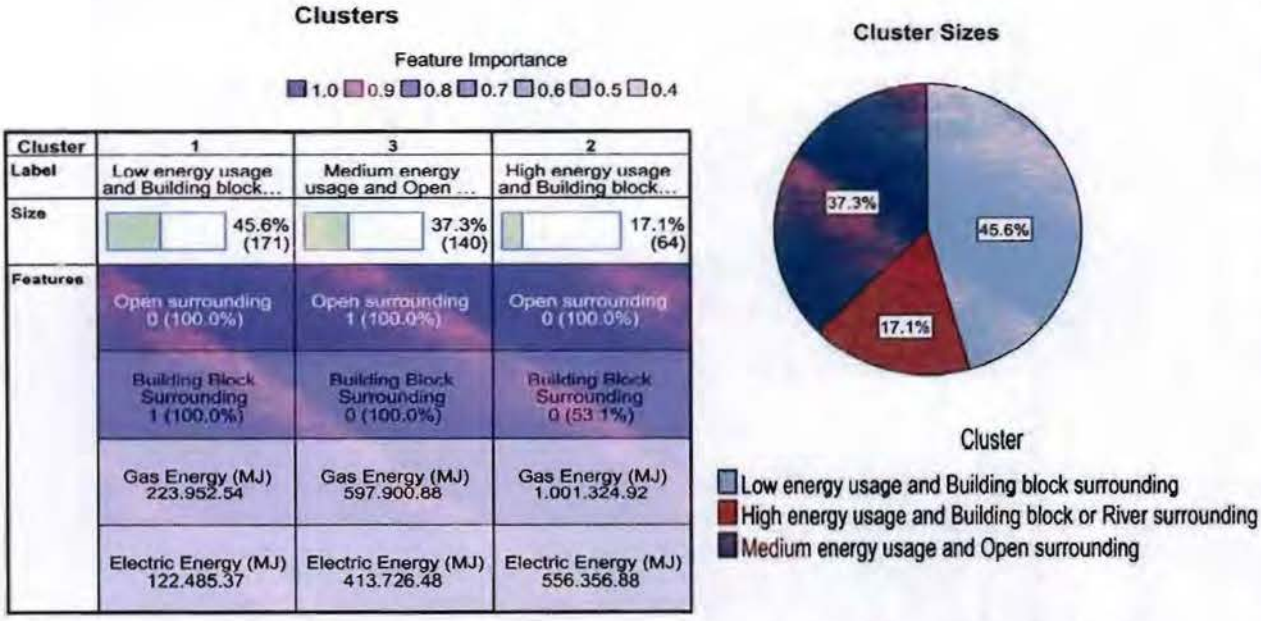

Figure 11: Two-step clustering results regarding energy usage and surrounding type

3.5.6. Clustering regarding energy usage and façade type

On the basis of the energy usage and façade types, four clusters have been derived after the cluster analysis on our dataset. These clusters are namely; high energy usage and brick façade cluster, low energy usage and brick façade cluster, medium energy usage and curtain wall façade cluster, medium energy usage and mixed façade cluster.

The reason behind the high energy usage or low energy usage with the brick façade type is not clear and should be investigated in further studies more. However, we can expect that the curtain wall façade or mixed façade types come with medium energy usage. For sure, we should not forget the effects of the other variables.

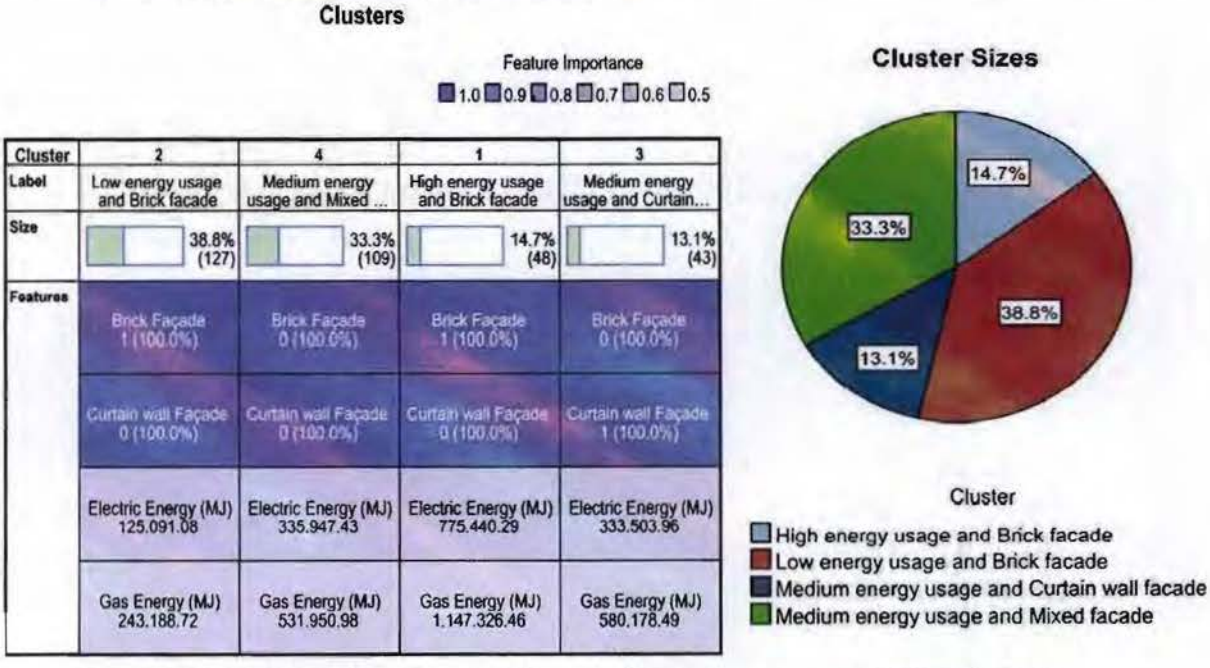

Figure 12: Two-step clustering results regarding energy usage and facade type

3.5.7. Clustering regarding energy usage and scale of the building

In the last section of the clustering we would like to consider the energy usage with the scale of the buildings. Total area and number of employees show the scale of the building or with the other words the company in the building. Figure 13 shows the analysis results.

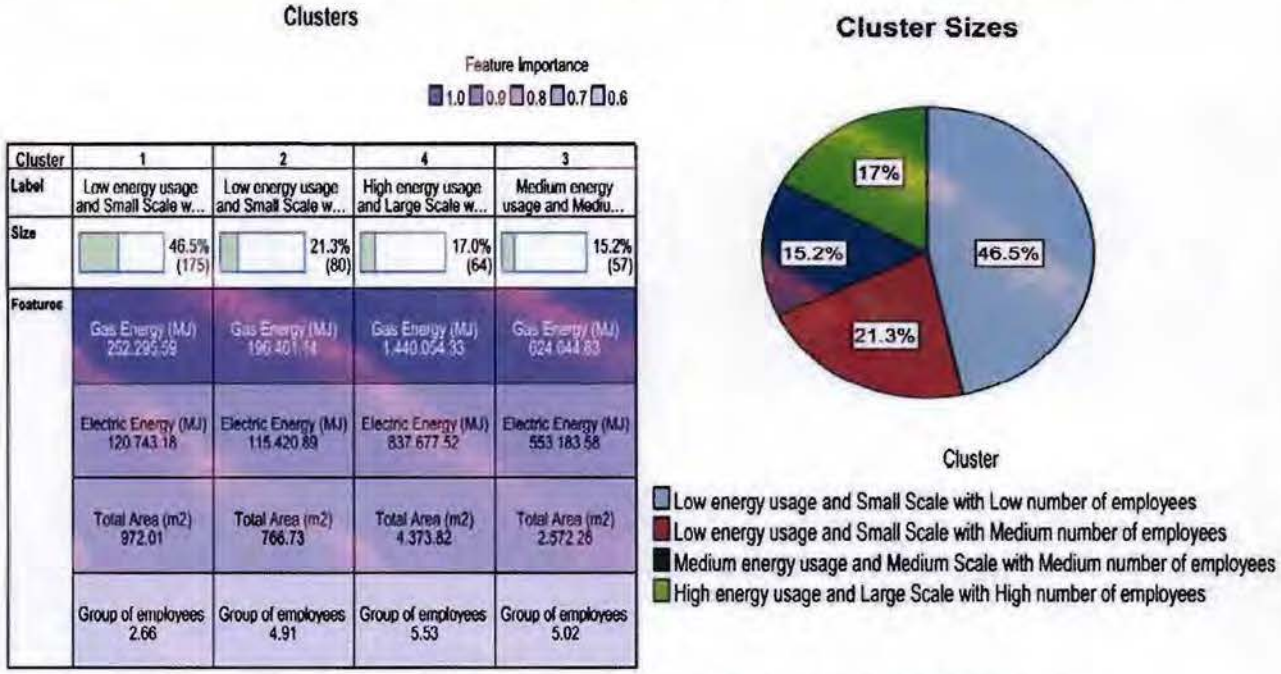

Figure 13: Two-step clustering results regarding energy usage and scale of the building

3.6. Conclusions

In this chapter we have done the cluster analysis on the energy usage dataset of the buildings in services sector in Eindhoven. We have used two different methods of the clustering namely k-means clustering and two-step clustering. The SPSS software has been used for the analysis. By means of these two methods different clusters of buildings have been built. The energy usage types are the main variables for performing the clustering.

In using the k-means clustering, the principal components analysis also has been preformed for the component's extraction from the database. Later on these components have been used for the clustering. Also, all of the variables in the dataset have been used in the other section for the clustering one more time.

In using the two-step clustering, the energy usage variables have been used in different sections and different combinations of the variables proposed for the clustering purposes. Consequently, different clusters of the buildings have been made which provide us with various perspectives about the buildings in this sector.

The results of this chapter can be used to assess the buildings more and performing different kinds of analyses to achieve the useful policies for the energy usage reduction in the services sector of city of Eindhoven.

Chapter 4: Energy usage prediction

In this chapter the energy usage prediction method is going to be described. First a short theory about the method that we would like to use is presented. Then the statistical analysis steps and clustering for the prediction are depicted in details. Next the outcomes from the analysis are included and the conclusions about the results provide the reader with clear view about the findings of this section.

4.1. Aim

The aim of this chapter is prediction of the energy usage of the buildings in services sector in the city of Eindhoven. For this issue, the dataset which we have built in previous steps is going to be used. For finding out the energy prediction formula and the important variables in energy usage prediction, the statistical analysis is done. The most useable method for prediction with the statistical analysis is linear regression model. We should use multiple regression models because we have more than one predictor for the energy usage. In the next part a short theory about this method has been depicted. Also, because of the non-uniformity of the dataset, the buildings are going to be put in different clusters. By this way the prediction equations can be derived more accurately.

4.2. Linear regression theoretical underpinnings

Linear regression is used to model the value of a dependent scale variable based on its linear relationship to one or more predictors [18]. When we have more than one predictor or independent variable, we call it multiple linear regressions.

In this method a straight line or a linear relationship is assumed between the dependent variable and the predictors. The relationship can be written down as follow:

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} + e_i \qquad (3)$$

Where

y_i is the value of the i^{th} case of the dependent scale variable
p is the number of predictors
b_j is the value of the j^{th} coefficient, $j = 0, \ldots, p$
x_{ij} is the value of the i^{th} case of the j^{th} predictor
e_i is the error in the observed value for the i^{th} case

As it was mentioned before, the regression analysis is done by using the SPSS software. In using this software for the analysis, we should consider some issues when we want to run the command. For the linear regression in the statistics part the "part and partial correlations" and "collinearity diagnostics" should be used in the analysis. This will help to have better outcomes. Also, pairwise cases should be excluded from the analysis.

4.3. Dependent and independent variables

For prediction of the energy usage in the services sector in Eindhoven the dependent and independent variables should be separated. In our case, electric energy, gas energy and

total energy are the dependent variables that we would like to predict them for the different buildings.

The independent variables or predictors are building age, height, location situations, surrounding situations, façade types, glass percentage, total area of the building and number of employees of the company which stay in the building.

An important issue in the analysis is that we have scaled the variables for better performance. In this way the variables differ in the same range and the outcome of the analysis is acceptable. For example, the total area of each building is divided by 100 and it is called total area (m2) with scaling. This also has been done for energy units.

The first step that we should perform before the regression analysis is checking the correlation between the predictors. We should try to minimize the correlation effects between the predictors to gain the most acceptable outcome. The next section is about this issue.

4.4. Correlation analysis

The correlations between the different predictors and the energy usage variables have been checked with the software for the entire dataset. We have divided this analysis to two parts; one for the categorical variables and the other one for the rest of the variables. The results of this analysis can be seen in appendixes section C: Correlation analysis. For each of the dependent variables, the correlation analysis is done with the predictors.

For the scale (continuous) variables, the correlation is significant between the energy and the total area, number of employees and height. For the categorical variables, the correlation is significant between the energy and the between building location, single building location, building block and open surrounding.

Table 28 shows the correlation coefficient and significance between different energy types and different predictors. It should be mentioned that only the important correlations have been summarized in this table.

As we know when the correlation coefficient is more between the dependent variable and independent variable, it means that this independent variable can be a suitable predictor for the dependent variable. However, this aspect should be proven through the regression analysis.

From the table 28, we can say that total area is the most important predictor. Also, the number of employees plays an important role for prediction. However, in the regression analysis the outcomes might be different in some cases. In the table 28, N is the number of cases which correlation has been done for them and correlation is significant at the 0.01 level (2-tailed).

Dependent variables		Predictors							
		Total area	Number of employees	Height	Building age	Between building location	Single location	Building block surrounding	Open surrounding
Electric energy	Pearson Correlation	0.733	0.499	0.289	-	-0.222	0.321	-0.335	0.250
	Sig. (2-tailed)	0.000	0.000	0.000	-	0.000	0.000	0.000	0.000
	N	376	376	331	-	375	375	375	375
Gas energy	Pearson Correlation	0.776	0.454	0.227	0.174	-0.136	0.234	-0.253	0.158
	Sig. (2-tailed)	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000
	N	376	376	331	366	375	375	375	375
Total energy	Pearson Correlation	0.807	0.502	0.269	0.123	-0.181	0.286	-0.304	0.207
	Sig. (2-tailed)	0.000	0.000	0.000	0.019	0.000	0.000	0.000	0.000
	N	376	376	331	366	375	375	375	375

Table 28: Correlation summary for the dependent variables and important predictors

The other important aspect of correlation analysis is finding out about the correlation between the predictors. From the correlation analysis results, it is clear that the correlation is significant between total area and employee numbers. However we would like to keep both of these variables in the prediction because both of them are highly correlated to the energy usage. Indeed, we define a new variable which is called "scale of the building" and it is defined as follow:

*Scale of the building = Total area of the building * Number of employees of the company* (4)

On the other hand, for the categorical variables high correlation is between the location situations and surroundings. Therefore, we are going to use only one of these variables in the prediction which has shown high correlation to the energy usage.

Furthermore, brick façade type is going to be used as the representative of the façade type because of the high correlation between different façade types which can be seen through analysis.

To sum up, for the prediction analysis, we are going to use the scale of the building (total area and number of employees), surrounding situation, building age, height, brick façade type and glass percentage as the predictors to predict electric energy, gas energy and total energy separately.

4.5. Energy usage prediction analysis

For the prediction analysis, we would like to consider the energy usage types separately. So the analysis is done for electric energy, gas energy and total energy in different sections.

At first the full database was used in the analysis. The results of the regression analysis for the entire buildings of the dataset show that for all the buildings together it is hard to find one equation for the prediction or find the important predictors. Then, we decided to divide the buildings on the basis of the sub-sectors. Again the results are not satisfactory. We will

show the results for different sub-sectors in next sections. Also, we used the clusters, which we have defined in previous section, for regression analysis but unfortunately the results were not good enough.

Finally, we decided to put the buildings into different clusters again and define one formula for each cluster. With this method each cluster has a specific energy prediction formula and specific important predictors. This issue is done by using the two-step clustering method. The used variables for clustering are different for each type of the energy and we will describe them in the relevant parts. The numbers of the clusters for the predictions are achieved by the best fit of the prediction equations.

In the following sections the sample results for some clusters are presented for each of the energy types to make the point clear enough. Also, the interpretations of the predictors and formula are described later in the current chapter.

4.5.1. Energy usage prediction for different sub-sectors

In this part the multiple regression analysis is preformed for the energy usage prediction for each of the sub-sectors of the services. As we mentioned before, the buildings have been divided into sub-sectors on the basis of the SBI 93 codes. In this part the results of the analysis for each of these sub-sectors are presented.

4.5.1.1. Energy usage predcition in retail sub-sector

The retail sub-sector has most of the buildings of our dataset. There are 204 buildings in this sub-sector. Firstly, the table 29 and the figures 14 to 16 show the statistics situations of the different energy usage variables in this sub-sector.

Statistics

		Electric Energy (MJ)	Gas Energy (MJ)	Total energy (MJ)
N	Valid	204	198	198
	Missing	185	191	191
Mean		294524.82	505545.67	803215.09
Median		146714.83	298624.91	468728.11
Std. Deviation		328097.899	506915.260	788372.135
Variance		1.076E11	2.570E11	6.215E11
Skewness		1.341	1.095	1.123
Std. Error of Skewness		.170	.173	.173
Kurtosis		.421	-.211	-.052
Std. Error of Kurtosis		.339	.344	.344
Minimum		1592	4757	15137
Maximum		1033006	1550050	2583056

Table 29: Statistics for energy usage in retail sub-sector

Considering the statistics table, in all energy types the mean is quite different from the median, suggesting that the distribution is asymmetric. Also, the histograms for all the energy types show that there is no normal distribution in our dataset. The difference between the normal distribution which is shown with a black line and the current distribution is obvious. These issues indicate that the prediction of the energy usage for this dataset is quite hard.

The multiple linear regression analysis has been implemented in the dataset in the retail subsector. The results of the analysis are not satisfactory as we expected before. However, through the analysis we could find out about the more important variables in the energy usage of the buildings in this sub-sector. The relevant tables have been included in the appendixes in part D: Sub-sectors regression analysis, 1-Retail sub-sector.

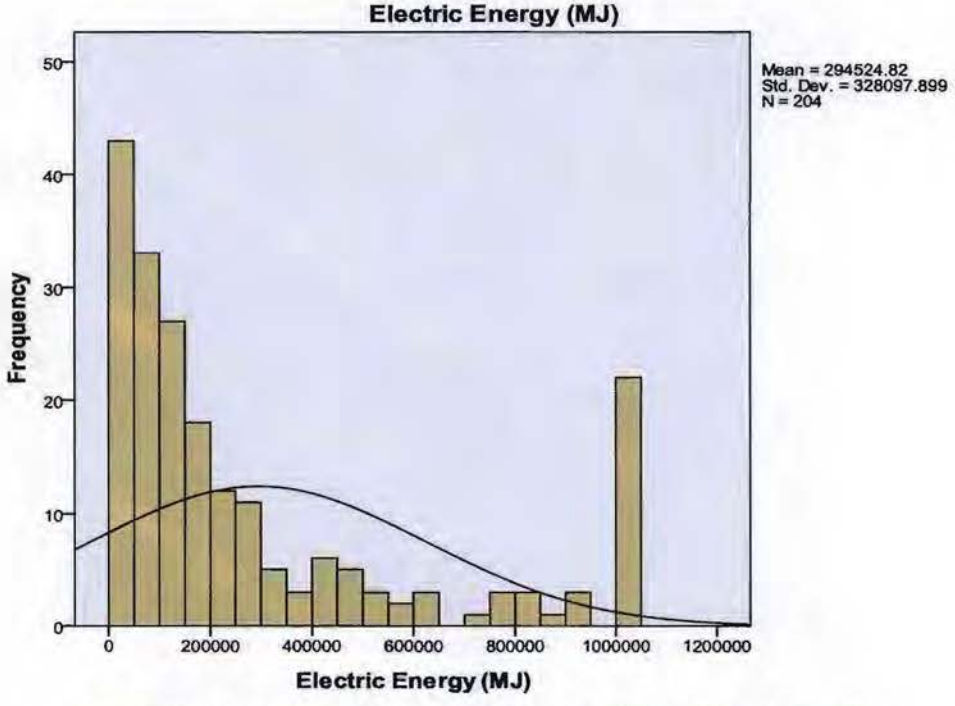

Figure 14: Histogram for electric energy usage in retail sub-sector

In terms of the electric energy usage, total area of the building, number of employees and the surrounding type are more important, respectively. Also, they all have the positive effects on the electric energy usage which means that when they increase, the electric energy usage will also increase.

Regarding the gas energy usage, total area of the building, building age and number of employees are more important, respectively. Again, they all have the positive effects on the gas energy usage which means that when they increase, the gas energy usage will have higher amounts too.

Considering the total energy usage, total area of the building, number of employees and building age are important predictors, respectively. All of these predictors have the positive

effects on the total energy usage. This means that by increasing these predictors, the total energy usage will increase consequently.

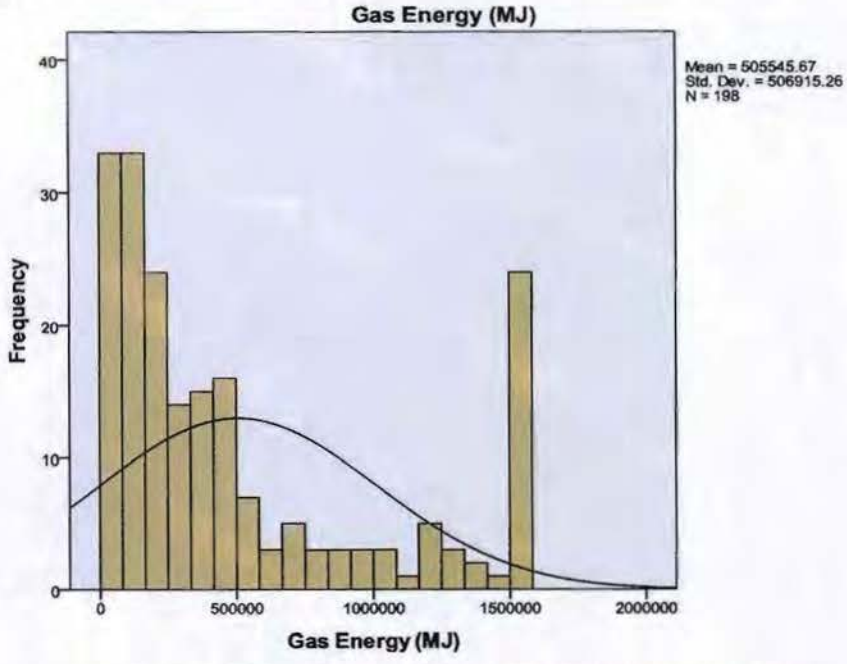

Figure 15: Histogram for gas energy usage in retail sub-sector

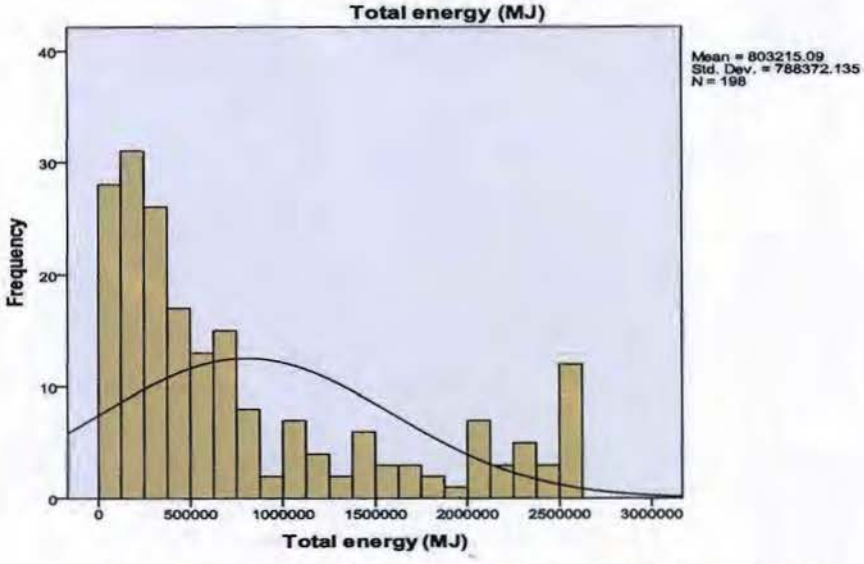

Figure 7: Histogram for total energy usage in retail sub-sector

4.5.1.2. Energy usage predcition in catering sub-sector

There are only 8 buildings in this sub-sector. The same statistics analysis has been done on these buildings. The results have been shown in table 30 and figures 17 to 19.

As it can be seen in the histograms, the normal distribution is not followed by the available data. So it can be predicted that in the regression analysis the desired results would not be easy to be reached.

Statistics

		Electric Energy (MJ)	Gas Energy (MJ)	Total energy (MJ)
N	Valid	8	8	8
	Missing	0	0	0
Mean		485893.90	687641.27	1173535.16
Median		429789.18	577082.93	1078348.19
Std. Deviation		452476.632	653277.972	1097431.409
Variance		2.047E11	4.268E11	1.204E12
Skewness		.186	.365	.271
Std. Error of Skewness		.752	.752	.752
Kurtosis		-2.175	-1.866	-1.992
Std. Error of Kurtosis		1.481	1.481	1.481
Minimum		19901	16481	36382
Maximum		1033006	1550050	2583056

Table 30: Statistics for energy usage in catering sub-sector

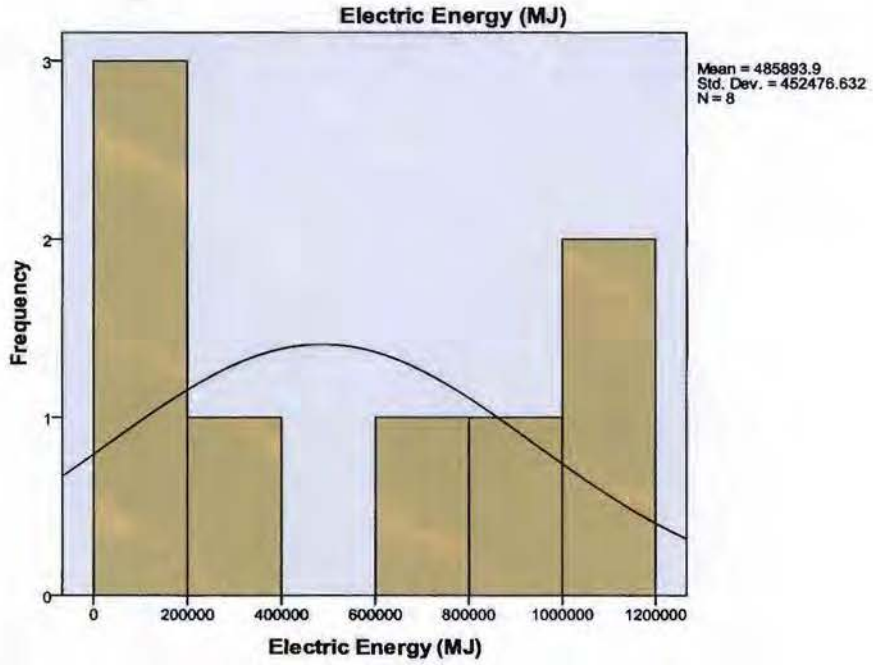

Figure 8: Histogram for electric energy usage in catering sub-sector

After running the regression analysis this issue is proved and we could not perform the regression analysis on this sub-sector because no variable was entered into the equation. Therefore, the prediction of the energy usage in this sector cannot be possible. This means that the energy usage behaviors of the buildings in this sub-sector are completely different from each other.

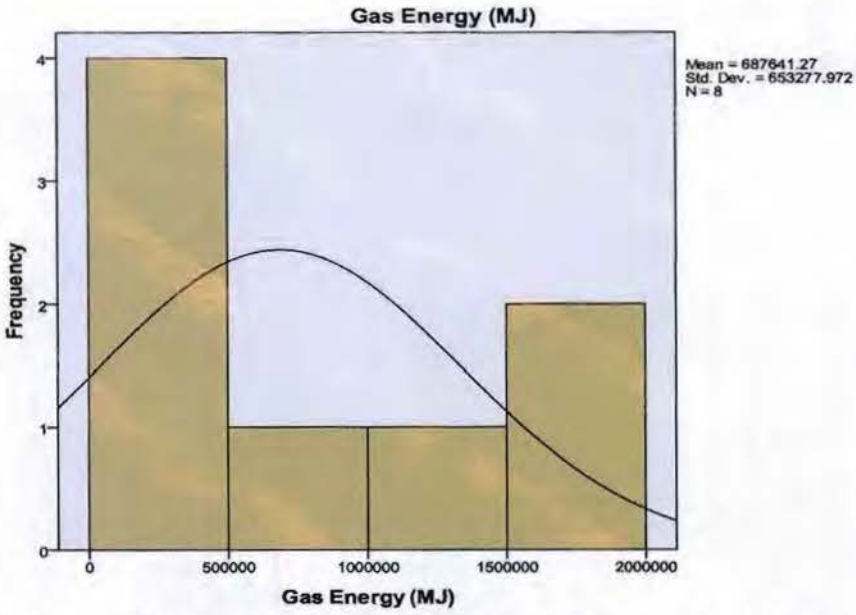

Figure 18: Histogram for gas energy usage in catering sub-sector

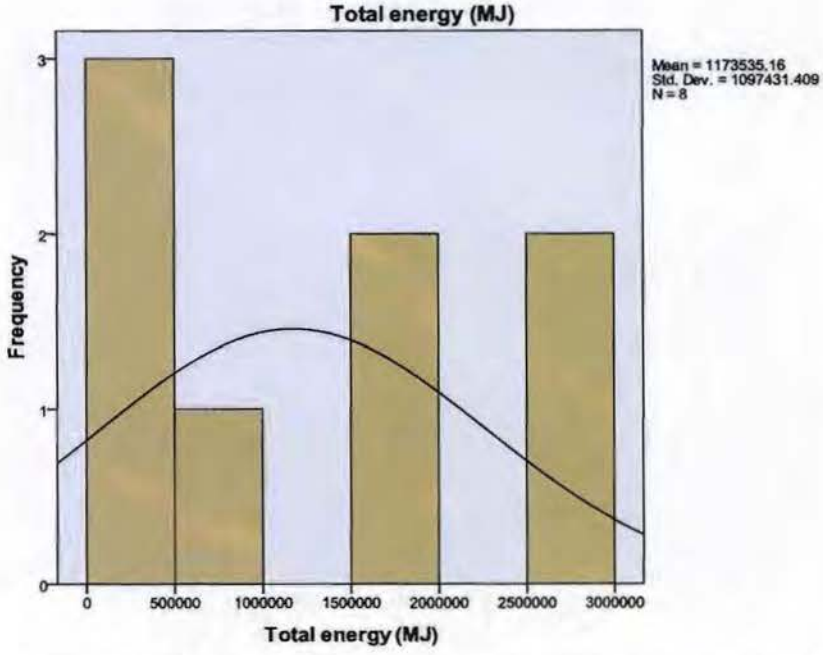

Figure 19: Histogram for total energy usage in catering sub-sector

4.5.1.3. Energy usage predcition in offices sub-sector

The next sub-sector is offices. This sub-sector has a high share of buildings which includes 133 buildings. The frequencies analysis is done on the energy usage variables in this dataset. The results can be seen below. (Table 31 and figures 20 to 22)

Statistics

		Electric Energy (MJ)	Gas Energy (MJ)	Total energy (MJ)
N	Valid	133	130	130
	Missing	0	3	3
Mean		334383.58	480498.50	806376.25
Median		151077.13	300551.12	454134.56
Std. Deviation		357075.572	498873.098	799556.626
Variance		1.275E11	2.489E11	6.393E11
Skewness		1.121	1.239	1.139
Std. Error of Skewness		.210	.212	.212
Kurtosis		-.254	.152	-.021
Std. Error of Kurtosis		.417	.422	.422
Minimum		18214	22247	61193
Maximum		1033006	1550050	2583056

Table 31: Statistics for energy usage in offices sub-sector

From the statistics table, it is clear that there is high difference between values of mean and median in each of the energy types. This indicates that the distribution is asymmetric. The histograms for each of the energy types also show that the distributions of the variables are different from normal distribution which is shown with black line. This issue means that the regression analysis might not work for the prediction of these variables.

The multiple regression analysis is run for the buildings in the offices sub-sector. The final equations for the energy usage in all energy types cannot provide us with satisfactory results and the error percentage is not acceptable. However, the important predictors in energy usage can be recognized.

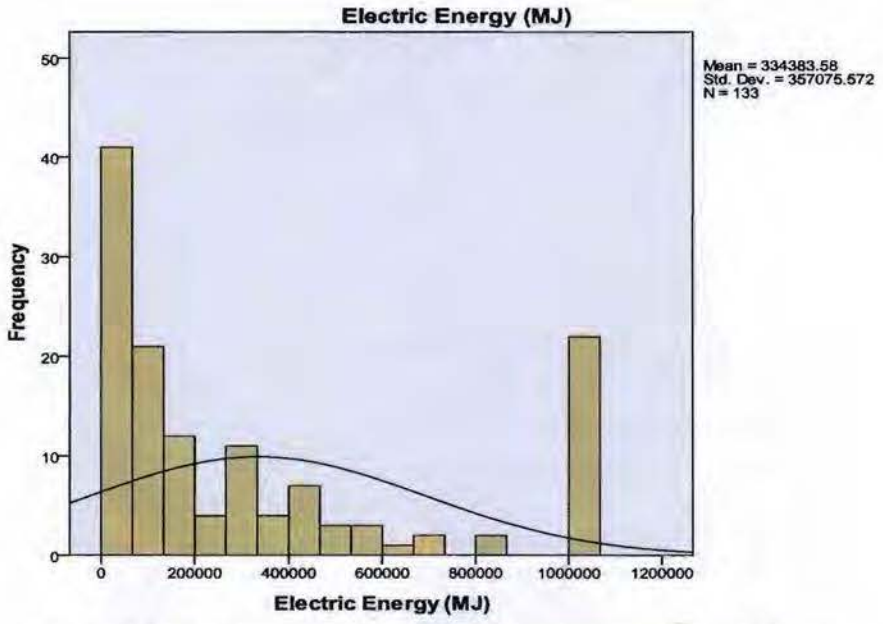

Figure 20: Histogram for electric energy usage in offices sub-sector

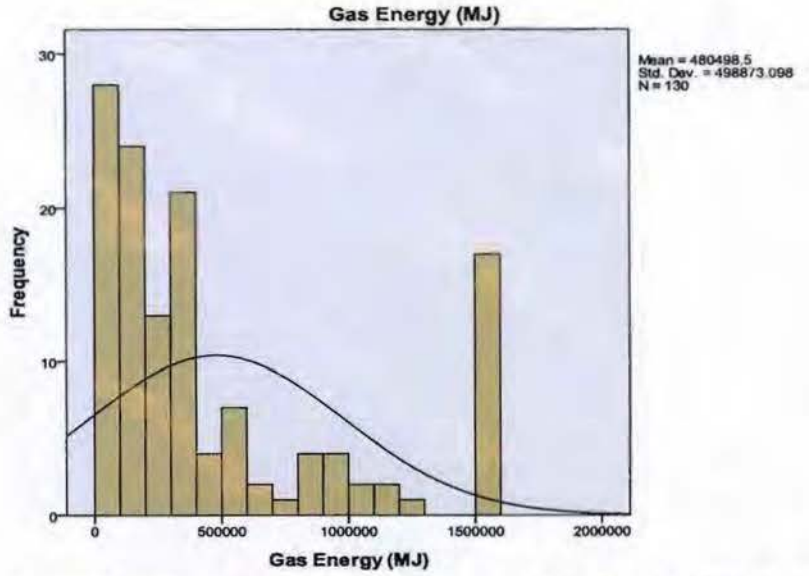

Figure 21: Histogram for gas energy usage in offices sub-sector

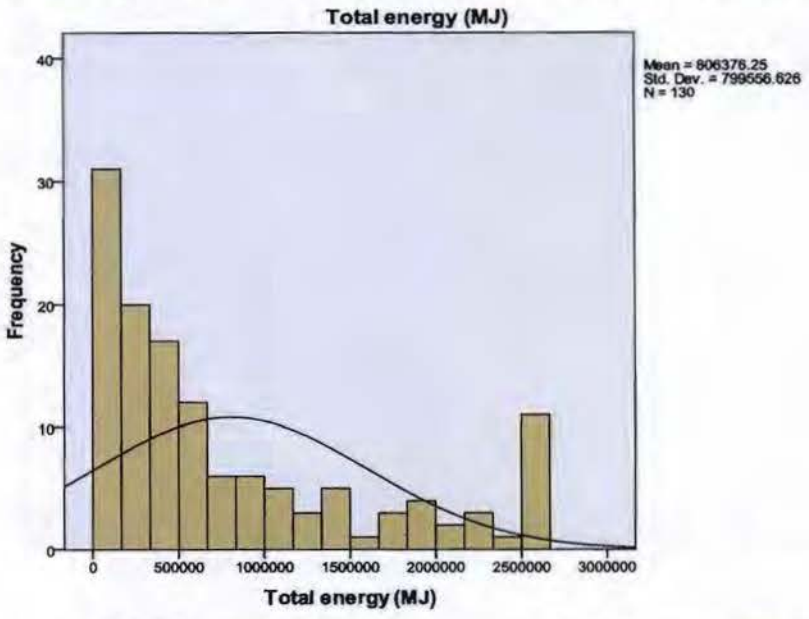

Figure 22: Histogram for total energy usage in offices sub-sector

For the electric energy usage, the total area of the building and the number of employees of the company are the most important predictors. When these variables have higher amounts, the electric energy usage also will have higher value.

Regarding the gas energy usage, the same predictors have the most important roles with also the positive influence on the energy usage. The total energy also can be predicted with the same predictors and with the same positive effects. The relevant tables have been included in the appendixes in part D: Sub-sectors regression analysis, 2-Offices sub-sector.

To sum up, we could say that the total area of the building and the number of employees of the company are the most important variables which can define the energy usage behavior

of a building in the offices sub-sector. Both of these variables have the positive effects which it means that by increasing them the energy usage will consequently be increased.

4.5.1.4. Energy usage predcition in education sub-sector

The education sub-sector has only 8 buildings. The frequencies analysis results for this sub-sector are pictured below.

Statistics

		Electric Energy (MJ)	Gas Energy (MJ)	Total energy (MJ)
N	Valid	8	8	8
	Missing	0	0	0
Mean		502066.02	489371.24	991437.26
Median		344973.94	384179.73	749066.14
Std. Deviation		379621.172	468901.511	792752.854
Variance		1.441E11	2.199E11	6.285E11
Skewness		.557	1.997	1.154
Std. Error of Skewness		.752	.752	.752
Kurtosis		-1.914	4.403	.605
Std. Error of Kurtosis		1.481	1.481	1.481
Minimum		133513	87272	277830
Maximum		1033006	1550050	2515107

Table 32: Statistics for energy usage in education sub-sector

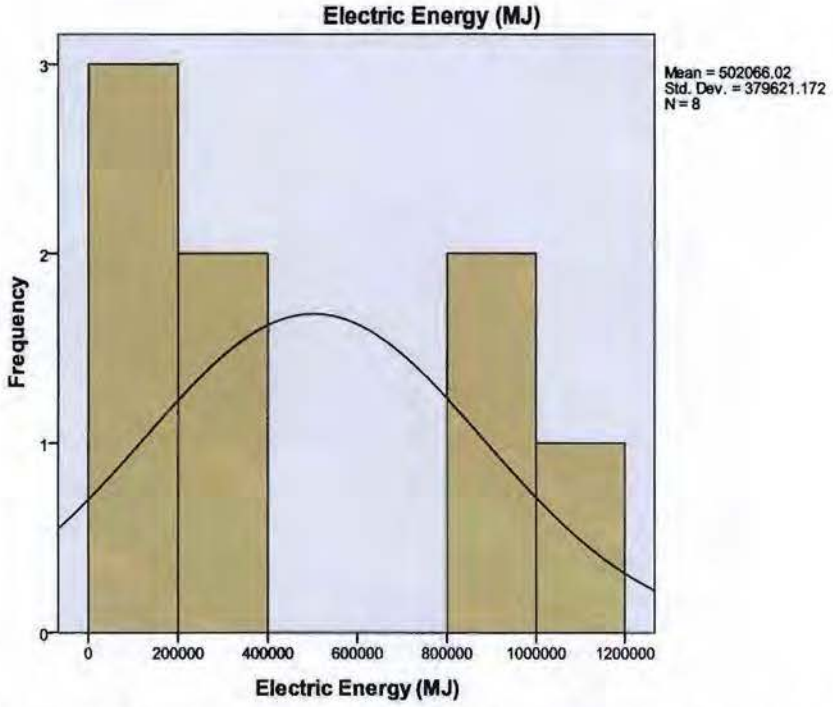

Figure 23: Histogram for electric energy usage in education sub-sector

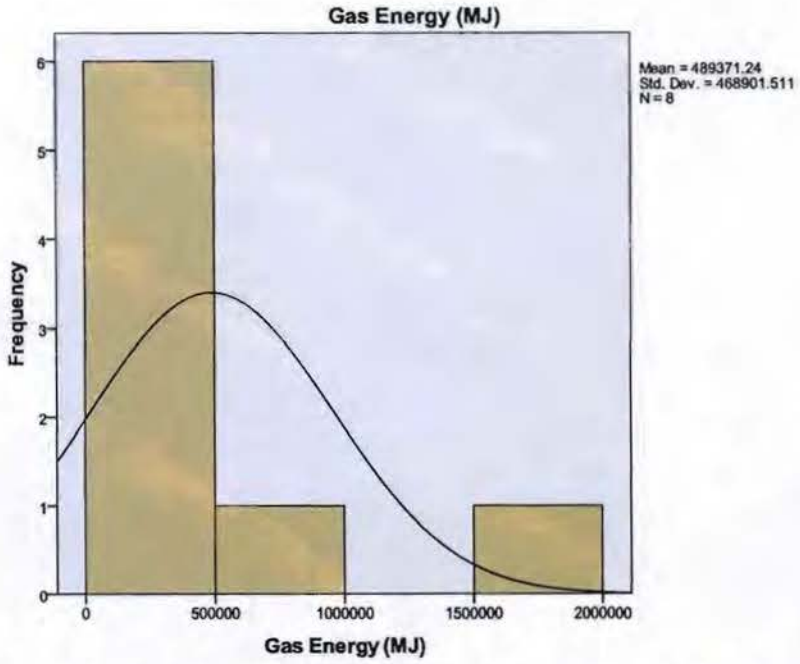

Figure 24: Histogram for gas energy usage in education sub-sector

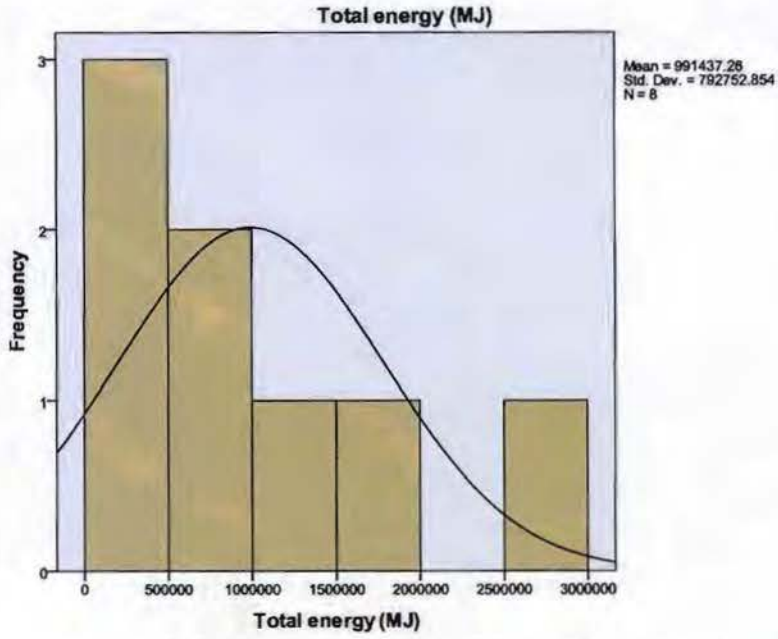

Figure 25: Histogram for total energy usage in education sub-sector

The big differences between mean and median for all of the energy types show that the distributions are asymmetric. The graphs also prove this fact.

The multiple regression analysis is preformed on this sub-sector. The results are not satisfactory enough in terms of the prediction formula. However, the important predictors can be derived only for the electric energy.

The total area of the building is the most important predictor for the electric energy. This predictor has the positive effect of the electric energy and when it is increased the electric energy is also higher.

The regression analysis has no results for the gas energy usage and total energy usage and this indicates the non-uniformity distribution of these variables. The relevant tables have been included in the appendixes in part D: Sub-sectors regression analysis, 3- Education sub-sector.

4.5.1.5. Energy usage predcition in health sub-sector

For the health sub-sector, there are 8 buildings. The statistical analysis table and figures are included below. (Table 33 and figures 26 to 28)

Statistics

		Electric Energy (MJ)	Gas Energy (MJ)	Total energy (MJ)
N	Valid	8	8	8
	Missing	0	0	0
Mean		303299.79	643023.90	946323.70
Median		146707.25	458849.56	605556.82
Std. Deviation		351093.133	514323.944	800033.199
Variance		1.233E11	2.645E11	6.401E11
Skewness		1.738	1.183	1.410
Std. Error of Skewness		.752	.752	.752
Kurtosis		2.037	-.098	1.593
Std. Error of Kurtosis		1.481	1.481	1.481
Minimum		81387	204489	285876
Maximum		1033006	1550050	2583056

Table 33: Statistics for energy usage in health sub-sector

From the table 33, it is obvious that the mean and median values differ significantly. So the distribution of the data is asymmetric. The visualizations of the distributions with the histograms also indicate the same aspect of the energy usage variables for the health sub-sector. It means that the regression analysis might have not desired outcomes for the prediction purposes.

Running the regression analysis for this sub-sector shows that the total area of the building is the most important predictor in all of the energy types. However, for the gas energy the building age is important as the second predictor. All of the predictors have the positive influence on the energy usage which means that increasing them leads to the higher energy usage.

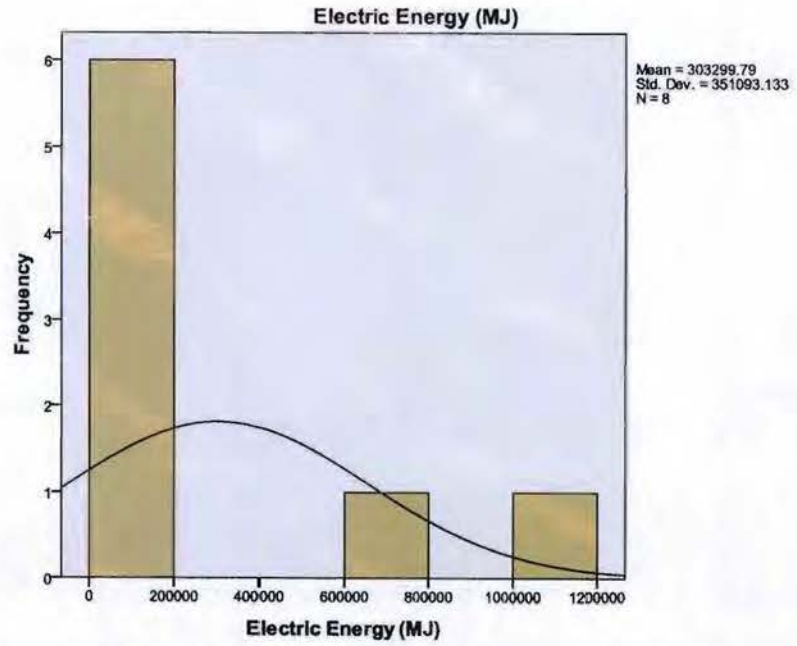

Figure 26: Histogram for electric energy usage in health sub-sector

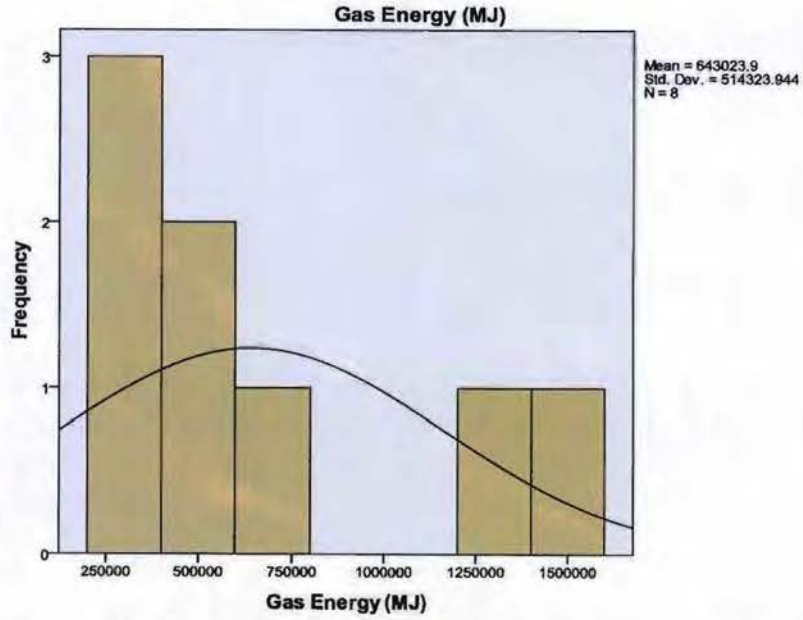

Figure 27: Histogram for gas energy usage in health sub-sector

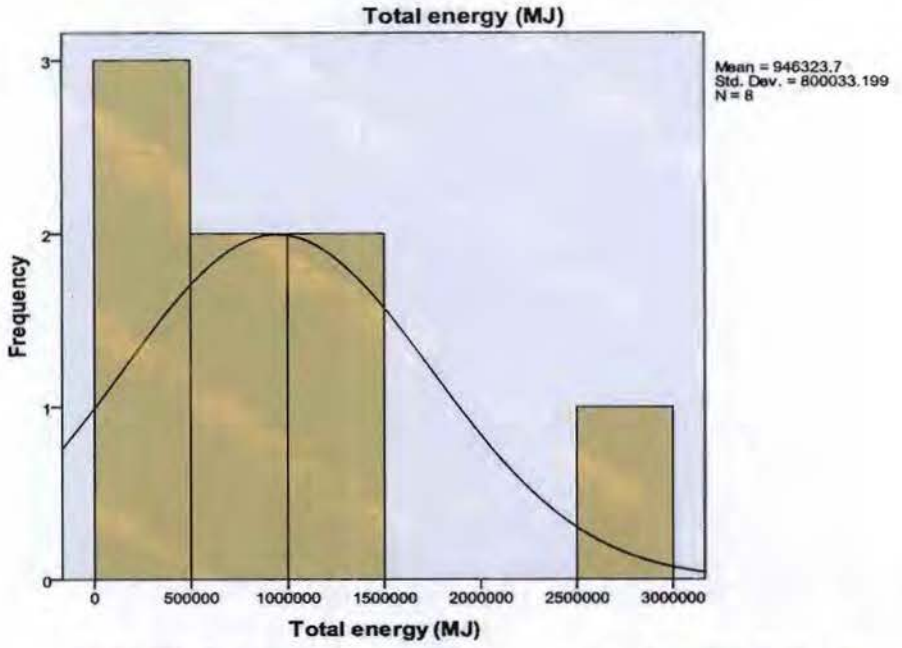

Figure 28: Histogram for total energy usage in health sub-sector

The relevant tables for the regression analysis have been included in the appendixes in part D: Sub-sectors regression analysis, 4- Health sub-sector.

4.5.1.6. Energy usage predcition in non-office based services sub-sector

In this sub-sector there are 26 buildings in total. The results of the frequencies analysis for them can be seen below. As the table and the figures show, the distribution is far from normal again.

Statistics

		Electric Energy (MJ)	Gas Energy (MJ)	Total energy (MJ)
N	Valid	26	24	24
	Missing	0	2	2
Mean		206937.65	438071.39	599480.59
Median		95316.44	338708.77	451597.52
Std. Deviation		259544.155	427137.996	595006.465
Variance		6.736E10	1.824E11	3.540E11
Skewness		2.143	1.556	1.880
Std. Error of Skewness		.456	.472	.472
Kurtosis		4.398	2.187	3.689
Std. Error of Kurtosis		.887	.918	.918
Minimum		17899	42246	61658
Maximum		1033006	1550050	2432991

Table 34: Statistics for energy usage in non-office based services sub-sector

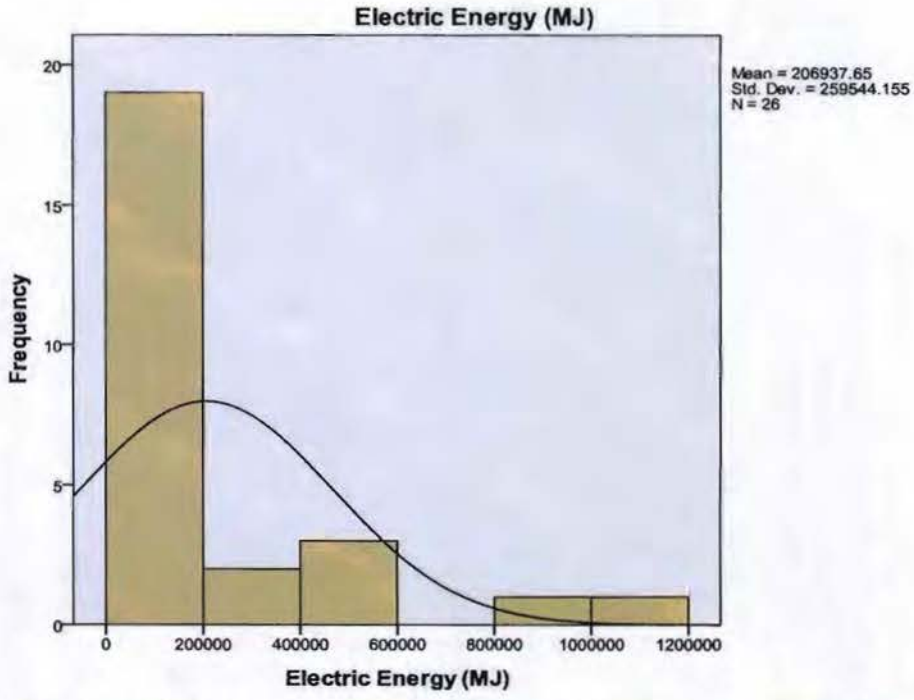

Figure 29: Histogram for electric energy usage in non-office based services sub-sector

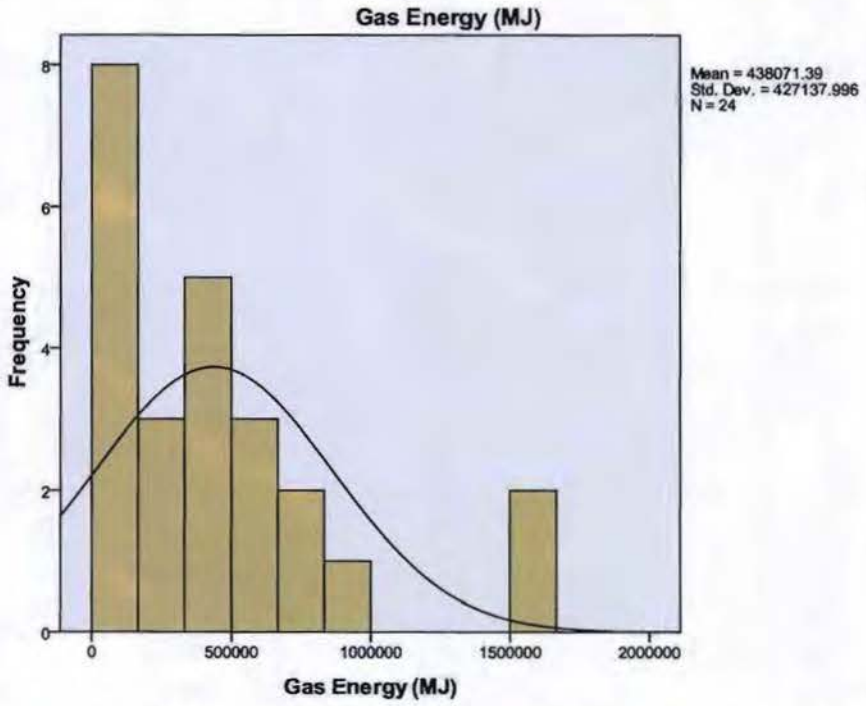

Figure 30: Histogram for gas energy usage in non-office based services sub-sector

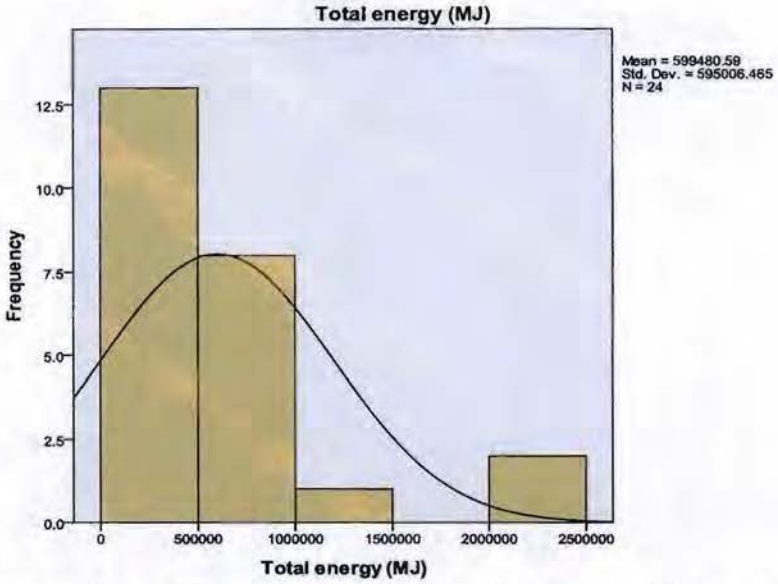

Figure 31: Histogram for total energy usage in non-office based services sub-sector

The results of the multiple regression analysis show that the number of employees is the most important predictor for all the energy types. The total area of the building is entered into the analysis for the gas energy prediction as the second predictor. Both of these predictors influence the energy usage in a positive way. Indeed, we can see that the energy usage in this sub-sector is more dependent on the number of the users of the energy in each building.

The relevant tables for the regression analysis have been included in the appendixes in part D: Sub-sectors regression analysis, 5- Non-office based services sub-sector.

4.5.2. Electric energy usage prediction

In this section the electric energy usage is considered for the prediction analysis. First we describe the clustering method for prediction and then the prediction results are shown.

4.5.2.1. Prediction clustering of electric energy usage

For clustering the buildings, by considering the energy usage as the dependent variable in the prediction, after several tests with different variables, a new variable has been defined in our dataset. This variable is related to the electric energy usage and the scale of the building (= total area * number of employees).

We have chosen the scale of the building on the basis of the high correlation coefficient to the electric energy consumption. The table 35 is the result of the correlation analysis for these two variables.

Correlations

		Electric Energy (scaled MJ)	Scale of building
Electric Energy (scaled MJ)	Pearson Correlation	1	.754**
	Sig. (2-tailed)		.000
	N	364	364
Scale of building	Pearson Correlation	.754**	1
	Sig. (2-tailed)	.000	
	N	364	364

**. Correlation is significant at the 0.01 level (2-tailed).

Table 35: Correlations between electric energy and scale of building

For the two-step clustering method, we have used "Electric energy / Scale of the building (E/SB)" as the continuous variable for classification of the buildings and using in the clustering.

On the basis of this clustering analysis, 20 clusters have been defined. There are in total 353 buildings in these clusters. For each of these clusters the multiple linear regression analysis has been implemented and the prediction equations have been defined. The results of the cluster analysis and regression analysis for the prediction of the electric energy can be seen in table 36.

The table shows the number of buildings in each cluster, mean amount of Electric energy / Scale of the building for each cluster, the R and R square from the regression analysis and finally the electric energy prediction equation for each cluster.

Scale of the building is derived here as the main predictor of the energy usage for all of the clusters. It is obvious from the table that when the building has higher total area and the company has more employees, then the electric energy usage is also higher.

Also some other variables have been included in the equations through the regression analysis. Building age, height and brick façade are also important predictors for prediction in some clusters.

However, the results of the analysis show that by having the scale of the building a good estimation of the electric energy usage can be derived. In the next section a detailed description for the prediction equation is depicted for a random cluster.

Cluster Number	Clustering information		Linear Regression information		
	Number of buildings	Mean Electric energy/ Scale of building	R	R square	Prediction equation for Electric energy
1	8	3.29	0.99	0.981	Electric energy = 1.951 + 2.336 * Scale of building
2	16	1.16	0.999	0.997	Electric energy = 0.182 + 1.15 * Scale of building
3	9	1.48	0.999	0.999	Electric energy = - 3.852 + 1.492 * Scale of building + 0.275 * Building age
4	11	1.98	0.998	0.996	Electric energy = 0.015 + 1.956 * Scale of building
5	19	0.76	0.999	0.998	Electric energy = - 4.759 + 0.79 * Scale of building + 0.708 * Height
6	15	0.94	0.999	0.998	Electric energy = - 0.547 + 0.966 * Scale of building
7	17	0.47	1	1	Electric energy = - 0.352 + 0.47 * Scale of building
8	21	0.53	0.999	0.997	Electric energy = 0.651 + 0.516 * Scale of building
9	21	0.62	0.998	0.996	Electric energy = - 0.666 + 0.64 * Scale of building
10	22	0.1	0.966	0.934	Electric energy = 0.098 + 0.096 * Scale of building
11	17	0.14	0.996	0.991	Electric energy = 0.591 + 0.130 * Scale of building
12	13	0.17	0.999	0.998	Electric energy = 0.158 + 0.164 * Scale of building
13	25	0.2	0.999	0.998	Electric energy = -0.291 + 0.207 * Scale of building
14	22	0.26	0.999	0.998	Electric energy = - 0.288 + 0.264 * Scale of building
15	20	0.23	1	0.999	Electric energy = - 0.183 + 0.232 * Scale of building
16	18	0.43	1	0.999	Electric energy = -0.098 + 0.423 * Scale of building
17	12	0.4	1	1	Electric energy = -1.341 + 0.390 * Scale of building + 0.248 * Height
18	27	0.36	1	1	Electric energy = - 0.397 + 0.372 * Scale of building
19	9	0.29	1	1	Electric energy = 0.411 + 0.282 * Scale of building
20	31	0.32	1	0.999	Electric energy = 2.682 + 0.316 * Scale of building + 1.578 * Brick Façade - 0.574 * Height
Total	**353**				

Table 36: Clustering and Prediction results for the electric energy

4.5.2.2. Prediction analyses in details for electric energy in cluster 5

For the electric energy prediction, we are going to describe the procedure for making the electric energy equation for cluster 5. In this cluster 19 buildings have been included after the cluster analysis. For these buildings the multiple linear regression analysis has been run in the SPSS software. The stepwise regression has been chosen to make the analysis clear and better. The outcome of the test can be seen in the following tables.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.999[a]	.998	.998	1.653110825
2	.999[b]	.998	.998	1.430519071

a. Predictors: (Constant), Scale of building

b. Predictors: (Constant), Scale of building, Height (cm)

c. Dependent Variable: Electric Energy (scaled MJ)

Table 37: Model summary of the regression analysis in cluster 5 for the electric energy

The model summary table reports the strength of the relationship between the model and the dependent variable. R, the multiple correlation coefficient, is the linear correlation between the observed and model-predicted values of the dependent variable. Its large value indicates a strong relationship. As you can see R is close to 1 and it shows the strong relationship.

In addition, R Square, the coefficient of determination, is the squared value of the multiple correlation coefficient. It shows that almost all of the variation in time is explained by the model. Because of the step-wise regression analysis we can see 2 models in the model summary table. It means that two variables are considered for the electric energy prediction.

The ANOVA table reports a significant F statistic, indicating that using the model is better than guessing the mean. As we can see in table 38, the significance is high and the F value is also high which shows the model works acceptable.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15995.784	1	15995.784	5853.311	.000[a]
	Residual	35.526	13	2.733		
	Total	16031.310	14			
2	Regression	16006.754	2	8003.377	3910.983	.000[b]
	Residual	24.557	12	2.046		
	Total	16031.310	14			

a. Predictors: (Constant), Scale of building

b. Predictors: (Constant), Scale of building, Height (cm)

c. Dependent Variable: Electric Energy (scaled MJ)

Table 38: ANOVA table of the regression analysis in cluster 20 for the electric energy

The other table is coefficients table. From the coefficients table, we can see that scale of building and heights have been chosen for the predictors. Also, the constant value has been proposed for the prediction.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-.842	.609		-1.383	.190					
	Scale of building	.791	.010	.999	76.507	.000	.999	.999	.999	1.000	1.000
2	(Constant)	-4.759	1.772		-2.686	.020					
	Scale of building	.790	.009	.998	88.275	.000	.999	.999	.997	.999	1.001
	Height (cm)	.708	.306	.026	2.315	.039	.061	.556	.026	.999	1.001

a. Dependent Variable: Electric Energy (scaled MJ)

Table 39: Coefficients table of the regression analysis in cluster 20 for the electric energy

We can make the prediction formula by using the unstandardized coefficients (Beta) from the coefficients table. This table indicates that the electric energy usage for this cluster follows the equation below:

*Electric energy = - 4.759 + 0.79 * Scale of building + 0.708 * Height* (5)

By using the above formula, electric energy usage can be predicted for the buildings in this cluster.

4.5.3. Gas energy usage prediction

The gas energy usage prediction is investigated in this part. Firstly, the cluster analysis is performed on the dataset to make appropriate clusters for the prediction and then the regression analysis is implemented for the prediction purposes.

4.5.3.1. Prediction clustering of gas energy usage

As it was mentioned in the previous part, scale of building is going to be used for the cluster analysis. This variable has high correlation coefficient with the gas energy usage as it is shown in table 40.

For the two step clustering method, we have used "Gas energy / Scale of the building (G/SB)" as the continuous variable for classification of the buildings. This issue has been derived from various tests with different combinations of the variables for the clustering.

Correlations

		Gas Energy (Scaled MJ)	Scale of building
Gas Energy (Scaled MJ)	Pearson Correlation	1	.757**
	Sig. (2-tailed)		.000
	N	364	364
Scale of building	Pearson Correlation	.757**	1
	Sig. (2-tailed)	.000	
	N	364	364

**. Correlation is significant at the 0.01 level (2-tailed).

Table 40: Correlations between gas energy and scale of building

After performing the cluster analysis on the entire dataset, 19 clusters have been defined for the gas energy prediction which with them the best results can be concluded. There are 363 buildings in total in these clusters. Then, the multiple regression analysis is run on these clusters. Table 41 shows the summary of the clustering analysis information and also the regression analysis results for the gas energy prediction in all the defined clusters.

The table shows the number of buildings in each cluster, mean amount of Gas energy / Scale of the building for each cluster, the R and R square from the regression analysis and at last the gas energy usage equation for each cluster.

The main predictor for the gas energy usage is scale of the building as we expected before. By having this independent variable, we can estimate the gas energy usage appropriately. On the other hand, other variables such as single building location, brick façade, glass percentage, open surrounding, height and building age also have been included in some of the gas energy usage equations. It means that in each cluster different kinds of predictors are dominant.

The next part describes the details of analysis for one cluster of the gas energy usage which is chosen randomly.

Cluster Number	Clustering information		Linear Regression information		
	Number of buildings	Mean Gas energy/ Scale of building	R	R square	Prediction equation for Gas energy
1	4	7.03	1	1	Gas Energy = - 17.099 + 7.733 * Scale of building + 15.180 * Single building location + 0.575 * Brick Façade
2	4	5.48	1	1	Gas Energy = 6.637 + 4.795 * Scale of building - 0.133 * Glass percentage
3	11	3.99	0.998	0.995	Gas Energy = - 0.888 + 4.178 * Scale of building
4	12	2.78	0.996	0.991	Gas Energy = - 16.712 + 2.798 * Scale of building + 2.923 * Height
5	22	1.96	0.988	0.975	Gas Energy = 2.424 + 1.844 * Scale of building
6	18	0.1	0.974	0.949	Gas Energy = - 1.457 + 0.126 * Scale of building
7	24	0.2	0.988	0.977	Gas Energy = 0.850 + 0.185 * Scale of building
8	26	0.29	0.993	0.987	Gas Energy = - 0.885 + 0.294 * Scale of building
9	20	0.59	1	0.999	Gas Energy = 0.446 + 0.588 * Scale of building
10	23	0.55	1	0.999	Gas Energy = -0.687 + 0.564 * Scale of building
11	38	0.49	0.999	0.999	Gas Energy = 0.624 + 0.472 * Scale of building
12	25	0.39	0.997	0.995	Gas Energy = - 2.862 + 0.363 * Scale of building + 0.316 * Building age
13	34	0.68	0.999	0.998	Gas Energy = 2.224 + 0.688 * Scale of building - 0.074 * Glass percentage
14	19	0.73	1	0.999	Gas Energy = 0.147 + 0.726 * Scale of building
15	19	0.82	0.999	0.998	Gas Energy = 0.405 + 0.805 * Scale of building
16	10	0.94	1	1	Gas Energy = 0.364 + 0.934 * Scale of building
17	17	1.05	0.999	0.997	Gas Energy = 1.765 + 1.006 * Scale of building
18	15	1.21	0.998	0.997	Gas Energy = 0.453 + 1.192 * Scale of building
19	22	1.45	0.999	0.998	Gas Energy = 0.341 + 1.451 * Scale of building - 2.289 * Open surrounding
Total	**363**				

Table 41: Clustering and Prediction results for the gas energy

4.5.3.2. Prediction analysis in details for gas energy in cluster 12

The cluster 12 is chosen for describing the details about the multiple regression analysis for finding the gas energy usage equation. In this cluster there are 25 buildings. The outcome of the regression analysis can be seen in the tables below.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.996[a]	.992	.991	5.900046023
2	.997[b]	.995	.994	4.757667012

a. Predictors: (Constant), Scale of building

b. Predictors: (Constant), Scale of building, Building Age

c. Dependent Variable: Gas Energy (Scaled MJ)

Table 42: Model summary of the regression analysis in cluster 12 for the gas energy

The high amounts of R, R square and adjusted R square in the model summary table show the high ability of the regression analysis in the prediction of the gas energy for this cluster. Also, the ANOVA table is shown in table 43. The high value of F shows the importance of the prediction.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	61750.094	1	61750.094	1773.891	.000[a]
	Residual	522.158	15	34.811		
	Total	62272.253	16			
2	Regression	61955.357	2	30977.678	1368.550	.000[b]
	Residual	316.896	14	22.635		
	Total	62272.253	16			

a. Predictors: (Constant), Scale of building

b. Predictors: (Constant), Scale of building, Building Age

c. Dependent Variable: Gas Energy (Scaled MJ)

Table 43: ANOVA table of the regression analysis in cluster 12 for the gas energy

The coefficients table is shown below too. As it is obvious from table 44, scale of the building and the building age are the important predictors for this cluster in terms of the gas energy prediction. The constant is also proposed by the model for the equation.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1.452	1.924		.755	.462					
	Scale of building	.372	.009	.996	42.118	.000	.996	.996	.996	1.000	1.000
2	(Constant)	-2.862	2.112		-1.355	.197					
	Scale of building	.363	.008	.972	47.221	.000	.996	.997	.900	.857	1.167
	Building Age	.316	.105	.062	3.011	.009	.429	.627	.057	.857	1.167

a. Dependent Variable: Gas Energy (Scaled MJ)

Table 44: Coefficients table of the regression analysis in cluster 12 for the gas energy

We can make the prediction formula by using the unstandardized coefficients (Beta) from the coefficients table. According to this table, the gas energy usage for this cluster follows the equation below:

*Gas Energy = - 2.862 + 0.363 * Scale of building + 0.316 * Building age* (6)

By using the above formula, one can estimate the gas energy usage for the buildings in this cluster with high accuracy.

4.5.4. Total energy usage prediction

The last part of the energy prediction is the total energy prediction. In this part the cluster analysis is run first to define the clusters of the buildings which are suitable for the prediction. Then the linear regression is used for the total energy prediction.

4.5.4.1. Prediction clustering of total energy usage

Before the beginning of the prediction analysis, the cluster analysis is done by using the scale of the building variable. The high correlation between this independent variable and the total energy is indicated in table 45.

Correlations

		Total energy (scaled) MJ	Scale of building
Total energy (scaled) MJ	Pearson Correlation	1	.805**
	Sig. (2-tailed)		.000
	N	364	364
Scale of building	Pearson Correlation	.805**	1
	Sig. (2-tailed)	.000	
	N	364	364

**. Correlation is significant at the 0.01 level (2-tailed).

Table 45: Correlations between total energy and scale of building

For the two step clustering method, after testing different variables for the clustering, we have used "Total energy / Scale of the building (T/SB)" as the variable as the continuous variable for classification of the buildings.

After the cluster analysis on the entire dataset, 21 clusters have been defined for the total energy prediction. There are 363 buildings in total in these clusters. The multiple linear regression method is used in each of these clusters for the prediction of the total energy usage of the buildings. Table 46 shows the summary of the clustering analysis information and also the regression analysis results for the total energy in different clusters.

The table shows the number of buildings in each cluster, mean amount of Total energy / Scale of the building for each cluster, the R and R square from the regression analysis and at last the total energy usage equation for each cluster.

As it could be seen in table 46, scale of the building is the most important predictor as we expected before. Height, building age, glass percentage and brick façade are the other predictors that are needed in some clusters for better estimation of the total energy usage.

Cluster Number	Clustering information		Linear Regression information		
	Number of buildings	Mean Total energy/ Scale of building	R	R square	Prediction equation for total energy
1	17	3.14	0.996	0.992	Total Energy = -21.035 + 3.459 * Scale of building + 2.963 * Height
2	26	2.44	0.994	0.989	Total Energy = 3.889 + 2.279 * Scale of building
3	9	4.49	0.996	0.992	Total Energy = 1.511 + 4.381 * Scale of building
4	8	5.67	0.998	0.996	Total Energy = 1.362 + 5.473 * Scale of building
5	9	7.91	1	0.999	Total Energy = 0.608 + 7.728 * Scale of building
6	4	6.43	1	1	Total Energy = 0.286 + 6.362 * Scale of building
7	22	1.61	0.998	0.997	Total Energy = 0.845 + 1.586 * Scale of building
8	20	1.93	0.997	0.994	Total Energy = 2.614 + 1.997 * Scale of building - 0.198 * Building age
9	31	1.06	0.999	0.998	Total Energy = 0.286 + 1.048 * Scale of building
10	22	1.16	1	0.999	Total Energy = -2.248 + 1.161 * Scale of building + 0.106 * Building age
11	15	1.26	0.999	0.999	Total Energy = - 0.452 + 1.278 * Scale of building
12	16	1.39	1	1	Total Energy = 0.571 + 1.348 * Scale of building + 2.087 * Brick Façade
13	15	0.49	0.996	0.992	Total Energy = - 2.183 + 0.506 * Scale of building
14	27	0.57	0.999	0.999	Total Energy = - 5.759 + 0.583 * Scale of building + 0.129 * Glass percentage
15	15	0.65	0.999	0.999	Total Energy = - 0.161 + 0.656 * Scale of building
16	11	0.19	0.983	0.966	Total Energy = - 0.133 + 0.201 * Scale of building
17	18	0.36	0.978	0.956	Total Energy = 0.269 + 0.351 * Scale of building
18	29	0.85	0.999	0.999	Total Energy = 1.282 + 0.828 * Scale of building
19	17	0.95	0.999	0.999	Total Energy = 0.476 + 0.943 * Scale of building
20	10	0.71	1	1	Total Energy = - 0.285 + 0.712 * Scale of building
21	22	0.77	0.999	0.999	Total Energy = - 1.390 + 0.783 * Scale of building
Total	**363**				

Table 46: Clustering and Prediction results for the total energy

4.5.4.2. Prediction analysis in details for total energy in cluster 10

The cluster 10 is chosen randomly for describing the details about the multiple regression analysis for finding the total energy usage equation. In this cluster there are 22 buildings. The outcome of the regression analysis can be seen in the tables below.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000[a]	.999	.999	3.098822380
2	1.000[b]	.999	.999	2.786782681

a. Predictors: (Constant), Scale of building

b. Predictors: (Constant), Scale of building, Building Age

c. Dependent Variable: Total energy (scaled) MJ

Table 47: Model summary of the regression analysis in cluster 10 for the total energy

In the model summary table, the high amounts of R, R square and adjusted R square show the high ability of the regression analysis in the prediction of the gas energy for this cluster. Also, the ANOVA table is shown in table 48. The high value of F shows the importance of the prediction.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	174601.866	1	174601.866	18182.580	.000[a]
	Residual	153.643	16	9.603		
	Total	174755.510	17			
2	Regression	174639.017	2	87319.509	11243.592	.000[b]
	Residual	116.492	15	7.766		
	Total	174755.510	17			

a. Predictors: (Constant), Scale of building

b. Predictors: (Constant), Scale of building, Building Age

c. Dependent Variable: Total energy (scaled) MJ

Table 48: ANOVA table of the regression analysis in cluster 10 for the total energy

The coefficients table is shown below too (table 49). As it is obvious from this table, scale of the building and the building age are the important predictors for this cluster in terms of the total energy prediction. The constant is also proposed by the model for the equation.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-.134	1.161		-.115	.910					
	Scale of building	1.167	.009	1.000	134.843	.000	1.000	1.000	1.000	1.000	1.000
2	(Constant)	-2.248	1.423		-1.580	.135					
	Scale of building	1.161	.008	.995	142.495	.000	1.000	1.000	.950	.911	1.097
	Building Age	.106	.048	.015	2.187	.045	.311	.492	.015	.911	1.097

a. Dependent Variable: Total energy (scaled) MJ

Table 49: Coefficients table of the regression analysis in cluster 10 for the total energy

We can make the prediction formula by using the unstandardized coefficients (Beta) from the coefficients table. This table indicates that the gas energy usage for this cluster follows the equation below:

*Total Energy = - 2.248 + 1.161 * Scale of building + 0.106 * Building age* (7)

By using the above formula, one can predict the total energy usage for the buildings in this cluster with high accuracy.

4.6. Conclusions

In this chapter we have done the analysis for prediction of the energy usage in the buildings of the services sector for city of Eindhoven. Firstly, a short theoretical description has been written down about the linear regression method. Then each of the sub sectors has been assessed to find out about the energy prediction equations and important predictors for energy usage. However, the results were not satisfactory enough because of the asymmetric distribution of the data.

Then we decided to divide the buildings in to the clusters. The prediction equations have been derived for different buildings in various clusters. Also, the important variables have been identified as the predictors of the energy usage. In total, we have divided the buildings to some clusters for each of electric energy, gas energy and total energy. In each of these clusters one equation governs the energy usage for the buildings.

The results of the analysis show that the scale of the building is the most important predictor for every types of energy usage. Scale of the building consists of the total area of the building and the number of employees of the company. The scale of the building and company highly affects the energy usage of that specific building. When we have the higher scale of the building and company, we should expect higher energy usage for that building.

In some equations, the other predictors also have been used. Considering the electric energy usage, building age and height of the building are also important predictors in some clusters.

When the building is older, the electric energy usage is also higher. Also, when the building is higher, the electric energy usage is more in some cases.

On the other hand, regarding the gas energy usage, surrounding type, façade type, height of the building, building age and glass percentage are also important predictors in some clusters.

For the total energy usage, height, building age, brick façade and glass percentage are important predictors in some clusters. The equations show that with higher amount for these independent variables, higher total energy will come.

To sum up, the results of analysis also confirm that the scale of the building in all the clusters and age of the building in some clusters are important predictors for the energy usage, as we expected before. Other variables also help us to make better estimations about the energy performance of a building.

Chapter 5: Conclusions, discussions and recommendations

In this chapter, the conclusions of the report are described. Also, the discussions and recommendations parts have been written down.

5.1. Conclusions

The main focus of this report was the energy usage in the services sector of the city of Eindhoven. Services sector contains some sub sectors such as retail, offices and etc. In the following sections the main conclusions for different steps of analysis are depicted.

5.1.1. Creating the dataset

As the beginning part of the research, the literature study was done about the available information and data for the energy usage of the buildings in the services sector. However, the existing data was not satisfactory and enough. Thus, we decided to make our own dataset.

In creating the dataset, several sources have been investigated and used. We tried to reach as much as possible information for the buildings which are important in terms of the energy usage. The building characteristics and the number of employees of the company were included in the dataset. Creating this dataset was time-consuming but it was inevitable as the initial step for this research and it can be also used for future studies.

The building related characteristics in the dataset are building age, height of the building, surrounding situations, location situations, façade types, glass percentage, surface area, number of floors and the full address of the building. Furthermore, the company name and the number of employees of the company for each building are included. The sector which the company is active in it is mentioned as well. In terms of the energy usage, each building has the information about the annual gas energy usage, electric energy usage and total energy usage, separately.

After making the dataset, data preparation steps were performed to make the data ready for the analysis. The outliers were detected and modified in the dataset. Also, the categorical variables had to be converted to the dummy coded variables. By doing these adjustments, the dataset was ready for the main statistical analyses of the research.

As it was mentioned in section 1.2, one of the research questions of this research was:

- Is it possible to make a dataset for the energy usage and related variables in the services sector?

The main conclusion that we could mention about this section will answer to this research question:

A dataset has been built which contains the building characteristics, company information and energy usage for the services sector in Eindhoven which includes 387 buildings in total.

The municipality of Eindhoven and researchers can consider this dataset as the basic step for further research and study about the energy usage of the services sector in the city of Eindhoven. By using this dataset, a clear view about the characteristics of the buildings, company information and the annual energy usage of the buildings are provided for the policy makers. However, this dataset is not complete enough. Some other variables could be added to this dataset which we will talk about them in the next sections in more details.

5.1.2. Cluster analysis

The other research question, which is mentioned in section 1.2., for this research is:

- How can we separate the buildings in the services sector to different clusters?

After creating the dataset, the cluster analysis was done on the dataset. The aim of this analysis was to divide the buildings in different clusters. The buildings in each cluster carry similar characteristics. The clusters have been made with different combinations of the variables. By this way, various perspectives have been achieved about the buildings. The energy usage amounts were the main focus of the clustering.

Two methods have been used in the clustering. The first method was k-means clustering method. By this method, two different kinds of clustering were done. The first one was on the basis of the extracted components from the Principal Components Analysis (PCA) method. By running the PCA method, three main components were extracted. These components were called energy usage and scale, surrounding type and building age and height. In the k-means clustering, 10 clusters were defined after running the cluster analysis with different k numbers. The clusters were labeled on the basis of different energy and total area levels of the buildings.

The other k-means clustering was with all the variables. Gas energy, electric energy and total area were the variables which finally we used in the clustering because of their higher importance compared to other variables. In this clustering, 8 clusters were defined. The energy and total area levels were identified for each cluster.

The other method of clustering, which was two step clustering method, was used to cluster the buildings through different combinations of the variables with the energy usage amounts. Seven different clustering were performed on the buildings. The buildings were put into different clusters in each of the analysis and the clusters were labeled according to their characteristics.

The main conclusion of this section can answer the above mentioned research question:

Different clusters of the buildings have been defined on the basis of the energy usage and the related variables. Each of these clusters provides different perspective for the buildings in the services sector of the Eindhoven. Two methods have been used which can offer better final results.

The municipality of Eindhoven can use the created clusters for study and assess the energy usage of the buildings. The different energy levels, which have been defined for the clusters, can help them to find out about the buildings which consume low, medium or high annual energy, separately for gas energy or electric energy. The reduction policies then can be defined specifically for each cluster of the buildings. Also, through the clustering of the energy with combination of different variables, the clear view is achieved for improving the performance of the buildings regarding the energy usage.

5.1.3. Prediction analysis

The prediction analysis was the other aspect of this research. We wanted to make the prediction equations for the energy usage and also recognize the important predictors of the energy in our dataset for the buildings in the services sector. The multiple linear regression analysis was used in this regard.

Two research questions, which we mentioned in section 1.2., are:

- What are the most important variables (predictors) for the energy usage of the buildings in the services sector?

- Could we define some equations for the energy usage of buildings in the services sector?

At first, the entire dataset was entered into the multiple regression analysis. However, the outcomes were not desirable and acceptable. Then, we decided to assess the buildings in each sub sector. Hence, the regression analysis was run for each sub sector, separately. The concluded results in this part also were not favorable enough but we found out about the important predictors in some sub sector.

Finally, we divided the buildings into different clusters, only for the purpose of prediction. Different variables were used for clustering and at last the scale of the building had the best results for the prediction analysis. Indeed, the buildings were grouped in different clusters on the basis of the "energy usage / scale of the building". For example, if we would like to find out about the formula for the gas energy usage, we use "gas energy usage / scale of the building" for the clustering the buildings. The results for the prediction in each cluster had high accuracy and they were acceptable. The important predictors were recognized and energy equations were defined for electric, gas and total energy separately in each cluster.

The main conclusion that we could mention about this section is:

The important predictors of the energy usage have been recognized for each sub sector and also for each cluster separately for gas energy usage, electric energy usage and total energy usage. Also, the energy usage equations have been derived for each of the energy types in each cluster.

Scale of the building is the most important predictor for all energy types. When the scale of the building is higher, the energy usage is higher consequently. Building age is also another important predictor. The same trend for the energy usage as the building age can be seen. Other variables like glass percentage, façade type, height of the building and surrounding type were important as well for prediction of the energy in some clusters.

By using the achieved conclusions, the municipality of Eindhoven can predict the energy usage of the buildings for the future. By this way, the policies can be made to reduce the energy usage for different buildings.

5.2. Discussions and recommendations

In this section we would like to discuss about the main limitations of the current research and also propose some recommendations for future research.

The first issue of the discussion is the data. We should mention that data gathering for this research was a hard and time-consuming task. Most of the organizations didn't response to the questions about the available data that they might have. The sources that we had to assess for the data gathering were not easy to reach. Also, combination of the available data into one single file was done manually which took lots of effort and time. Indeed, we could mention that the lack of data was the main limitation about this research.

Even after creating the dataset in the analysis the data didn't perform as we expected before. The reason was the asymmetric distribution of the data in our dataset. It means that the buildings in the services sector have different energy usage behavior, which don't follow the normal distribution, and it makes the prediction analysis hard.

In the distribution of the energy usage of the sub sectors, such as figure 14, at the end of the graph a high peak can be seen. These high peaks are in fact the outliers of the dataset which we had edited them before. Some buildings have extremely high energy usage. We tried to delete these peaks from the dataset and run the analysis again to maybe achieve better results. Unfortunately after deleting them, the results for the prediction were not better.

Furthermore, we have many outliers in our dataset. For example, considering the electric energy usage, the number of the outliers is equal to 48 (table 6) which is more than 10% of the total data. We tried to keep them in the dataset with the adjustments techniques. However, deleting them from the dataset didn't give us any better final results.

The effort should be made for adding more energy related variables to the created dataset. These variables can be the energy usage of the building in hourly basis, the user behavior of the buildings (such as the patterns of the energy usage for the employees) and the number of appliances of the company. Some more building characteristics also should be investigated for the services sector's buildings. Insulations of walls, floors, roofs, windows and pipes would be

important. Also, more buildings should be included in the dataset which will make the results of the analysis even better.

The clusters which we have made in this project for the prediction analysis should be investigated in details to find out why they have similar behavior in terms of the energy usage. Some specific buildings can be chosen from the dataset and from each cluster to check the energy usage management system of them in reality.

References

[1] Kasteren, van, H., Konz, W., van Schijndel, P., Smeets, R., Wentink, C., (December 2008), *"SOLET report: Energiek Brabant – Een scenariostudie naar de energievoorziening van Noord-Brabant in 2040"*.

[2] Kerssemeeckers, M., (July 2001), *"ICARUS – 4 , Sector study for the services sector"*, Ecofys Energy and Environment Utrecht

[3] Ang, B.W., (1995), *"Multilevel decomposition of industrial energy consumption"*, Energy Economics, 17 (1), pp. 39–51

[4] Ang, B.W., Lee, P.W., (1996), *"Decomposition of industrial energy consumption: the energy coefficient approach"*, Energy Economics, 18, pp. 129–143

[5] Gaglia, A. Balaras, C. Mirasgedis, S, Georgopoulou, E. Sarafidis, Y. Lalas, D., (2007), *"Empirical assessment of the Hellenic non-residential building stock, energy consumption, emissions and potential energy savings"*, Energy Conversion and Management 48, 1160–1175

[6] Choudhary, R, (2012), *"Energy analysis of the non-domestic building stock of Greater London"*, Building and Environment 51, 243-254

[7] Liu, C., (2006), *"A Study on Decomposition of Industry Energy Consumption"*, International Research Journal of Finance and Economics, Issue 6

[8] Pan, Y., Yin, R., Huang, Z., (2008) *"Energy modeling of two office buildings with data center for green building design"*, Energy and Buildings 40, 1145–1152

[9] van Loon, P., (March 2012), *"Target group clustering for applications of energy effective renovation concerning privately owned dwellings"*, CME master program Thesis, Eindhoven University of Technology

[10] http://www.endinet.nl/, Online, last access 5 June 2012

[11] http://www.kadaster.nl/BAG/bagviewer/, Online, last access 5 June 2012

[12] Grubbs, F. E., (1996), *"Procedures for detecting outlying observations in samples"*, Technometrics 11, 1–21.

[13] Hoaglin, D.C., Iglewicz, B., and Tukey, J.W., (1986), *"Performance of some resistant rules for outlier labeling"*, Journal of American Statistical Association, 81, 991-999.

[14] Hoaglin, D. C., and Iglewicz, B., (1987), *"Fine tuning some resistant rules for outlier labeling"*, Journal of American Statistical Association, 82, 1147-1149.

[15] David W. Stockburger, (1998), "*Multivariate statistics: concepts, models and applications*", Missouri State University

[16] Jain, A. K., Murty, M. N., Flynn, P. J., (1999), " *Data clustering: a review*", Journal of ACM computing surveys (CSUR), Volume 31, Issue 3, 264-323

[17] Weijermars, W., (2006), "*Analysis of urban traffic patterns using clustering*", Universiteit Twente, PhD Thesis

[18] PASW statistics 18 core system user's guide

[19] Bates, L.K., (2006), "*Does Neighborhood Really Matter? Comparing historically defined neighborhood boundaries with housing submarkets*", Journal of Planning Education and Research, 26, 5-17.

[20] Wu, C., Rashi Sharma, R., (2011), "Housing submarket classification: The role of spatial contiguity", Applied geography, 32, 746-56.

Appendixes

A. Data preparation: detecting outliers

1. Building Age:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Building Age	377	97.2%	11	2.8%	388	100.0%

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Building Age	5.00	5.80	11.00	18.00	35.00	44.20	54.10
Tukey's Hinges	Building Age			11.00	18.00	35.00		

Variable	Table used in SPSS	25 percentile = Q1	75 percentile = Q3	g	Lower Limit	Upper Limit
Building Age	Percentiles	11	35	1.5	0	71
				Number of modified cases =		3
		Q3-Q1 =	24			
		g'	36			

2. Height:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Height (cm)	338	87.1%	50	12.9%	388	100.0%

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Height (cm)	3.66	4.54	5.49	6.71	7.01	7.62	9.14
Tukey's Hinges	Height (cm)			5.49	6.71	7.01		

Variable	Table used in SPSS outcome	25 percentile = Q1	75 percentile = Q3	g	Lower Limit	Upper Limit
Height	Percentiles	5.49	7.01	1.5	3.21	9.29
				Number of modified cases =		16
		Q3-Q1 =	1.52			
		g'	2.28			

3. Glass percentage:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Glass Percentage	338	87.1%	50	12.9%	388	100.0%

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Glass Percentage	15.00	20.00	30.00	40.00	50.00	60.00	70.00
Tukey's Hinges	Glass Percentage			30.00	40.00	50.00		

Variable	Table used in SPSS outcome	25 percentile = Q1	75 percentile = Q3	g	Lower Limit	Upper Limit
Glass Percentage	Percentiles	30	50	1.5	0	80
		Q3-Q1 =	20	*Number of modified cases =*		4
		g'	30			

4. Total area:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Total Area (m2)	387	99.7%	1	.3%	388	100.0%

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Total Area (m2)	214.00	358.80	550.80	1118.00	2534.00	5047.92	8037.40
Tukey's Hinges	Total Area (m2)			550.80	1118.00	2531.50		

Variable	Table used in SPSS outcome	25 percentile = Q1	75 percentile = Q3	g	Lower Limit	Upper Limit
Total Area (M2)	Percentiles	550.8	2534	1.5	0	5508.8
		Q3-Q1 =	1983.2	*Number of modified cases =*		34
		g'	2974.8			

5. Electric energy:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Electric -Energy (KWH)	387	99.7%	1	.3%	388	100.0%

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Electric -Energy (KWH)	7095.34	9736.21	16834.77	41965.87	124879.44	354204.07	529799.32
Tukey's Hinges	Electric -Energy (KWH)			16912.15	41965.87	124564.33		

Variable	Table used in SPSS outcome	25 percentile = Q1	75 percentile = Q3	g	Lower Limit	Upper Limit
Electric Energy (KWH)	Percentiles	16834.77	124879.44	1.5	0	286946.445
				Number of modified cases =		48
		Q3-Q1 =	108044.67			
		g'	162067.005			

6. Gas energy:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gas Energy (M3)	376	96.9%	12	3.1%	388	100.0%

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Gas Energy (M3)	958.30	1356.89	2972.67	7542.41	16944.50	46165.04	97686.01
Tukey's Hinges	Gas Energy (M3)			2982.33	7542.41	16938.00		

Variable	Table used in SPSS outcome	25 percentile = Q1	75 percentile = Q3	g	Lower Limit	Upper Limit
Gas Energy (M3)	Percentiles	2972.67	16944.5	1.5	0	37902.245
				Number of modified cases =		44
		Q3-Q1 =	13971.83			
		g'	20957.745			

B. Principal Component Analysis

Correlation Matrix

		Building Age	Height (cm)	Building Block Surrounding	Open surrounding	Total Area (M2) scaled	Electirc Energy (scaled MJ)	Gas Energy (Scaled MJ)
Correlation	Building Age	1.000	-.322	-.021	.046	.110	.030	.174
	Height (cm)	-.322	1.000	-.127	.100	.291	.280	.227
	Building Block Surrounding	-.021	-.127	1.000	-.834	-.286	-.333	-.253
	Open surrounding	.046	.100	-.834	1.000	.170	.253	.158
	Total Area (M2) scaled	.110	.291	-.286	.170	1.000	.738	.776
	Electirc Energy (scaled MJ)	.030	.280	-.333	.253	.738	1.000	.760
	Gas Energy (Scaled MJ)	.174	.227	-.253	.158	.776	.760	1.000

Component Matrix[a]

	Component		
	1	2	3
Building Age	.088	.018	.872
Height (cm)	.393	.150	-.716
Building Block Surrounding	-.622	.727	.009
Open surrounding	.531	-.801	.005
Total Area (M2) scaled	.845	.344	.072
Electirc Energy (scaled MJ)	.863	.256	.010
Gas Energy (Scaled MJ)	.835	.368	.167

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Component Correlation Matrix

Component	1	2	3
1	1.000	-.261	-.076
2	-.261	1.000	.046
3	-.076	.046	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Component Score Covariance Matrix

Component	1	2	3
1	.997	-.479	1.944
2	-.479	1.058	-.552
3	1.944	-.552	2.929

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.
Component Scores.

C. Correlation analysis

Correlations							
		Electric Energy (scaled MJ)	Building Age	Height (cm)	Glass Percentage	Total Area (M2) with scaling	Group of employees
Electric Energy (scaled MJ)	Pearson Correlation	1	.028	.289**	.060	.733**	.499**
	Sig. (2-tailed)		.592	.000	.281	.000	.000
	N	376	366	331	327	376	376
Building Age	Pearson Correlation	.028	1	-.320**	.061	.095	-.024
	Sig. (2-tailed)	.592		.000	.279	.071	.643
	N	366	366	324	318	366	366
Height (cm)	Pearson Correlation	.289**	-.320**	1	-.007	.301**	.220**
	Sig. (2-tailed)	.000	.000		.908	.000	.000
	N	331	324	331	286	331	331
Glass Percentage	Pearson Correlation	.060	.061	-.007	1	.072	-.028
	Sig. (2-tailed)	.281	.279	.908		.193	.617
	N	327	318	286	327	327	327
Total Area (M2) with scaling	Pearson Correlation	.733**	.095	.301**	.072	1	.443**
	Sig. (2-tailed)	.000	.071	.000	.193		.000
	N	376	366	331	327	376	376
Group of employees	Pearson Correlation	.499**	-.024	.220**	-.028	.443**	1
	Sig. (2-tailed)	.000	.643	.000	.617	.000	
	N	376	366	331	327	376	376
**. Correlation is significant at the 0.01 level (2-tailed).							

Correlations								
		Electric Energy (scaled MJ)	Bewteen Buildings Location	Single Building Location	Building Block Surrounding	Open surrounding	Curtain wall Façade	Brick Façade
Electric Energy (scaled MJ)	Pearson Correlation	1	-.222**	.321**	-.335**	.250**	.017	-.046
	Sig. (2-tailed)		.000	.000	.000	.000	.754	.402
	N	376	375	375	375	375	327	327
Bewteen Buildings Location	Pearson Correlation	-.222**	1	-.459**	.253**	-.191**	.106	.117*
	Sig. (2-tailed)	.000		.000	.000	.000	.055	.034
	N	375	375	375	375	375	327	327
Single Building Location	Pearson Correlation	.321**	-.459**	1	-.251**	.156**	.006	-.088
	Sig. (2-tailed)	.000	.000		.000	.002	.910	.114
	N	375	375	375	375	375	327	327
Building Block Surrounding	Pearson Correlation	-.335**	.253**	-.251**	1	-.830**	.011	-.016
	Sig. (2-tailed)	.000	.000	.000		.000	.846	.780
	N	375	375	375	375	375	327	327
Open surrounding	Pearson Correlation	.250**	-.191**	.156**	-.830**	1	.028	.036
	Sig. (2-tailed)	.000	.000	.002	.000		.615	.512
	N	375	375	375	375	375	327	327
Curtain wall Façade	Pearson Correlation	.017	.106	.006	.011	.028	1	-.418**
	Sig. (2-tailed)	.754	.055	.910	.846	.615		.000
	N	327	327	327	327	327	327	327
Brick Façade	Pearson Correlation	-.046	.117*	-.088	-.016	.036	-.418**	1
	Sig. (2-tailed)	.402	.034	.114	.780	.512	.000	
	N	327	327	327	327	327	327	327
**. Correlation is significant at the 0.01 level (2-tailed).								
*. Correlation is significant at the 0.05 level (2-tailed).								

Correlations							
		Gas Energy (Scaled MJ)	Building Age	Height (cm)	Glass Percentage	Total Area (M2) with scaling	Group of employees
Gas Energy (Scaled MJ)	Pearson Correlation	1	.174**	.227**	.086	.776**	.454**
	Sig. (2-tailed)		.001	.000	.121	.000	.000
	N	376	366	331	327	376	376
Building Age	Pearson Correlation	.174**	1	-.320**	.061	.095	-.024
	Sig. (2-tailed)	.001		.000	.279	.071	.643
	N	366	366	324	318	366	366
Height (cm)	Pearson Correlation	.227**	-.320**	1	-.007	.301**	.220**
	Sig. (2-tailed)	.000	.000		.908	.000	.000
	N	331	324	331	286	331	331
Glass Percentage	Pearson Correlation	.086	.061	-.007	1	.072	-.028
	Sig. (2-tailed)	.121	.279	.908		.193	.617
	N	327	318	286	327	327	327
Total Area (M2) with scaling	Pearson Correlation	.776**	.095	.301**	.072	1	.443**
	Sig. (2-tailed)	.000	.071	.000	.193		.000
	N	376	366	331	327	376	376
Group of employees	Pearson Correlation	.454**	-.024	.220**	-.028	.443**	1
	Sig. (2-tailed)	.000	.643	.000	.617	.000	
	N	376	366	331	327	376	376
**. Correlation is significant at the 0.01 level (2-tailed).							

Correlations								
		Gas Energy (Scaled MJ)	Bewteen Buildings Location	Single Building Location	Building Block Surrounding	Open surrounding	Curtain wall Façade	Brick Façade
Gas Energy (Scaled MJ)	Pearson Correlation	1	-.136**	.234**	-.253**	.158**	.049	-.054
	Sig. (2-tailed)		.008	.000	.000	.002	.375	.331
	N	376	375	375	375	375	327	327
Bewteen Buildings Location	Pearson Correlation	-.136**	1	-.459**	.253**	-.191**	.106	.117*
	Sig. (2-tailed)	.008		.000	.000	.000	.055	.034
	N	375	375	375	375	375	327	327
Single Building Location	Pearson Correlation	.234**	-.459**	1	-.251**	.156**	.006	-.088
	Sig. (2-tailed)	.000	.000		.000	.002	.910	.114
	N	375	375	375	375	375	327	327
Building Block Surrounding	Pearson Correlation	-.253**	.253**	-.251**	1	-.830**	.011	-.016
	Sig. (2-tailed)	.000	.000	.000		.000	.846	.780
	N	375	375	375	375	375	327	327
Open surrounding	Pearson Correlation	.158**	-.191**	.156**	-.830**	1	.028	.036
	Sig. (2-tailed)	.002	.000	.002	.000		.615	.512
	N	375	375	375	375	375	327	327
Curtain wall Façade	Pearson Correlation	.049	.106	.006	.011	.028	1	-.418**
	Sig. (2-tailed)	.375	.055	.910	.846	.615		.000
	N	327	327	327	327	327	327	327
Brick Façade	Pearson Correlation	-.054	.117*	-.088	-.016	.036	-.418**	1
	Sig. (2-tailed)	.331	.034	.114	.780	.512	.000	
	N	327	327	327	327	327	327	327
**. Correlation is significant at the 0.01 level (2-tailed).								
*. Correlation is significant at the 0.05 level (2-tailed).								

Correlations							
		Total energy (scaled) MJ	Building Age	Height (cm)	Glass Percentage	Total Area (M2) with scaling	Group of employees
Total energy (scaled) MJ	Pearson Correlation	1	.123*	.269**	.080	.807**	.502**
	Sig. (2-tailed)		.019	.000	.148	.000	.000
	N	376	366	331	327	376	376
Building Age	Pearson Correlation	.123*	1	-.320**	.061	.095	-.024
	Sig. (2-tailed)	.019		.000	.279	.071	.643
	N	366	366	324	318	366	366
Height (cm)	Pearson Correlation	.269**	-.320**	1	-.007	.301**	.220**
	Sig. (2-tailed)	.000	.000		.908	.000	.000
	N	331	324	331	286	331	331
Glass Percentage	Pearson Correlation	.080	.061	-.007	1	.072	-.028
	Sig. (2-tailed)	.148	.279	.908		.193	.617
	N	327	318	286	327	327	327
Total Area (M2) with scaling	Pearson Correlation	.807**	.095	.301**	.072	1	.443**
	Sig. (2-tailed)	.000	.071	.000	.193		.000
	N	376	366	331	327	376	376
Group of employees	Pearson Correlation	.502**	-.024	.220**	-.028	.443**	1
	Sig. (2-tailed)	.000	.643	.000	.617	.000	
	N	376	366	331	327	376	376
*. Correlation is significant at the 0.05 level (2-tailed).							
**. Correlation is significant at the 0.01 level (2-tailed).							

Correlations								
		Total energy (scaled) MJ	Bewteen Buildings Location	Single Building Location	Building Block Surrounding	Open surrounding	Curtain wall Façade	Brick Façade
Total energy (scaled) MJ	Pearson Correlation	1	-.181**	.286**	-.304**	.207**	.039	-.054
	Sig. (2-tailed)		.000	.000	.000	.000	.485	.329
	N	376	375	375	375	375	327	327
Bewteen Buildings Location	Pearson Correlation	-.181**	1	-.459**	.253**	-.191**	.106	.117*
	Sig. (2-tailed)	.000		.000	.000	.000	.055	.034
	N	375	375	375	375	375	327	327
Single Building Location	Pearson Correlation	.286**	-.459**	1	-.251**	.156**	.006	-.088
	Sig. (2-tailed)	.000	.000		.000	.002	.910	.114
	N	375	375	375	375	375	327	327
Building Block Surrounding	Pearson Correlation	-.304**	.253**	-.251**	1	-.830**	.011	-.016
	Sig. (2-tailed)	.000	.000	.000		.000	.846	.780
	N	375	375	375	375	375	327	327
Open surrounding	Pearson Correlation	.207**	-.191**	.156**	-.830**	1	.028	.036
	Sig. (2-tailed)	.000	.000	.002	.000		.615	.512
	N	375	375	375	375	375	327	327
Curtain wall Façade	Pearson Correlation	.039	.106	.006	.011	.028	1	-.418**
	Sig. (2-tailed)	.485	.055	.910	.846	.615		.000
	N	327	327	327	327	327	327	327
Brick Façade	Pearson Correlation	-.054	.117*	-.088	-.016	.036	-.418**	1
	Sig. (2-tailed)	.329	.034	.114	.780	.512	.000	
	N	327	327	327	327	327	327	327
**. Correlation is significant at the 0.01 level (2-tailed).								
*. Correlation is significant at the 0.05 level (2-tailed).								

D: Sub-sectors regression analysis

1- Retail sub-sector:

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.715[a]	.512	.509	22.99698019
2	.757[b]	.572	.567	21.59021766
3	.768[c]	.590	.582	21.22147920

a. Predictors: (Constant), Total Area (M2) with scaling

b. Predictors: (Constant), Total Area (M2) with scaling, Group of employees

c. Predictors: (Constant), Total Area (M2) with scaling, Group of employees, Open surrounding

d. Dependent Variable: Electirc Energy (scaled MJ)

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.784[a]	.615	.612	31.569
2	.795[b]	.632	.628	30.935
3	.803[c]	.644	.637	30.527

a. Predictors: (Constant), Total Area (M2) with scaling

b. Predictors: (Constant), Total Area (M2) with scaling, Building Age

c. Predictors: (Constant), Total Area (M2) with scaling, Building Age, Group of employees

d. Dependent Variable: Gas Energy (Scaled MJ)

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.807[a]	.651	.648	46.748
2	.822[b]	.676	.672	45.163
3	.832[c]	.693	.687	44.104

a. Predictors: (Constant), Total Area (M2) with scaling

b. Predictors: (Constant), Total Area (M2) with scaling, Group of employees

c. Predictors: (Constant), Total Area (M2) with scaling, Group of employees, Building Age

d. Dependent Variable: Total energy (scaled) MJ

2- Offices sub-sector:

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.795[a]	.632	.628	21.78936315
2	.814[b]	.663	.656	20.94543415

a. Predictors: (Constant), Total Area (M2) with scaling

b. Predictors: (Constant), Total Area (M2) with scaling, Group of employees

c. Dependent Variable: Electirc Energy (scaled MJ)

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.822[a]	.675	.672	28.592
2	.831[b]	.691	.684	28.057

a. Predictors: (Constant), Total Area (M2) with scaling

b. Predictors: (Constant), Total Area (M2) with scaling, Group of employees

c. Dependent Variable: Gas Energy (Scaled MJ)

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.854[a]	.729	.726	41.873
2	.869[b]	.755	.749	40.049

a. Predictors: (Constant), Total Area (M2) with scaling

b. Predictors: (Constant), Total Area (M2) with scaling, Group of employees

c. Dependent Variable: Total energy (scaled) MJ

3- Education sub-sector:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.769[a]	.591	.523	26.23002975

a. Predictors: (Constant), Total Area (M2) with scaling

b. Dependent Variable: Electirc Energy (scaled MJ)

4- Health sub-sector:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.755[a]	.569	.483	25.24100362

a. Predictors: (Constant), Total Area (M2) with scaling

b. Dependent Variable: Electirc Energy (scaled MJ)

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.966[a]	.933	.919	14.63106793
2	.993[b]	.986	.979	7.484710797

a. Predictors: (Constant), Total Area (M2) with scaling

b. Predictors: (Constant), Total Area (M2) with scaling, Building Age

c. Dependent Variable: Gas Energy (Scaled MJ)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.952[a]	.906	.887	26.84278930

a. Predictors: (Constant), Total Area (M2) with scaling

b. Dependent Variable: Total energy (scaled) MJ

5- Non-office based services sub-sector:

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.705[a]	.497	.469	18.92103059
2	.807[b]	.652	.611	16.18661118

a. Predictors: (Constant), Group of employees

b. Predictors: (Constant), Group of employees, Total Area (M2) with scaling

c. Dependent Variable: Electirc Energy (scaled MJ)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.662[a]	.439	.407	32.879

a. Predictors: (Constant), Group of employees

b. Dependent Variable: Gas Energy (Scaled MJ)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.726[a]	.527	.501	42.039

a. Predictors: (Constant), Group of employees

b. Dependent Variable: Total energy (scaled) MJ

Summary

TOWARDS SMART ENERGY CITY; ANALYSIS OF THE ENERGY USAGE IN SERVICES SECTOR IN THE CITY OF EINDHOVEN

Construction Management and Urban Development 2010-2012

Author: Iman Karimi

Graduation program:

Construction Management and Urban Development

Graduation committee:

Prof.dr.ir. B. de Vries
Dr. Qi Han
Dr.ir. Erik Blokhuis

Date of graduation:

28-08-12

ABSTRACT

This paper is about the energy usage of the services sector in Eindhoven. Firstly, one dataset is created which contains the annual energy usage of the buildings of this sector together with the energy usage variables. The cluster analyses have been done on this dataset to put the buildings in different groups on the basis of different criteria. In addition, the multiple regression analyses are used for the energy usage prediction of the buildings. The results of the research are the energy usage dataset for the services sector, different clusters of the buildings, important predictors for the energy usage in different sub sectors and different clusters and finally the energy usage equations in different clusters.

Keywords: services sector of Eindhoven, energy usage, energy usage predictors, clustering, regression analysis

INTRODUCTION

The energy sector faces numerous problems, e.g. climate change, environmental and human accidents, reliability of energy supply and oil dependency. It is therefore time to launch a fundamental change with respect to our energy supply. The drivers for such a change originate in broad societal ambitions, and materialize in policy that is mostly formulated at national and cross-border levels. For instance, the European Commission has formulated major objectives for future energy systems, e.g. to reduce carbon emissions by 20%, to increase the share of renewable energies by 20%, and to increase the energy efficiency by 20% before 2020. This fundamental change implies transitions towards new sustainable energy systems, in which energy reduction ambitions play a major role.

One important sector in terms of energy usage is the services sector. Within the province of Noord-Brabant this sector uses around 20% of the total energy usage. In the table 1 the energy usage of different sectors in this province has been shown [1].

Eindgebruikers	Energie gebruik 2040 (PJ)	Extra potentiele besparing (%)	Extra potentiele besparing (PJ)
Huishoudens	63 - 98	40 - 60%	25 - 59
Diensten	73 - 108	40 - 60%	30 - 65
Landbouw	13 - 27	> 50%	6 - 14
Industrie	84 - 146	10 - 20%	8 - 29
Transport	73 - 126	20 - 65%	12 - 82
Totaal	***375 - 625***	***160 - 255%***	***83 - 249***

Table 50: Primary yearly energy use according to end users and potential energy savings in Noord_Brabant (Source: SOLET report, 2008)

As you can see in this table, the potential for energy saving in the sector of services is 40% to 60% or 30-65 Peta Joules (PJ). This provides the municipality with great opportunity for making policies to save energy. Indeed, it is obvious that study the energy usage in the services sector is really important. This research has focused specifically on this issue.

PROBLEM DESCRIPTION

On average, residential energy use covers a relative large part of the total energy use in cities (approximately 40%). Indeed, modeling the energy use of households by considering the occupant's behavior is really important. In this regard we could find many articles which have studied this issue. However, the energy use of commercial and non-commercial sector (for example the services sector) accounts for almost the same amount of city energy. The research after the energy usage in the services sectors remains a shallow area of the research. Through the literature study, we could not find a suitable article which describes the energy use patterns in this sector clearly. Also, the data about the energy usage in details is not enough and satisfactory in the services sector. Indeed, studying the energy usage variables in the services sector could be a vital initial step for beginning of the research in this area.

This research aims to study the energy usage in the services sectors by considering the important variables for the energy usage prediction and also clustering the different building types on the basis of these variables.

STATE-OF-THE ART

The energy use of industrial companies has been researched extensively. Especially, the contribution of Ang (1995) [2] and Ang and Lee (1996) [3] are interesting; they focus on decomposition of industrial energy composition, aiming to study the impacts of structural change and changes in sectoral energy efficiencies. However, by looking through the available

articles in this field, research after services sector energy use is often executed on national scale, not discussing individual cases. Also, there is not any dataset available which includes information about the energy usage and the important variables in this regard. Indeed, creating a dataset which has the information about the companies in the services sector and buildings characteristics and employees numbers would be very beneficial for also future studies in this field.

The focus of this research is on the services sector of the city of Eindhoven. The services sector is divided to retail, catering, offices, health, education and non offices based services sub sectors. After creating the dataset for the buildings of this sector, by using the statistical analysis, we could find the most important variables which are effective in energy usage. This issue has not been considered in the district level or city level previously and by considering the energy usage of the individual buildings. Furthermore, cluster analysis of the buildings in the services sector in the city of Eindhoven is a new subject of study. The clustering has been done on the dwellings in the city of Eindhoven previously [4]. However, there is not any clustering for the services sector's buildings.

RESEARCH DESIGN

In this research three main steps have been made. Firstly, a dataset has been built which contains the energy usage and the related variables of the buildings. After that the cluster analysis has been used to put the buildings in different groups from various perspectives. Finally, the multiple regression analysis has been implemented on the dataset for the purpose of prediction of the energy usage formula. Figure 1 can clear the process better.

CREATING THE DATASET

The dataset has been created on the basis of available data from various sources. The information which has been gathered in the dataset contains the annual energy usage of the buildings (separately for electric and gas energy), building characteristics (total floor area, glass percentage, façade type, height, building age, number of floors, surrounding situation (open, building block or river) and location situation (single, side or between)) and company characteristics (address, number of employees, SBI 93 sector). In total there are 387 buildings in the dataset. However, some information is missing for some of the buildings. This dataset is the initial step for performing this research.

After creating the dataset, the data preparation procedure is followed. The missing data has been recognized. The outliers of the dataset also need to be modified. For outlier adjustments, the following formulas have been used [5]:

$$\text{Upper boundary} = Q3 + (1.5 * (Q3 - Q1)) \quad (1)$$
$$\text{Lower boundary} = Q1 - (1.5 * (Q3 - Q1)) \quad (2)$$

The data with more values compared to upper boundary and less values compare to lower boundary are modified with replacing them with the boundary value. On the other hand, the

categorical variables of the dataset should be turned to dummy variables. This issue is done for surrounding situations, building locations and façade types.

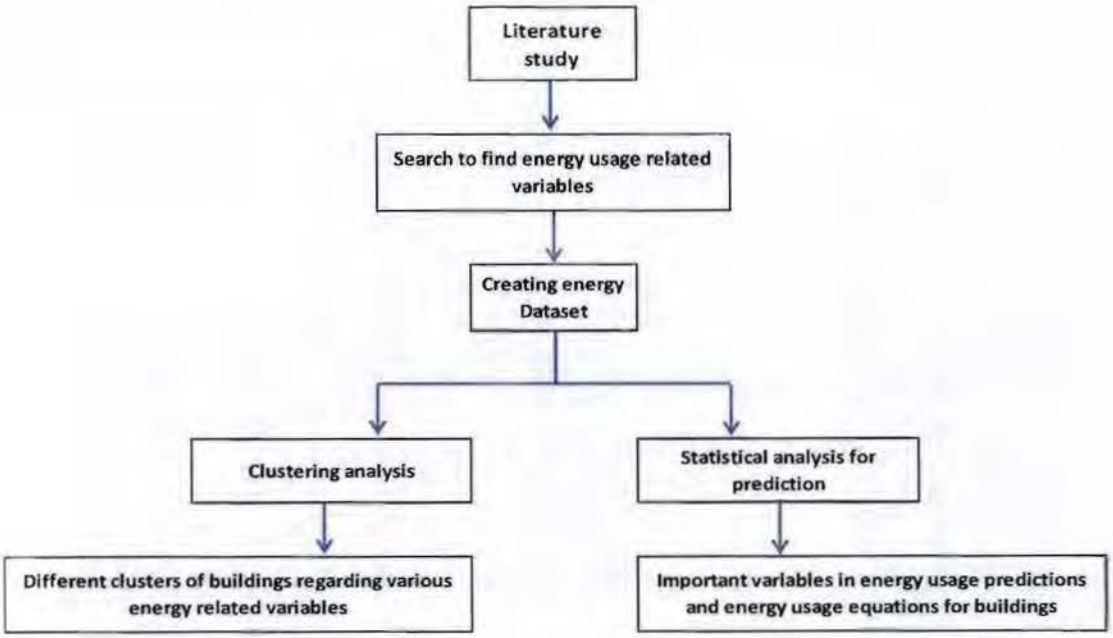

Figure 9: research design

CLUSTER ANALYSIS

The cluster analysis is one purpose of this research. In this project, we have used two methods for the cluster analysis; k-means clustering and two step clustering. The SPSS software is used for the analysis.

K-means clustering

For running the k-means cluster analysis, first we should reduce the number of the variables of the dataset by using the Principal Components Analysis (PCA). With the PCA method, we can extract the components from the dataset. These components represent also the other variables with high accuracy. The PCA has been run for our dataset and finally three components have been extracted. These components show high correlations with the energy usage, total area, surrounding type, building age and height of the building. Regarding the correlations, the first component is called energy usage and scale, the second one is called surrounding type and the third one is labeled building age and height. These components represent 82 percent of the variables of our dataset.

After the extracted components are known, the k-means cluster analysis is implemented by using them. In running the k-means clustering, the number of k should be chosen upfront. In our analysis we have used the k number as 4, 6, 8 and 10. For checking the homogeneity of the clusters, we have used the weighted average standard deviation (WASD) measure [6]. The

standard deviation per cluster regarding a characteristic (s or SD) is the square root of the variance (s2) per cluster regarding that characteristic. The equation below shows the way of calculation for WASD:

$$WASD_{\text{per characteristic}} = \frac{\sum_{i=1}^{i=n}(N_i * SD_i)}{N} = \frac{\sum_{i=1}^{i=n}\left(N_i * \sqrt{\frac{\sum_{j=1}^{j=N_i}(x_j - \bar{x})^2}{N_i}}\right)}{N} \quad (3)$$

A low WASD indicates a high homogeneity within the cluster. We have run the analysis for different k numbers and for each of the clusters, the WASD have been calculated for the energy use and scale component (because this component is the main focus of the clustering). The results of the calculation can be seen in table 2.

Number of K	WASD										
	Cluster 1	cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Average
4	21.12	29.69	26.06	33.76							27.6575
6	23.72	14.79	23	19.22	25.06	8.05					18.9733333
8	10.4	14.55	8.08	12.86	6.2	26.73	7.14	12.03			12.24875
10	5.21	13.35	5.74	23.72	8.56	11.1	10.08	5.82	5.13	7.07	9.578

Table 2: WASD for different k numbers in each cluster

As you can see in this table, the k=10 has the best results. By assuming the k=10, the k-means cluster analysis is run. The table 3 shows the division of the buildings for this clustering. The table indicates that 324 buildings have been divided to 10 clusters. On the basis of the final cluster centers, the clusters have been labeled for the energy usage and total area. In table 4, you can see the information about each of these clusters.

Two steps clustering

The two step clustering is the other method of clustering in our project. By using this method, we can divide the buildings on the basis of different variables. The electric and gas energy usage are the variables that have been used for the clustering. On the basis of these variables and by trial and error method for the best number for clustering, three clusters have been defined. The results can be seen in figure 2. As it can be seen in the figure, three different clusters are on the basis of the energy usage; low energy usage, medium energy usage and high energy usage.

ENERGY USAGE PEREDICTION ANALYSIS

One purpose of this research is the energy usage prediction analysis for the services sector of the Eindhoven. The multiple regression analysis has been used to reach this goal by means of the SPSS software. This method has been used first for the entire dataset. Unfortunately, the results were not acceptable for all the buildings at the same time together. Then, the analysis is done for the sub sectors separately, which the outcomes were not favorable enough as well.

Number of Cases in each Cluster		
Cluster	1	17.000
	2	41.000
	3	22.000
	4	92.000
	5	31.000
	6	32.000
	7	27.000
	8	16.000
	9	12.000
	10	34.000
Valid		324.000
Missing		63.000

Table 3: number of cases in each cluster for the k-means clustering with k=10

Cluster number	Average Total Area (m2)	Average Electric energy (MJ)	Average Gas energy (MJ)	Total area level	Electric energy level	Gas energy level
1	4,606	695,010	1,521,963	High	High	High
2	1,551	202,020	510,547	Medium	Medium	High
3	1,051	182,451	335,154	Medium	Low	Medium
4	954	134,366	231,938	Low	Low	Medium
5	940	166,714	206,782	Low	Low	Medium
6	1,397	243,883	355,503	Medium	Medium	Medium
7	4,312	903,966	1,218,739	High	High	High
8	3,523	695,029	1,079,237	High	High	High
9	4,578	797,608	1,013,510	High	High	High
10	671	89,699	190,544	Low	Low	Low

Table 4: levels of energy usage and total area for each cluster

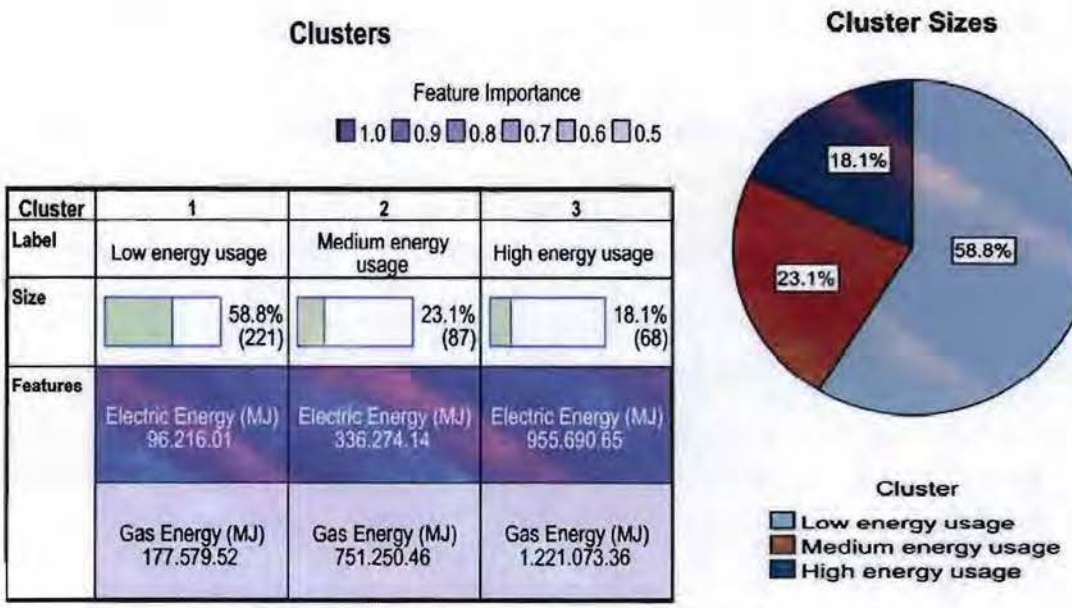

Figure 2: Two-step clustering results regarding energy usage

Then, the analysis performed on the previous made clusters, which again the results were not satisfactory. For reaching the desirable outcomes, the cluster analysis is performed again on our dataset to find out about the suitable cluster of the buildings, only for the purpose of prediction. Here, we are going to show the results of analysis for one sub sector and for the total energy prediction.

In the multiple linear regression method, a straight line or a linear relationship is assumed between the dependent variable and the predictors. The relationship can be written down as follow [7]:

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} + e_i \qquad (4)$$

Where y_i is the value of the i^{th} case of the dependent scale variable, p is the number of predictors, b_j is the value of the j^{th} coefficient, $j = 0, \ldots, p$, x_{ij} is the value of the i^{th} case of the j^{th} predictor, e_i is the error in the observed value for the i^{th} case.

The important aspect in the regression analysis is the correlation analysis between the predictors and the dependent variable and the predictors themselves. From the correlation analysis results, it is clear that the correlation is significant between total area and employee numbers. However, we would like to keep both of these variables in the prediction because both of them are highly correlated to the energy usage. Indeed, we define a new variable which is called "scale of the building" and it is defined as follow:

*Scale of the building = Total area of the building * Number of employees of the company* (5)

ENERGY USAGE PREDICTION IN RETAIL SUB SECTOR

The retail sub-sector has most of the buildings of our dataset. There are 204 buildings in this sub-sector. The distribution of the total energy has been shown in figure 3.

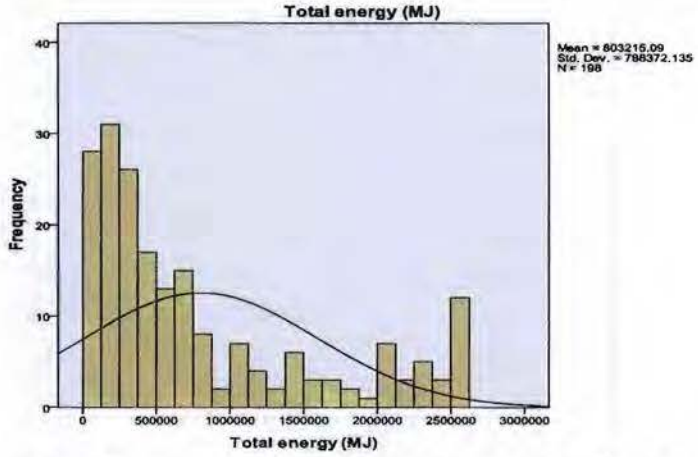

Figure 3: Histogram for total energy usage in retail sub-sector

Considering the statistics table for this sub sector, for the energy usage the mean is quite different from the median, suggesting that the distribution is asymmetric. Also, the histogram above shows that there is no normal distribution in our dataset for the total energy. The difference between the normal distribution which is shown with a black line and the current distribution is obvious. These issues indicate that the prediction of the energy usage for this dataset is quite hard. The regression analysis is run for this sub sector. Considering the total energy usage, total area of the building, number of employees and building age are important predictors, respectively. All of these predictors have the positive effects on the total energy usage. This means that by increasing these predictors, the total energy usage will increase consequently. However, the energy usage equation does not give us good outcomes.

ENERGY USAGE PREDICTION WITH CLUSTERING

Before the beginning of the prediction analysis, the cluster analysis is done by using the scale of the building variable. For the two step clustering method, after testing different variables for the clustering, we have used "Total energy / Scale of the building (T/SB)" as the continuous variable for classification of the buildings. After the cluster analysis on the entire dataset, 21 clusters have been defined for the total energy prediction. There are 363 buildings in total in these clusters. The multiple linear regression method is used in each of these clusters for the prediction of the total energy usage of the buildings. Table 5 shows the summary of the clustering analysis information and also the regression analysis results for the total energy in different clusters. The table shows the number of buildings in each cluster, mean amount of Total energy / Scale of the building for each cluster, the R and R square from the regression analysis and at last the total energy usage equation for each cluster. As it could be seen in table 5, scale of the building is the most important predictor as we expected before. Height, building age, glass percentage and brick façade are the other predictors that are needed in some clusters for better estimation of the total energy usage.

CONCLUSIONS, DISCUSSIONS AND RECOMMENDATIONS

In this research the energy usage of the services sector in the city of Eindhoven is assessed. Firstly, a dataset is created which contains the energy usage amount of the buildings, some building characteristics and company information. By using this dataset, the cluster analysis is performed to put the buildings in different clusters on the basis of different variables. Two clustering methods are used, namely k-means clustering and two step clustering. Also, the multiple regression analysis is used for the energy usage prediction. As the results of the analyses, different clusters of the buildings have been defined. The important predictors of the energy usage have been recognized. According to the results, scale of the building is the most important predictor. Furthermore, building age, glass percentage, surrounding type, façade type and height of the building are the other useful predictors. The energy usage equations are also derived for different clusters.

The municipality of Eindhoven can use the built dataset as the initial step for the data gathering about the energy usage of the services sector. Also, by the created clusters, beneficial perspectives are made for the policy makers about the energy levels of the buildings.

Cluster Number	Clustering information		Linear Regression information		
	Number of buildings	Mean Total energy/ Scale of building	R	R square	Prediction equation for total energy
1	17	3.14	0.996	0.992	Total Energy = -21.035 + 3.459 * Scale of building + 2.963 * Height
2	26	2.44	0.994	0.989	Total Energy = 3.889 + 2.279 * Scale of building
3	9	4.49	0.996	0.992	Total Energy = 1.511 + 4.381 * Scale of building
4	8	5.67	0.998	0.996	Total Energy = 1.362 + 5.473 * Scale of building
5	9	7.91	1	0.999	Total Energy = 0.608 + 7.728 * Scale of building
6	4	6.43	1	1	Total Energy = 0.286 + 6.362 * Scale of building
7	22	1.61	0.998	0.997	Total Energy = 0.845 + 1.586 * Scale of building
8	20	1.93	0.997	0.994	Total Energy = 2.614 + 1.997 * Scale of building - 0.198 * Building age
9	31	1.06	0.999	0.998	Total Energy = 0.286 + 1.048 * Scale of building
10	22	1.16	1	0.999	Total Energy = -2.248 + 1.161 * Scale of building + 0.106 * Building age
11	15	1.26	0.999	0.999	Total Energy = - 0.452 + 1.278 * Scale of building
12	16	1.39	1	1	Total Energy = 0.571 + 1.348 * Scale of building + 2.087 * Brick Façade
13	15	0.49	0.996	0.992	Total Energy = - 2.183 + 0.506 * Scale of building
14	27	0.57	0.999	0.999	Total Energy = - 5.759 + 0.583 * Scale of building + 0.129 * Glass percentage
15	15	0.65	0.999	0.999	Total Energy = - 0.161 + 0.656 * Scale of building
16	11	0.19	0.983	0.966	Total Energy = - 0.133 + 0.201 * Scale of building
17	18	0.36	0.978	0.956	Total Energy = 0.269 + 0.351 * Scale of building
18	29	0.85	0.999	0.999	Total Energy = 1.282 + 0.828 * Scale of building
19	17	0.95	0.999	0.999	Total Energy = 0.476 + 0.943 * Scale of building
20	10	0.71	1	1	Total Energy = - 0.285 + 0.712 * Scale of building
21	22	0.77	0.999	0.999	Total Energy = - 1.390 + 0.783 * Scale of building
Total	**363**				

Table 5: Clustering and Prediction results for the total energy

The results of the clustering and energy prediction, including the important predictors, can help the authorities to make the energy reduction policies for the services sector. However, the results can be improved in future with better dataset.

For the future studies, the effort should be made for adding more energy related variables to the current dataset. These variables can be the energy usage of the building in hourly basis, the user behavior of the buildings (such as the patterns of the energy usage for the employees) and the number of appliances of the company. Some more building characteristics also should be investigated for the services sector's buildings. Insulations of walls, floors, roofs, windows and pipes would be important. Also, more buildings should be included in the dataset which will make the results of the analysis even better. The clusters which we have made in this project for the prediction analysis should be investigated in details to find out why they have similar behavior in terms of the energy usage. Some specific buildings can be chosen from the dataset and from each cluster to check the energy usage management system of them in reality.

REFERENCES

[1] Kasteren, van, H., Konz, W., van Schijndel, P., Smeets, R., Wentink, C., (December 2008), *"SOLET report: Energiek Brabant – Een scenariostudie naar de energievoorziening van Noord-Brabant in 2040"*.

[2] Ang, B.W., (1995), *"Multilevel decomposition of industrial energy consumption"*, Energy Economics, 17 (1), pp. 39–51

[3] Ang, B.W., Lee, P.W., (1996), *"Decomposition of industrial energy consumption: the energy coefficient approach"*, Energy Economics, 18, pp. 129–143

[4] van Loon, P., (March 2012), *"Target group clustering for applications of energy effective renovation concerning privately owned dwellings"*, CME master program Thesis, Eindhoven University of Technology

[5] Hoaglin, D.C., Iglewicz, B., and Tukey, J.W., (1986), *"Performance of some resistant rules for outlier labeling"*, Journal of American Statistical Association, 81, 991-999.

[6] Wu, C., Rashi Sharma, R., (2011), "Housing submarket classification: The role of spatial contiguity", Applied geography, 32, 746-56.

[7] PASW statistics 18 core system user's guide

IMAN KARIMI
i.karimi@student.tue.nl

2002-2006 Bachelor in Civil Engineering in Arak University, Iran
2006-2009 Master in Civil Engineering in TMU, Iran
2005-2009 Karbordkar construction contractors, Tehran, Iran
2009-2010 Dalgostar Contractors, Tehran, Iran
2010-2012 Master in Construction Management and Engineering, Eindhoven University of Technology
2012-2014 PDEng in Comprehensive design in civil engineering, Delft University of Technology