Data Mining Approaches for Calculating the Energy Consumption of Buildings

Within the context of the Zero Budget Sustainability service

By

M. Maaskant MSc.

Faculty of Industrial Engineering and Innovation Sciences
TECHNISCHE UNIVERSITEIT EINDHOVEN
In collaboration with PLOOS ENERGIEVERLENING

First supervisor: Prof. dr. Ir. Uzay Kaymak
Supervisor company: ing. Paul Vermeulen
Second reader: dr. Anna Wilbik

JANUARY 2016

Word count: 23322
Abstract

Zero Budget Sustainability is a high potential service offered by Ploos Energieverlening. In this service, external financers invest in energy saving measures when the owner of the building lacks budget. The resulting savings are used to repay the investor. Currently, both financers and building owners lack confidence in the forecasted savings. It has been indicated that an independent scientific approach to estimate energy consumption of buildings can help to improve confidence.

Based on a scientific literature study and expert knowledge, an artificial neural network has been built to calculate the energy consumption of commercial buildings on 15 minute time intervals. These estimations are based on climate variables, behavioral data and building characteristics including 20 specified low-level energy saving measures. The dataset is further improved by a fuzzy cluster analysis. The model has also been tested on a dataset containing residential buildings.

The obtained results show that electricity consumption can be estimated with no more than 8.23 percent symmetric errors. Gas consumption can be estimated with 26 percent symmetric errors. Over a longer period the difference between estimated and observed consumption is 0.3 percent for electricity and 4.3 percent for gas. For residential buildings, absence of behavioral data about the residents resulted in less accurate predictions although adding re-supplied electricity from solar panels made an improvement. During the research project, meaningful recommendations and practical implications for the model were given to the company.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Background</strong></td>
<td></td>
</tr>
<tr>
<td>1.1 Company description</td>
<td>7</td>
</tr>
<tr>
<td>1.2 Data mining and Big Data</td>
<td>8</td>
</tr>
<tr>
<td>1.3 Problem description</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Zero Budget Sustainability</td>
<td>10</td>
</tr>
<tr>
<td>1.5 Buildings’ energy consumption estimation</td>
<td>11</td>
</tr>
<tr>
<td>1.6 Research questions</td>
<td>12</td>
</tr>
<tr>
<td>1.6.1 Sub questions</td>
<td>13</td>
</tr>
<tr>
<td>1.7 Methodology</td>
<td>14</td>
</tr>
<tr>
<td>1.8 Project plan</td>
<td>15</td>
</tr>
<tr>
<td>1.9 Thesis structure</td>
<td>16</td>
</tr>
<tr>
<td><strong>2 Literature review</strong></td>
<td></td>
</tr>
<tr>
<td>2.1 Search method</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Investment in energy consumption reduction</td>
<td>18</td>
</tr>
<tr>
<td>2.2.1 Relation to ZBS</td>
<td>19</td>
</tr>
<tr>
<td>2.3 Energy consumption estimation techniques</td>
<td>20</td>
</tr>
<tr>
<td>2.4 Machine learning</td>
<td>21</td>
</tr>
<tr>
<td>2.4.1 Supervised learning</td>
<td>22</td>
</tr>
<tr>
<td>2.4.2 Reinforcement learning</td>
<td>23</td>
</tr>
<tr>
<td>2.4.3 Unsupervised learning</td>
<td>24</td>
</tr>
<tr>
<td>2.5 Clustering</td>
<td>25</td>
</tr>
<tr>
<td>2.5.1 Partitioning criteria</td>
<td>26</td>
</tr>
<tr>
<td>2.5.2 Cluster analysis methods</td>
<td>27</td>
</tr>
<tr>
<td>2.5.3 Determining the number of clusters</td>
<td>28</td>
</tr>
<tr>
<td>2.5.4 Validation indices</td>
<td>29</td>
</tr>
<tr>
<td>2.6 Case-based reasoning</td>
<td>30</td>
</tr>
<tr>
<td>2.7 Neural Networks</td>
<td>31</td>
</tr>
<tr>
<td>2.7.1 Learning</td>
<td>32</td>
</tr>
<tr>
<td>2.8 Summary</td>
<td>33</td>
</tr>
<tr>
<td><strong>3 Methodology</strong></td>
<td></td>
</tr>
<tr>
<td>3.1 The current calculation module</td>
<td>34</td>
</tr>
<tr>
<td>3.2 CRISP-DM</td>
<td>35</td>
</tr>
<tr>
<td>3.2.1 Data Preparation</td>
<td>36</td>
</tr>
<tr>
<td><strong>4 The dataset</strong></td>
<td></td>
</tr>
<tr>
<td>4.1 Data understanding</td>
<td>37</td>
</tr>
<tr>
<td>4.1.1 External data</td>
<td>38</td>
</tr>
<tr>
<td>4.1.2 Internal data</td>
<td>39</td>
</tr>
<tr>
<td>4.2 Data preparation</td>
<td>40</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Ploos Energieverlening logo</td>
<td>8</td>
</tr>
<tr>
<td>1.2 Facilicom logo</td>
<td>9</td>
</tr>
<tr>
<td>1.3 Overview of Big Data</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Zero Budget Sustainability structure</td>
<td>12</td>
</tr>
<tr>
<td>1.5 ZBS in terms of time and costs</td>
<td>12</td>
</tr>
<tr>
<td>1.6 An overview of the steps in CRISP-DM</td>
<td>15</td>
</tr>
<tr>
<td>2.1 Number of publications for the search query: &quot;Buildings energy consumption neural networks&quot;</td>
<td>18</td>
</tr>
<tr>
<td>2.2 Schematic overview of reverse factoring</td>
<td>19</td>
</tr>
<tr>
<td>2.3 Overview of techniques for estimating residential energy consumption</td>
<td>21</td>
</tr>
<tr>
<td>2.4 Structure of the customer characterization framework, obtained from (Figueiredo, Rodriguez, Vale, &amp; Gouveial, 2005)</td>
<td>23</td>
</tr>
<tr>
<td>2.5 K-fold cross-validation</td>
<td>26</td>
</tr>
<tr>
<td>2.6 Confusion matrix for binary classification problems</td>
<td>27</td>
</tr>
<tr>
<td>2.7 The standard reinforcement learning model</td>
<td>28</td>
</tr>
<tr>
<td>2.8 Graphical representation of fuzzy clustering</td>
<td>32</td>
</tr>
<tr>
<td>2.9 Images that show why iVAT is preferred to VAT</td>
<td>33</td>
</tr>
<tr>
<td>2.10 The CBR cycle, obtained from (Watson, 1999)</td>
<td>35</td>
</tr>
<tr>
<td>2.11 An example of an artificial neuron</td>
<td>37</td>
</tr>
<tr>
<td>2.12 A perceptron</td>
<td>37</td>
</tr>
<tr>
<td>2.13 A perceptron with constant bias</td>
<td>38</td>
</tr>
<tr>
<td>2.14 Layered feed-forward neural network</td>
<td>39</td>
</tr>
<tr>
<td>2.15 Relation between errors on the training and validation set</td>
<td>41</td>
</tr>
<tr>
<td>3.1 Example of an energy consumption analysis</td>
<td>42</td>
</tr>
<tr>
<td>3.2 Four level breakdown of the CRISP-DM methodology, obtained from (Smart Vision Europe, 2013)</td>
<td>44</td>
</tr>
<tr>
<td>3.3 Breakdown of the data preparation step</td>
<td>45</td>
</tr>
<tr>
<td>4.1 Locations of the weather stations</td>
<td>47</td>
</tr>
<tr>
<td>4.2 Location P.132708.15, 2015 week 21</td>
<td>48</td>
</tr>
<tr>
<td>4.3 Old classification of measures</td>
<td>52</td>
</tr>
<tr>
<td>4.4 New classes</td>
<td>53</td>
</tr>
<tr>
<td>5.1 iVAT images with normalization (left) and without normalization (right)</td>
<td>56</td>
</tr>
<tr>
<td>5.2 An office building in the dataset</td>
<td>58</td>
</tr>
<tr>
<td>5.3 Satellite image of the apartment buildings in the dataset</td>
<td>58</td>
</tr>
<tr>
<td>5.4 Regression results for the basic gas model</td>
<td>59</td>
</tr>
<tr>
<td>5.5 Model without time dummies</td>
<td>60</td>
</tr>
<tr>
<td>5.6 Typical pattern of gas consumption during the heating season</td>
<td>60</td>
</tr>
</tbody>
</table>
List of Tables

2.1 The Fuzzy C-Means Algorithm ........................................... 32
2.2 VAT Recording Algorithm ............................................. 34
2.3 Inputs, predicted values, observed values and squared errors for the model of Figure 2.13 ........................................... 38
2.4 Inputs, predicted values, observed values and mean squared error for the updated weight ........................................... 40

4.1 Climate parameters ....................................................... 48
4.2 Energy measurements .................................................... 49
4.3 Overview of non real-time data ......................................... 49

6.1 Results for the Pearson’s and Spearman’s CCV using non-normalized and normalized data ........................................... 63
6.2 Xie-Beni index values ..................................................... 63
6.3 Neural network: results for electricity prediction ......................... 64
6.4 Neural network: results for electricity prediction using different samples ........................................... 65
6.5 Neural network: results for electricity prediction using occupation rate ........................................... 65
6.6 Neural network: results for gas prediction ................................ 66
6.7 Neural network: results for gas prediction based on daily summations ........................................... 66
6.8 Neural network: results for gas prediction based on daily summations using different distance measures ........................................... 67
6.9 Neural network: results for gas prediction using occupation rate ........................................... 67
6.10 Results for energy prediction based on daily summations .................... 67
6.11 Results for electricity prediction of residential buildings .................... 68
6.12 Variables, scores and ranks ............................................. 70

A.1 Climate parameters for Eindhoven on 18-06-2014 ....................... 91
A.2 Amount of energy saving measures in the dataset ......................... 92
A.3 Energy labels in the dataset ............................................ 92
A.4 Building types in the dataset ........................................... 92
Chapter 1

Background

Over the past years, energy consumption by buildings has steadily increased to between 20 percent and 40 percent in developed countries, being 40 percent of all energy consumption and 36 of all CO$_2$ emissions in the EU. While there is a trend focusing on reducing energy use by cars and making industries more energy efficient, buildings even account for a higher share of energy use than the transport and the industry sector (Pérez-Lombard, Ortiz, & Pout, 2008). When considering the energy usage by buildings, heating, ventilation and air-conditioning (HVAC) systems account for almost half of energy consumption in buildings and between 10 percent and 20 percent of all energy used in developed countries. Although new buildings in the EU are required to comply with standards set by the European Commission, it is clear that many older buildings are far less energy efficient and often require large investments to reduce energy consumption. Energy budgets often limit investments in energy efficiency, even when it is considered as an investment opportunity with a positive net present value (Jackson, 2010). This makes clear that large reductions in CO$_2$ that are possible and profitable only occur when the investment budget is sufficient. Ploos Energieverlening (former Ploos van Amstel Milieuconsulting) has introduced a service called Zero Budget Sustainability. This service is aimed at customers that want to reduce their energy consumption but lack investment opportunities. To overcome this problem, external partners are approached that want to invest in these energy consumption reducing measures. When energy consumption forecasts can be estimated with high accuracy, the risk, interest and maturity estimations of the investment can also be improved. Potential customers and investors are then faced with reliable and credible forecasts that can convince them to use the Zero Budget Sustainability service offered by Ploos Energieverlening.

Although this service was introduced in 2011, so far only two customers have decided to use the service. Multiple reasons have contributed to this problem. For external financers, the current model does not provide adequate information about the financial risks and the maturity of the investment. Also, the service is not promoted well enough to potential customers resulting in lack of interest. Ploos aims at promoting the properties of their service by showing that data mining can improve the estimations concerning energy saving measures. Therefore, this thesis is written to prove that data Mining can indeed improve energy consumption forecasts after the proposed investments have been implemented and also to promote the Zero Budget Sustainability service.

Each educational semester, Ploos employs a student to tackle data related problems. Previous work in energy consumption forecasting has been done by Pedroni in 2015. This study was aimed at creating a model to provide insight in energy consumption for customers but also to predict future energy consumption. This was mainly on the short term, for example predicting the consumption for the next week. Following the results of this work, the company has decided to extend the possibilities of data mining and place it in a specific context, as the benefits
of this approach have become clear.

This thesis is written for three groups of readers. First, there is the academic field. It is aimed to extend the scientific literature by answering the research question and provide new insights in the subject of data mining. Second, the company (that is described in Section 1.1) can use the report and the model in their approach towards customers. Finally, customers can use the report and the results of the model in their decision to apply the ZBS service.

Beforehand, it has to be noted that some parts of this thesis can be difficult to read without background in mathematics or software coding. It is therefore advisable for those readers to focus more on Chapter 1 which describes the problem and Chapter 7 where the conclusion is stated. These chapters are written such that the core of the research project is clear to the reader.

1.1 Company description

Ploos Energieverlening (former Ploos van Amstel Milieuconsulting) has been founded in 1995 by Joop Ploos van Amstel. At the moment, the company is only involved in energy related subjects and aims to reduce the energy usage and costs of their customers. Moreover, Ploos supports their customers with specific advise and services. In general, the company suggests to invest in those projects where each Euro invested has the maximum yield. Besides that, Ploos also aims to reduce CO₂ emissions. This is clearly displayed on their website, which shows the amount of CO₂ their customers are saving on each specific day. Ploos director Joost Ploos van Amstel has described that "current technology and financial resources are sufficient to reach the 2020 global CO₂ targets".

Ploos’ customers are very diverse, ranging from supermarkets and butcher shops to healthcare suppliers and distribution partners. If a customer decides to use one of the services offered by Ploos, energy saving targets are agreed on beforehand. This gives customers insight and confidence in the minimum amount they will save on their energy consumption. Ploos guarantees this amount before the proposed service is agreed on. The company logo is displayed in Figure 1.1.

![Figure 1.1: Ploos Energieverlening logo](image)

Ploos has developed a web application service as well as other online products. These products are offered to customers to help them to comply with standards and regulations, checking and processing invoices, benchmarking and many other services. Also, the budget neutral (Zero Budget Sustainability) service is offered. In short, ZBS offers customers the possibility to invest in energy saving while maintaining the level of their current energy budget. If the
customer and financer agree, there is even a possibility to reduce the energy budget immediately to somewhere between the original and the proposed future budget. More about ZBS can be found in Section 1.4.

In December 2015, the company has been taken over by the Facilicom Services Group. Facilicom is a large facility service supplier that is active in cleaning, catering services and security. By acquiring Ploos, the company aims to become active in supplying energy services as well. For Ploos, the main benefits of the take-over are access to a larger network of potential customers and integration of services in the future. The Facilicom logo is displayed in Figure 1.2.

![Facilicom logo](image)

**Figure 1.2: Facilicom logo**

### 1.2 Data mining and Big Data

During the last few years, the subject of *Big Data* has received much attention. In recent years, the possibilities of Big Data have increased due to improved computation methods, better analytics and more availability of data. Despite the attention from media, many people still believe that Big Data simply means "a lot of data". In fact, Gartner\(^1\) describes Big Data as: "High-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable insight, decision making and process automation". This is graphically illustrated in Figure 1.3.

![Overview of Big Data](image)

**Figure 1.3: Overview of Big Data**

\(^1\)http://www.gartner.com/it-glossary/big-data/
Figure 1.3 shows that over time, the velocity, volume and variety of data have increased. This also holds within the energy sector. For example, the implementation of smart meters has increased the velocity of consumption volumes from once a year to one measurement each 15 minutes (which can be considered as nearly real time). One can imagine that this increase of datapoints also affects the volume of data stored. This requires more storage capacity and improved computation methods for deriving information and actionable results. Finally, the data variety has evolved from simple data tables to all kinds of data storage, ranging from pictures to sound recordings. Today, even handwritten letters can be processed!

Increase in data volume, variety and velocity creates possibilities for a wider range of more accurate models. Whereas in the past the energy consumption was predicted on a yearly or monthly base, increase of data velocity (with thanks to smart meters) has enabled analysts to estimate energy consumption on a 15 minute interval.

Data mining has been described as "knowledge discovery in databases" (Fayyad, Piatetsky-Shapiro, & Smyth, 1997). This scientific field aims to develop methods and techniques to map low-level data into compact forms such as a report. The most important reason is that this low-level data is too voluminous for understanding. By using a descriptive approximation the data can be made more useful. An automated approach of data processing is needed because doing it manually requires too much time, is not cost-effective and is impractical. Moreover, analysts today are confronted with a data overload that requires different approaches that those who have been used in the past.

"Knowledge is power" is a quote that is used for a long time. This also holds for energy related data. When used for the right purposes, knowledge of energy data on a personal level can help to achieve a cost reduction and can be used for optimizing energy savings. However, if not properly used privacy issues can be violated and persons or companies can be harmed. Therefore, it is important to note that all information used in this thesis and in general by Ploos is treated with respect for privacy and security. Moreover, data is used in such a way that it will always benefit both Ploos and its customers.

1.3 Problem description

Energy budgets often limit investments in energy efficiency, even if the investment has a positive net present value. Until now, investments in energy efficiency involve uncertainty considering the exact financial gains. In financial terms, uncertainty implies risk in investments. Another problem is that few comparable investment opportunities are available when considering non-financial investments such as energy efficiency for buildings. The complexity of energy efficiency investments generally forces firms to apply simple payback rules to evaluate investments. As a result, bounded rationality is one of the major problems when it comes to these investments, as well as satisficing behavior. Firms consider investment in energy consumption reduction simply not as a profitable investment. Considering ZBS, five problems have been crucial for the lack of success of the service. First, external financers are conservative for investing. Many small and medium enterprises face lack of access to external financial resources today. Second, their trust in the ZBS service is insufficient and financial risks concerning the project are estimated to be high, partly due to the bankruptcy risk of the customer. This problem
could be overcome by the introduction of higher risk premiums, but it would reduce profitability for customers and thus is not a preferable option. Three main reasons for lack of interest of potential customers in the service are present: potential customers are not convinced that the proposed energy saving measures will turn out to realize the projected savings. They are also reticent in transferring their invoices to an external party. Finally, energy bills make up approximately one or two percent of most firm’s costs. As a result, energy expenses are not taken serious as a possible cost reduction. Both financers and customers have indicated that a more scientific approach to the results that ZBS proposes is needed to convince them of the advantages it offers. Important to note is that both financers and customers value independent research by and external institute as crucial for their decision to apply ZBS to their situation. In order to illustrate the problems regarding ZBS, a hypothetical scenario has been written. This scenario can be found in Appendix A.2.

Current literature on energy forecasting mainly focusses on predicting energy consumption under comparable circumstances. However, Zero Budget Sustainability is intended to improve energy efficiency. As a result of that, circumstances and building’s characteristics will change dramatically after the energy consumption reduction measures as proposed are implemented. This requires a different approach compared to most literature that is now dominant in the field. In other words, suitable techniques for the required problem have to be discovered and analyzed.

1.4 Zero Budget Sustainability

With the current technical solutions, it is possible to realize the global CO\textsubscript{2} targets. However, due to insufficient energy budgets and lack of interest and trust of both customers and financers many energy efficiency improvements are not realized. Therefore, the Zero Budget Sustainability (ZBS) construction has been created. ZBS enables that energy saving investments can be implemented without increasing the energy budget or other out of pocket costs. If the organization uses the online tool Central Invoice Processing (CIP), the complete energy budget is transferred to Ploos and is used to pay energy suppliers. It has to be noted that the energy budget will be managed by a foundation called Stichting Derdegelden. This foundation can spend the energy budget only on paying energy suppliers, financers and suppliers of energy consumption reducing investments. After the energy saving investments have been implemented, Ploos continues to receive the energy budget even though energy consumption and costs are decreased. The energy budget is then used to pay the suppliers and investors of the energy saving investments. This will continue until the contract is discontinued, which depends on the investments combined with the realized savings in energy. Also, 0.5 percent of the energy budget will be used to create a guarantee fund. This fund can be used to pay investors back in case one of the customers faces bankruptcy.

Besides that, ZBS offers customers to provide project management and a CO\textsubscript{2} ambition level. In case the customer chooses to use the project management service, Ploos will handle all the energy contracts and makes sure that the contracts fit within the ZBS targets. They also compare different energy suppliers and oversee the implementation process. There are six different CO\textsubscript{2} ambition levels, from 0 percent in equal steps to 100 percent. The percentage implies which share of the avoided inflation in the energy budget can be allocated to CO\textsubscript{2} reduction. Ploos
ensures that at the end of the ZBS process, energy costs are at a maximum of 50 percent compared to the original budget, although under specific circumstances a lower reduction is also possible. Figure 1.4 shows the structure of ZBS. In general, this scheme makes clear that Ploos acts as an intermediary for invoices and finances. Also, the foundation that takes care of invoices and finance is displayed.

Figure 1.4: Zero Budget Sustainability structure

In terms of time and costs, ZBS can be described by the following example. Consider a building that can be made more energy efficient by placing solar panels on the roof. The owner of the building decides to use the ZBS construction to finance the investment. Figure 1.5 illustrates this.

Figure 1.5: ZBS in terms of time and costs
Before the ZBS contract is signed, the energy expenditures are at a level of 100. The amount that is actually spent on energy bills is indicated by the red line in Figure 1.5. At time $t_0$, the ZBS contract is signed and the solar panels are placed on the roof of the building. This takes until time $t_1$, when the installation is complete. In the time between $t_0$ and $t_1$, costs are made to install the solar panels and these are made by the external financer. At time $t_2$, the investment has been paid off so now the energy bill can be lowered to the level that Ploos has calculated. This is also the time when the ZBS contract is ended. From now, the energy costs remain at the level of 50 until the end of the technology life $t_3$. The blue surface indicates the amount that is actually spent on energy. Note that for simplicity this graph is based on a few assumptions. Energy savings start when all investments have been made. Also, inflation and the cost of capital are not taken into account. More about this can be read in Section 2.2.

1.5 Buildings’ energy consumption estimation

In order to derive the best possible results in energy consumption forecasting, two attributes are of main importance; selecting the right techniques and methodologies and selecting the right predictors. When considering the variables, difference can be made between internal and external variables. Internal variables consist of factors that are company and/or building specific. Another important variable is considered the nature of the company. In a consultant bureau, fewer employees are expected to be at the office because consultants are often working at their customers’ office. Therefore, occupancy is expected to be of influence on energy consumption. This can be described by occupation rate. However, this data is supplied by customers and is sometimes difficult to obtain, as will be shown later.

Building characteristics are also important for energy consumption forecasting. For example, the energy label a building receives is a good predictor for energy usage per square meter. However, the energy label by itself is expected to be insufficient as a predictor so total floor space is expected to influence total energy consumption as well. External variables consist of weather measurements. Variables that can be included are temperature, solar radiation, humidity ratio and wind speed (Li, Su, & Chu, 2011).

Energy consumption can be split into three components, being electricity, gas and heat. Note that when heat is supplied, no gas fired heating equipment is available. It is clear that not all variables will influence both electricity and gas (or heat) consumption. Therefore, it might be necessary to split total energy consumption into these three components in order to derive optimal results.

Finally, it has to be noted that total energy costs are dependent on two components; energy consumption and the energy price. More information about the energy price can be found in Appendix A.4. This thesis will focus only on energy consumption estimation and not on energy price forecasting. The main reason for this choice is that energy consumption can be partially influenced itself by the company that is using the energy. The energy price is mostly dependent on external factors such as the global oil price and energy taxes. Companies can do little to reduce the price they pay for their energy except for choosing the cheapest energy supplier. Therefore, gas consumption and electricity consumption will be the dependent variables in this thesis.
1.6 Research questions

Ploos has stated that potential customers and financers have to be convinced by the advantages of the ZBS service. To achieve this goal, a working model has to be created that can estimate future energy consumption of a building after the proposed energy saving measures have been implemented. This thesis will therefore focus on applying the best suited techniques and the best of the available data that will enable the estimation of future energy consumption of buildings where the ZBS concept is applied. Important aspects are that the models can be applied within the time span of this thesis, that they are easily understandable, produce good results and short computation times. The research question will therefore be:

- Can the energy consumption of buildings be predicted with high accuracy within the context of ZBS?

1.6.1 Sub questions

To answer the research question, two sub questions need to be addressed. The first sub-question is the following:

- Which variables are good predictors for energy consumption based on the available data?

This sub-question will be answered based both on scientific literature as well as expert knowledge obtained from Ploos employees. Another approach to select the data is feature selection (Guyon & Elisseeff, 2003). However, for this thesis it has been chosen to use a domain knowledge approach since much knowledge is available via Ploos employees. Another reason is that some data is hard to obtain, it requires special attention and takes a lot of time to obtain. Indeed, if deep domain knowledge is available, selecting a set of ad hoc features is preferred (Guyon & Elisseeff, 2003). It has to be noted that this question also heavily depends on the data that is available and the data that is supplied by external sources. For example, organizational variables might not be supplied because of privacy considerations. Another sub-question that has to be addressed is:

- Which models are applicable to the described problem?

This question will be answered using scientific articles and books. Several techniques will be studied and previous research projects will be described. Then, the most suitable techniques will be selected and applied to the problem.

1.7 Methodology

To address the research question and the sub-questions from Section 1.6 a data mining approach is required. Many different data mining methodologies exist, including the most common used methods CRISP-DM and SEMMA. This thesis is build according to the CRISP-DM (Cross-Industry Standard Process for Data Mining) process. CRISP-DM is by far the most used method among data mining practitioners (Harper & Pickett, 2006) and also the most preferred one (Nadali, Kakhky, & Nosratabadi, 2011). The steps in the CRISP-DM will be described briefly in Chapter 3. Figure 1.6 shows the steps CRISP-DM framework.
During the first phase, a literature study is conducted which indicated the most suitable models that can be applied. During this phase expert knowledge was used to determine the right predictors. This was followed by the next two phases, in which the data was studied and prepared for analysis. Later, the model was build. In the next step, the models were evaluated for their performance. In this particular phase, the main research question was answered. The final phase is aimed at delivering a report, presenting the results and defend the thesis.

![Figure 1.6: An overview of the steps in CRISP-DM](image)

CRISP-DM was proposed as an industry-wide standard methodology for data mining in the 1990’s. Figure 1.6 shows the process that has six steps and starts with business understanding and ends with deployment. As can be obtained from Figure 1.6 the process can have multiple iterates. This is due to the nature of data mining, which is driven by the knowledge and expertise of the analyst. It is important to pay extra attention to the earlier steps because the latter steps are built on the outcomes of the earlier ones.

### 1.8 Project plan

This thesis adresses the problems as described in Section 1.3. More specific, the aim is to build a model that can predict energy consumption of buildings with high accuracy. Then, the results are compared to the calculation module that Ploos is currently using. If the results are comparable, confidence of both customers and financers in ZBS can be increased and the results can be used to promote the service, which is important for Ploos. To predict energy consumption, the focus is on the most relevant information that is available. In the model, climate data, customer information and building properties are used. Information about the input data and the models is obtained from expert knowledge and scientific literature. Although financial aspects are taken into consideration, it is specifically not the aim of this thesis to predict the energy price. Therefore, the energy price is taken as given. Furthermore, the focus of this thesis is on commercial buildings since from these buildings more relevant information is available.
Data is obtained only for buildings in which energy saving measures have been implemented. It was found that the status of the measures are not up to date so the information in the system had to be updated before a start could be made with analyzing the data. The planning of the project is aimed at following the steps of CRISP-DM sequentially, although that is not possible in all cases. During the writing of the literature review, part of the data understanding and data preparation phase has been performed as well. This has been done because the gathering of data was expected to be difficult and time consuming. Overall, the time span of writing the thesis was eight months.

1.9 Thesis structure

The next chapters of this thesis are structured as follows. Chapter 2 describes the literature review that has been conducted in order to answer the main research question and the sub-questions. Chapter 3 explains the CRISP-DM methodology that is applied for this thesis. Furthermore, Chapter 4 describes the dataset that was available for research purposes. Chapter 5 goes into the details of the applied models. Chapter 6 gives the results and finally, Chapter 7 describes the conclusion and recommendations that follow from this research project. Also, an Appendix A is included to provide additional details next to the main body.
Chapter 2

Literature review

This chapter describes relevant literature. First, the search method that is used to obtain scientific literature is described. Then, the problem statement will be described into further detail which shows the managerial need for this thesis. This is done by relating the concept of ZBS to problems and solutions from scientific literature. It will be shown that the problems in energy investment can be overcome by introducing the ZBS concept. Furthermore, previous research is described from which suitable techniques are obtained, followed by a review of the concept of machine learning. Finally, the most relevant techniques are presented which are clustering, case-based reasoning and artificial neural networks.

2.1 Search method

Building’s energy consumption estimation is a field that has been studied for many decades. In large cities, buildings consume up to 60 percent of all energy (Steemers, 2003). As a result, hundreds of studies related to energy consumption of buildings can be found, some more relevant than others. Therefore, it is important to perform search using a structured approach.

To obtain scientific articles, a search query in a scientific database has to be performed, or printed material has to be accessed. For this thesis, three databases have been used. First, the library of Eindhoven University of Technology (TU/e) has its own collection of scientific articles that are accessible for students and lecturers. The online database is called “focus” and contains many full-text articles, E-books, references to publications outside TU/e library collection, as well as indications of the availability of printed books. Other databases that have been used are Google Scholar and ScienceDirect. This database mainly contains articles related to engineering, life sciences, health sciences and social sciences.

A starting point for a literature review is to look at review articles. In most cases, these articles briefly describe the history of the field but also the latest developments including the ”state of the art” techniques. For example, the article by Zhao and Magoules from 2012 provides an overview of recent work in building’s energy consumption prediction, including the most relevant techniques that have been used. This article was found by typing the search terms ”review energy consumption prediction” into TU/e’s focus database. These search terms also obtained the article by Pérez-Lombard et al., from 2008. In return, by using these review articles, more specific search terms can be obtained. For example, the article by Monfet, Corsi, Choinière, and Arkhipova from 2014 was found by simply adding the technique (case-based reasoning) to the search query.

Another approach for obtaining information is to have a look at cited articles. By using this search method, a better overview of the scientific field is obtained and the development of mod-
els can be studied into more detail. Also, information regarding data was extracted this way. For example, the article by Li et al. describes relevant climate variables and was found using the list of cited articles in the article of Monfet et al.

Finally, some information regarding the publication can be of importance. If a paper is cited more often, this can be a sign of the quality or relevance, although in general older papers have been cited more than recent ones. Also, the publication date of the article is taken into consideration. When looking for the state of the art technology, recent articles are more important simply because they describe the latest developments in the field.

Many online databases show publication years for the results of a specific search query, although not all results are relevant. This can be used to get an overview of when a specific subject was in the center of attention. For example, according to the TU/e database, neural networks got most attention in 2012. This example is displayed in Figure 2.1.

2.2 Investment in energy consumption reduction

Energy consumption reduction results in savings on energy costs. A profitable energy efficiency investment exists if the discounted value of savings exceeds the investment costs. As a result, the Net Present Value (NPV) of an investment can be calculated using the traditional capital budgeting investment formula [Jackson, 2010]:

\[
NPV = \sum_{t=1}^{T} \frac{S}{(1+i)^t} - I
\]

In which \(S\) represents savings, \(i\) is the interest rate, \(I\) equals the investment cost and \(T\) represents the technology life. When a risk factor \(r\) is incorporated into the investment decision, the NPV equation changes to:

\[
E(NPV) = \sum_{t=1}^{T} \frac{E(S)}{(1+i+r)^t} - E(I)
\]

Note that for the Net Present Value, the savings, the technology life and the investment cost expectations are given. This implies that uncertainty affects all variables in the equation.
Also, risk factor $r$ is difficult if not impossible to estimate (Jackson, 2010). Note that investments are considered as initial. Experience obtained by Ploos indicates that this is a realistic assumption. For example, placing solar panels on a roof requires a few days while the life cycle is estimated to be 15 years. Within this context, even if some of the payments follow on a later date, investments can be considered as initial.

This Net Present Value estimation is also made in the current calculation module used by Ploos. This will be described in more detail in Subsection 3.1. In practice, most firms are conservative when considering risky investments. The fact that the savings are difficult to calculate and the maturity of the investment is unknown simply drives firms to invest in projects that have less risky outcomes. Most firms apply simple payback rules of thumb when considering the energy efficiency investment environment. Indeed, many firms are not using optimal calculation methods for energy efficiency improvements (DeCanio, 1998).

A solution to the unknown risk-problem can be found in supply chain finance, by means of the factoring principle. In factoring, firms sell their creditworthy accounts receivable at a discount (Klapper, 2006). Factoring is a financial service and includes credit protection, accounts receivable bookkeeping, collection services and financing. Many factors will only buy a complete portfolio because the risk of default is lower compared to only a single asset. However, creating a diversified portfolio requires factors to collect credit information and credit risk has to be calculated (Klapper, 2006).

In general, factoring has many barriers to overcome. One solution for this problem is the so-called Reverse Factoring. This construction implies that the lender only buys accounts receivables from specific informationally transparent and high quality buyers. Traditional factoring and reverse factoring have in common that the credit risk is equal to the default risk of the high quality customer, and not the risky firm. An example of factoring in practice is GMA factoring. Figure 2.2 gives a schematic view on reverse factoring.

![Figure 2.2: Schematic overview of reverse factoring](http://www.factoring-invoices.com/)
Literature suggests that lack of information transparency is one of the most critical issues in reverse factoring (Klapper, 2006), (Iacono, Reindorp, & Dellaert, 2015).

2.2.1 Relation to ZBS

Although ZBS is not completely comparable to factoring or reverse factoring, some similarities are obvious. First, instead of accounts receivable in factoring, ZBS is aimed at reducing costs by generating savings on energy expenditures. The factor in factoring would be the external financer. The small and medium enterprises that are considered as risky in the literature (Klapper, 2006), (Iacono et al., 2015) would in this case be represented by the company that wants to reduce energy consumption. Ploos in this case does not have an equivalent in the theory regarding factoring, because Ploos is just an intermediator or facilitator.

In this context, it is likely that information asymmetries or lack of information considering investment opportunities in energy consumption reduction, (Jackson, 2010) are also a major part of the current problems that ZBS is facing. Therefore, this thesis aims to improve energy consumption reduction estimations by applying the best possible techniques to the available data. When considering the specific case of ZBS, the equations in Section 2.2 can be applied to the described problems. Since customers and investors are interested in more accurate calculations, the technology life $E(T)$ can be regarded, as well as the expected investment costs $E(I)$. However, these variables are not considered as extremely uncertain because for most investments the costs are known as well as the technology life. And then there is the risk factor $r$. Risk is estimated separately for each specific case and as was stated by Jackson (2010), difficult to calculate. It also requires company specific information regarding finance that is not available. However, by creating the guarantee fund as described in Section 1.3 the risk factor $r$ is reduced. Therefore, this thesis only focuses on estimating the expected savings $E(S)$ from investments.

2.3 Energy consumption estimation techniques

Building's energy consumption forecasting has been a field of study for a long time. Following the oil crisis of 1973, much attention has been devoted to increase the energy efficiency of buildings (Van Raaij & Verhallen, 1983), (Hong, Chou, & Bong, 2000). A sharp increase in energy prices during this period resulted in an immediate decline of energy consumption of both commercial and residential buildings (Hirst & Jackson, 1977). One of the first attempts to predict energy consumption of buildings was the DOE-2 program. This whole-building simulation program was able to predict energy consumption of a building using weather parameters and building characteristics (Z. Yang & Becerik-Gerber, 2014). Since its introduction, it has been optimized and new features have been added. Especially since the early 1990’s, building simulation programs have been in the interest of many researchers. Most studies differ between residential and commercial buildings. Within residential buildings, difference is made between a top-down and a bottom-up approach (Swan & Ugursal, 2009). Top-down models use long-term macroeconomic indicators such as GDP\(^2\), housing construction rates and climatic conditions to determine the effect on energy consumption. Bottom-up approaches rely on information from less than a sector as a whole, such as variables measured at a household level.

\(^2\)Gross Domestic Product
Within the top-down approach, differences can be made between econometric and technological models. Bottom-up techniques include statistical and engineering models. An overview of the techniques is provided in Figure 2.3.

![Figure 2.3: Overview of techniques for estimating residential energy consumption](image)

Advantages of top-down approaches and specifically econometric models are that they are easy to develop and only need limited information provided by macroeconomic indicators, which are widely available. A disadvantage of these methods is that they do not look at the individual level and often lack the impact of future technological improvements.

As the primary information source of statistical models within the bottom-up approach is energy data provided by energy suppliers, lack of availability is sometimes a drawback of these models. For engineering methods such as DOE-2, detailed building information needs to be available. These models calculate or simulate the energy consumption and do not take into account historical energy consumption. A major drawback is that engineering methods make assumptions about occupants’ behavior, rather than taking it into account. Indeed, as an early study by Pettersen indicated that for residential buildings, inhabitant’s behavior is most important for estimating energy consumption. This study already took into account climate data, building characteristics and inhabitant’s behavior. When the behavior is unknown, climate variables are far less important compared to a situation where behavior is predictable. Later works indeed showed comparable conclusions.

Besides residential buildings, effort is also spent at estimating the energy consumption of commercial buildings. For commercial buildings, roughly the same techniques are applicable and are used although behavior is generally more predictable and thus less important as input parameter. Instead, the type of customer is regarded as a good predictor, as Ploos uses this to differ within their customer base.

Another difference in approach is that some researchers focus on improving energy efficiency rather than forecasting using constant parameters. Ahmed, Korres, Ploenings, Elhani, & Menzel have performed research in which building characteristics and climate conditions have been connected to energy efficiency. In this study, building characteristics are connected to the buildings’ performance by using data mining techniques. It was aimed to predict thermal classes and indoor daylight levels for optimal room scheduling. Room scheduling can
then be used to schedule meetings to the most energy efficient room. Weather data includes historical weather values and weather forecast data. Three different models were used to predict low-energy comfort rooms. A Naive Bayes classifier turned out to be the best predictor, outperforming decision tree and support vector machine.

Dalamagkidis, Kolokotsa, Kalaizakis, and Stavrakakis used a linear reinforcement learning controller (LRLC) that uses thermal comfort, indoor air quality and energy consumption as signals to the agent. After training the model on four years of data, results are comparable to those of On/Off controllers and Fuzzy-PD controllers. However, the model still makes significant errors such as turning on the heat in the summer or the cooling in the winter.

Instead of focusing on energy consumption, some researchers took into account the energy price as a measure to improve energy efficiency. Neill, Levorato, Goldsmith, and Mitra presented a Consumer Automated Energy Management System (CAES). This algorithm used reinforcement learning as a technique to estimate the future energy price and schedules residential energy consumption by switching on and off devices. Another advantage of this algorithm is that it reduces peaks in energy consumption. As a result, energy costs of households are reduced by 16 to 40 percent.

Since the 1970’s, architects and engineers have used calculation modules to determine the indoor daylight of buildings. In the 1980’s, adaptations of the DOE-2 program were created to determine the effect of daylighting on peak electrical demand (Choi, Johnson, & Selkowitz 1984). The first attempt to use daylight levels to determine energy consumption was made using an adapted version of the DOE-2 program (Winkelmann & Selkowitz 1985). In this analysis, the energy consumption and cost-related effects of daylighting strategies are calculated. After more than two decades, the DOE-2 program has been replaced by a new energy simulation tool, being EnergyPlus (Crawley et al. 2001). This software package is still used and maintained today.

Since the 1990’s, artificial neural networks (ANN’s) have gained attention as a technique for energy consumption estimation (Kalogirou 2001). ANN’s are applicable for many energy systems including the estimation of a building’s thermal load, predicting air flow and the prediction of energy consumption of buildings. Compared to the earlier engineering methods, ANN’s are able to produce equal and sometimes better results than simulation tools (Neto & Fiorelli 2008). ANN’s can also be used as a technique for long-term energy consumption estimation on a macroeconomic scale (Ekonomou 2010), (Sözen, Akcayol, & Arcaklioglu 2006).

Compared to traditional statistical methods, ANN’s have some advantages including that they can execute more tasks than a linear program, their learning capability and their ability to be executed in any application (Ahmad et al. 2014). Recently, artificial neural networks have become the most widely used artificial intelligence models in predicting energy consumption of buildings (Zhao & Magoules 2012). More about neural networks can be found in Section 2.7. Earlier work has indicated that ANN’s are the most suitable for buildings with non-linear energy use patterns because they can approximate almost any continuous function, even with only one hidden layer (Chen, Billings, & Grant, 1990). Non-linear consumption patterns are a common feature for buildings (Karatasou & Santamouris 2010). In these ANN’s, many variables can be included.
An early study by Park, El-Sharkawi, Marks, Atlas, & Dambor (1991) showed that ANN's are able to predict daily load forecasting and clearly outperform previous approaches including time series and regression. This study used historical load and climate inputs as predictors. Instead of predicting the exact daily load, Figueiredo et al. (2005) presented an electricity consumer characterization framework. In this framework, a knowledge discovery in databases procedure is made, supported by data mining techniques. The presented framework is composed from two modules, the load profiling module and the classification module. Classification is based on load profiles. The structure and process of the framework are shown in Figure 2.4.

Figure 2.4: Structure of the customer characterization framework, obtained from Figueiredo et al. (2005).

Load profiles are based on consumption patterns. In the classification step, the major goals are the inference of a rule set to characterize each class and to support the attribution of new customers to the classes obtained by the load profiling module (Figueiredo et al., 2005). The classification model is built to assign new consumers to the existing ones. More about classification and clustering can be found in Section 2.5.

Wong, Wan, & Lam (2010) performed research on energy consumption prediction of buildings located in Hong Kong. A neural network was created that used four different weather parameters, four building design parameters and a day type were included. Output is measured in electricity consumed for cooling, heating, lighting and total electricity usage. ANN’s have proved to be able to reach high accuracy for predicting energy consumption.

Another study by Ekici & Aksoy (2009) takes into account two different types of inputs for
buildings. First, physical environmental records such as temperature, solar radiation and wind speed. Second, artificial designing parameters (transparency ratio, orientation, materials used etc.) are considered. In this project, it was shown that ANN’s can achieve a success of 90-99 percent in predicting heating load.

Finally, Grey models are models where information of a system is only partially known. It can be used when the dataset is incomplete or when there is uncertainty about the data (Zhao & Magoules, 2012). But grey models have not gained as much attention compared to other techniques used for energy consumption estimation.

(Monfet et al., 2014) state that although ANN’s produce good results for the prediction of energy consumption, Case-based reasoning (CBR) has some strong advantages over neural networks. Case-based reasoning is described in more detail in Section 2.6. First, the knowledge of the CBR case library can be updated more easily. In an CBR system, the new case only has to be added to the case library whereas a neural network requires to be retrained. Furthermore, the results of CBR are easier to understand and can be presented without deep understanding of the subject (Turban, Sharda, & Delen, 1997). Another advantage is that CBR can handle large amounts of inputs and outputs while an ANN can be troublesome in handling many features. Finally, CBR is better in handling missing information. An application of CBR is given by (Koo & Hong, 2015) who developed an energy performance curve which is categorized into energy labels.

The right technique depends on the purpose of the research project and the availability of data. More information about the dataset can be found in Chapter 4. It has been obtained that clustering techniques can be used to improve the dataset. Therefore, clustering is described in more detail in Section 2.5. Regarding the classification and regression techniques that are discussed, case-based reasoning is applicable to the research problem of this thesis. However, as ANN’s are able to predict future energy consumption with high accuracy, both techniques are discussed in Section 2.6 and 2.7.

2.4 Machine learning

Machine learning, sometimes refered to as statistical learning, computational learning and pattern recognition (Alexander, 2013) basically implies that computers are enabled to acquire knowledge from data obtained from historical happenings (Turban et al., 1997). The main advantage of machine learning is that they overcome deficiencies of manual knowledge acquisition techniques. This is done by automating the learning process. Machine learning can be described as a family of artificial intelligence technologies that is concerned with the design and development of algorithms. These algorithms allow computers to learn based on historical data.

2.4.1 Supervised learning

If instances are given with known labels (the corresponding correct outcomes), learning is called supervised (Kotsiantis, Zaharakis, & Pintelas, 2006). For example, online fraud can be detected based on variables such as time of the first transaction, time between transactions and the value of the transactions. A set of outcomes of fraud cases is also available. It is then
possible to test to what extend the model is able to predict fraud cases based on the historical
data that is available.

Classification implies identifying to which of a set of categories (sub-populations) a new ob-
servation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Within classification, difference can be made between binary classification and multi-class classification problems (Kotsiantis 2007). Binary classification is composed of variables that have only two possible outcomes. Multi-class classification contains variables that have more than two, but a finite number of possible outcomes. Those outcomes are referred to as labels.

For energy data, classification has been used for several purposes. An example is the divi-
sion of temperature into thermal classes (hot, warm, neutral, cold etc.) and indoor daylight
levels (enough, not enough) ((Ahmed et al., 2011). Another application is given by (Figueiredo et al., 2005), where consumers are classified according to load profiles. In other words, a finite number of outcomes. However, classification is not used to predict the exact energy consump-
tion of buildings since these outcomes can take an infinite amount of values.

When a problem can have an infinite number of outcomes, it is regarded to as regression. Regression is a learning function where data items are mapped to a real-valued prediction variable (Fayyad et al. 1997). A good regression model is able to predict future results. For example, estimating future energy consumption of a building falls in the category regression. Many examples can be found in the scientific literature, ranging from simple statistical re-
gression (Pettersen 1994) to more advanced methods as ANN’s (Park et al., 1991), (Ahmad et al., 2014) (and many more), and Case-Based reasoning (Monfet et al. 2014) and (Koo & Hong, 2015).

**Training and testing**

Supervised learning systems require training and testing. The description of the supervised learning problem is obtained from the article by Dietterich:

In the standard supervised learning problem, a learning program is given examples in the form of \((x_1, y_1), \ldots, (x_m, y_m)\) for an unknown function \(y = f(x)\). The \(x_i\) values are mostly in the form of \(<x_{i,1}, x_{i,2}, \ldots, x_{i,n}>\). Values of the components can be discrete or have real numbers. The \(y\) values are drawn from a discrete set of classes \((1, \ldots, K)\) when a classification problem is considered or have real numbers in case of a regression. Given a set of \(S\) training examples, the learning algorithm outputs a classifier or a value. This output is an hypothesis about the true function \(f\). For a given number of new \(x\) values, it predicts the corresponding \(y\) values, based on what the system has obtained from the training set \(S\).

**Sample split and cross-validation**

The data is partitioned into two mutually exclusive subsets called training set and test set. In case of a neural network, the data is partitioned into three subsets, being the training, testing and validation set. This validation set is used to avoid overfitting of the network to the training set. During the training, the error on the validation set is monitored until it is increasing,
which is a sign of overfitting. Neural networks are described in more detail in section 2.7. A common approach is to divide the dataset into a training set and a test set, roughly by one-third two-third. A drawback of this approach is the problem of overfitting, which implies that the model is biased to the training set. Therefore, cross-validation is the most widely used error prediction technique (Efron et al., 2004). For computational reasons, 5-fold or 10-fold is preferred (Fushiki, 2011). In general, this is referred to as K-fold cross-validation. In this methodology, the dataset is randomly split into $k$ mutually exclusive subsets of about the same size. The number of $k$ implies the number of testing and training sessions. In each session, the model is trained on all but one of the subsets and tested on the remaining one. Then, the overall accuracy is calculated by:

$$ CVA = \frac{1}{k} \sum_{i=1}^{k} A_i $$

In which $CVA$ is the cross-validation accuracy, $k$ is the number of folds and $A_i$ the accuracy measure. Figure 2.5 provides more insight in K-fold cross-validation.

**Figure 2.5: K-fold cross-validation**

**Accuracy estimation**

For classification problems, a *confusion matrix* (or classification matrix) is used to estimate the accuracy. The confusion matrix for a binary classification problem (where only two classes are present) is displayed in Figure 2.6.
Where True Positive (TP) and True Negative (TN) are correctly predicted and False Negative (FN) and False Positive (FP) are incorrectly predicted. If the classification problem is multi-class, the matrix has as many rows and columns as there are classes. In both cases, the accuracy is measured by:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

There are mainly two reasons for measuring the accuracy of classification problems. It gives a certain confidence in the prediction quality for future cases and it can also be used to compare different classification models.

For regression problems, many commonly used accuracy measures exist (Hyndman & Koehler, 2005). Difference can be made between scale-dependent measures, measures based on percentage errors, measures based on relative errors and relative measures.

Scale-dependent measures are helpful for comparing different models on the same dataset, but can not be used if the data has different scales. Of the scale-dependent measures, the Root Mean Square Error (RMSE) is the most preferred one, although it is sensitive to outliers (Hyndman & Koehler, 2005). The RMSE is common for measuring the accuracy of energy consumption prediction (Tso & Yau, 2007) (Monfet et al., 2014). The RMSE is described by:

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2}
\]

In which \( y_t \) denotes an observation at time \( t \) and \( \hat{y}_t \) denotes the forecasted value at time \( t \). For energy savings data, the Coefficient of Variance Root Mean Square Error (CV-RMSE) is used to determine goodness of fit (Haberl, Claridge, & Culp, 2005). This should be below 25 percent. The CV-RMSE is described by:

\[
\text{CV-RMSE} = \sqrt{\frac{1}{\overline{y}_t} \sum_{t=1}^{n} (y_t - \overline{\hat{y}}_t)^2}
\]

In which \( \overline{\hat{y}}_t \) is the mean value of the dependent variable test set. Percentage errors are scale-independent and can be used to compare the performance of models across different data
sets (Hyndman & Koehler, 2005). One disadvantage of percentage error measures is that for \( y_t = 0 \), they are skewed. The Mean Absolute Percentage Error (MAPE) is described by:

\[
MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|
\]

Another disadvantage of the MAPE is that a heavier penalty is put on positive errors than on negative errors. Therefore, the Symmetric Mean Absolute Percentage Error (SMAPE) can be used. The advantage of SMAPE is that it can handle negative values. A disadvantage is that for values of zero, which is common for energy consumption, it returns the maximum error of two. It is calculated by:

\[
SMAPE = \frac{1}{n} \sum_{t=1}^{n} 2 \ast \left| \frac{y_t - \hat{y}_t}{y_t + |y_t|} \right|
\]

### 2.4.2 Reinforcement learning

Reinforcement learning is situated between supervised learning and unsupervised learning. The difference compared to supervised learning is that there are no historical cases, only a set of possible outcomes. The machine learns from its new experiences. Reinforcement learning can be described as the problem faced by an agent that must learn behavior using trial-and-error interactions with a dynamic environment (Kaelbling, Littman, & Moore, 1996). Two main strategies exist for solving reinforcement learning problems. One is to search in the space of behaviors to find a behavior that performs well in the environment. The other is to use statistics and dynamic programming methods. This strategy estimates the utility of taking actions in states of the world (Kaelbling et al., 1996). Figure 2.7 shows the standard reinforcement learning model.

![Figure 2.7: The standard reinforcement learning model](image)

In the standard reinforcement model, obtained from Kaelbling et al. an agent is connected to its environment via perception and action. The agent receives as input \( i \) on each step of interaction and receives some information of the current state \( s \) of the environment. The agent then chooses action \( a \) to generate as output. This action changes the state of the environment.
and this state is communicated back to the agent through a reinforcement signal $r$. The agent’s behavior $B$ is supposed to choose actions that increase the long-run sum of values of the reinforcement signal. It learns by systematic trial and error. The $I$ is considered as an input function which determines how the agent views the environment state.

Reinforcement learning is used in energy research manly to improve efficiency. Examples such as Dalamagkidis et al. show that it can be used to reach an optimal situation using thermal comfort, indoor air quality and energy consumption. Neill et al. created the CAES algorithm that is able to reduce energy costs and smooths energy consumption.

2.4.3 Unsupervised learning

When outcomes of a set of data is unknown, unsupervised learning is used to discover knowledge (Turban et al., 1997). In unsupervised learning, the machine simply receives inputs but never receives target outputs (Ghahramani, 2004). The machine’s goal is to develop a formal framework based on the notion that it is the machine’s goal to build representatives of the input which can be used for decision making. Other possibilities are predicting future inputs and efficiently communicating inputs to another machine (Ghahramani, 2004). Clustering is an example of unsupervised learning and a good method to improve grip on for example time series data obtained from smart meters (Lavin & Klabjan, 2015).

2.5 Clustering

Clustering implies data mining methods for classifying items, events or concepts into groupings named clusters (Turban et al., 1997). Cluster analysis is an exploratory data analysis tool for solving classification problems. Its main purpose is to classify cases into clusters such that the data cases have strong similarity with other cases in the cluster and weak similarity with cases outside the cluster. The different clusters then describe the classes to which their members belong. Clustering can be used to improve the dataset, for example to define the number of classes as inputs for a model, which is common for data mining problems with a large dataset (Figueiredo et al., 2005).

When considering a population of $n$ cases (or elements) with $m$ attributes, many applications are possible. Attributes can be binary, categorical or numerical. Applications of a cluster analysis can range from clustering customers into groups to developing a business strategy (Michaud, 1997). Clustering can also be used for outlier detection (Turban et al., 1997).

For energy data, effort has mostly been spent on creating energy profiles of a specific date or customer but recently new applications such as identifying opening hours and potentials for energy efficiency have been developed (Lavin & Klabjan, 2015).

2.5.1 Partitioning criteria

Given $n$ elements with $m$ attributes, as in Section 2.5, elements within the same cluster should have small distances while elements of different clusters should have large distances (Michaud, 1997). For a given objective function $F(P)$, partitioning criteria are useful as objective functions while searching for an optimal solution. Multiple measures can be used to compute the distance
between objects including \textit{euclidean}, root mean square and \textit{Mikowski} distances \cite{Liao2005}. Of these three methods, the euclidean distance is the most wide used and has previously achieved good results in commercial energy consumption studies. Let $p = (p_1, p_2, ..., p_n)$ and $q = q_1, q_2, ..., q_n$ be two data points with $n$ dimensions. The \textit{Mikowski distance} is given by:

\begin{equation}
    d_M = \left( \sum_{i=1}^{n} (q_i - p_i)^q \right)^{1/q}
\end{equation}

The Mikowski distance is a generalization of the Euclidean distance. The Euclidean distance between datapoints is given by:

\begin{equation}
    d_E = \sqrt{ \sum_{i=1}^{n} (q_i - p_i)^2 }
\end{equation}

Another distance measure is the \textit{standardized Euclidean distance}, where each difference between datapoints is devided by the standard deviation of that particular variable. Finally, the \textit{Manhattan distance}, sometimes refered to as \textit{Cityblock distance} is a common used distance measure. It is given by:

\begin{equation}
    d_C = \sum_{i=1}^{n} |q_i - p_i|
\end{equation}

Depending on the kind of heuristics in the search proces, common approaches exist for the number of possible partitions: hierarchical, \textit{K-means} and \textit{Fuzzy C-means}.

\subsection{Cluster analysis methods}

Three common methods for partition based clustering methods will be decribed; hierarchical clustering, \textit{K-means} clustering and \textit{Fuzzy C-means}.

\textbf{Hierarchical clustering}

When considering hierarchical clustering, $n$ clusters are created which contain of one element in agglomerative clustering. In each step, a pair of clusters is merged into one. The cluster-pair that merges is determined by the best objection function value that is obtained from the merge. The hierarchical method is highly dependent on the choice of the partitioning criteria. The process stops when a reasonable number of clusters is obtained (which may be based on expert knowledge) or when the next iteration results in a large increase in the objective function. Divisive hierarchical clustering, in contrast to agglomerative clustering, starts with one cluster that contains $n$ elements. Each iteration then splits one cluster into two until a pre-defined stopping criterion is satisfied \cite{Michaud1997}.

\textbf{K-means}

The K-means algorithm was developed in 1955 and is still widely used \cite{Jain2010}. The algorithms assigns each element to one of the $K$ fixed clusters. Within a representation of $n$ elements, the K-means algorithm aims at grouping objects such that the similarities between objects within the cluster are high while similarities with objects outside the cluster are low. The cluster centers are chosen at random. The elements are assigned to the cluster whose \textit{centroid} is closest. The number of clusters is kept constant while the partition is improved
iteratively. The elements are handled one-by-one and reassigned to a cluster such that the partitioning criterion is improved most. A common ending criterion is when there are no elements reassigned (Michaud, 1997). Like hierarchical clustering, the outcome of K-means depends on the partitioning criteria.

For K-means, the optimal number of clusters can be determined using multiple criteria. In most cases, different numbers of clusters are used. Broad methods to determine the number of clusters are (Ciang & Mirkin, 2010):

- Variance-based approach: using intuitive or model-based functions of the sum of within cluster distances get extreme values
- Structural approach: Comparing within cluster cohesion versus between-cluster separation at different $K$
- Consensus distribution approach: Choosing $K$ according to the distribution of the consensus matrix for sets of K-means clustering at different $K$
- Hierarchical approach: Choosing $K$ by using results or agglomerative clustering procedure
- Resampling approach: Choosing $K$ according to the similarity of K-means clustering results on randomly perturbed or sampled data

**Fuzzy C-means**

Fuzzy clustering differs from ordinary clustering in the sense that all observations $x_j$ belong to all clusters, but with different degrees $u_{i,j}$, $U = [u_{i,j}] \in [0,1]^{c \times n}$ (Höppner & Klawonn, 2003). Let $X = [x_1, ..., x_n]$ be a dataset which is partitioned into $c$ prototypes (clusters) with centers $v = (v_1, ..., v_c)$. Then, the Fuzzy C-means algorithm minimizes the objective function of weighted distances:

$$J_{FCM}(X; V, U) = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{i,j}^{m} \|x_j - v_i\|^2$$

Subject to the constraints:

- $\forall i \in \mathbb{N}_{\leq c}: \sum_{j=1}^{n} u_{i,j} > 0$ and
- $\forall j \in \mathbb{N}_{\leq n}: \sum_{i=1}^{c} u_{i,j} = 1$

In which the parameter $m$ is called the fuzzifier and determines the fuzziness of the partition. The first constraint ensures that no cluster is empty and results in a partition into $c$ predefined clusters. The second constraint implies that all the data points have the same weight in the dataset. Fuzzy clustering is graphically displayed in Figure 2.8.
Figure 2.8: Graphical representation of fuzzy clustering

In Figure 2.8, three clusters with cluster centers $V_1$, $V_2$, and $V_3$ are shown. Also, two data points including membership grades are drawn. Datapoints can represent an observation at a specific time. The grades indicate the degree to which each data point belongs to the clusters. For this example, cluster profiles based on temperature and gas consumption are created. The Fuzzy C-Means algorithm is displayed in Table 2.1.

Table 2.1: The Fuzzy C-Means Algorithm

<table>
<thead>
<tr>
<th>The Fuzzy c-Means Algorithm (FCM-AO)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Store</strong></td>
</tr>
<tr>
<td><strong>Pick</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Guess</strong></td>
</tr>
<tr>
<td><strong>Iterate</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Use</strong></td>
</tr>
</tbody>
</table>

The Fuzzy C-means algorithm basically implies that the initial cluster centers are calculated, followed by a calculation of the distances. Furthermore, the partition matrix $U_{i,j}$ is updated. If improvement of the clustering procedure is possible, the cluster centers are changed such that the distances between the data points and the cluster centers is reduced. This process is repeated until a predefined stopping criteria is satisfied. This can be for example that the change in the partition matrix is smaller than a predefined $\epsilon$. 

32
2.5.3 Determining the number of clusters

For Fuzzy clustering, the number of clusters is pre-defined. To determine the number of clusters the VAT (Visual Assessment of cluster Tendency) and iVAT (Improved Visual Assessment of cluster Tendency) are popular methods (L. Wang, Nguyen, Bezdek, Leckie, & Ramamohanarao, 2010), where the iVAT method is preferred because it makes obtaining the number of clusters easier (Havens & Bezdek, 2012). An example of VAT and iVAT images of dissimilarity data is displayed in Figure 2.9.

![VAT and iVAT images](image)

Figure 2.9: Images that show why iVAT is preferred to VAT

In figure 2.9, pictures c) and d) show the VAT representation of picture a) and b) while picture e) and f) show the iVAT representation. These images clearly display that the iVAT algorithm can be preferred to VAT algorithm. In the VAT and iVAT algorithms, each pixel in grayscale $I(D^*)$ displays the dissimilarity between two objects where white pixels represent low similarity and black pixels high similarity. Because each object is exactly similar to itself (i.e. has a dissimilarity of 0), the diagonal elements of the picture are submatrices of relatively similar objects. The off-diagonal elements are scaled in the range [0,1]. Therefore, the cluster tendency can be obtained from the number of dark blocks on the diagonal. Let $D$ be a $N \times N$ dissimilarity matrix. The VAT Recording Algorithm is displayed in Table 2.2.
Table 2.2: VAT Recording Algorithm

<table>
<thead>
<tr>
<th>VAT Recording Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong></td>
</tr>
<tr>
<td><strong>Data:</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>for $r = 2, \ldots, N$ do</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>Obtain the ordered dissimilarity matrix $D^<em>$ using the ordering array $P$ as: $D^</em><em>{pq} = D</em>{P(p), P(q)}$. $1 \leq p, q \leq N$.</td>
</tr>
<tr>
<td><strong>Output:</strong></td>
</tr>
</tbody>
</table>

In general, the arg max and arg min functions in Step 2 and Step 3 are set valued so that the procedure selects any of the optimal arguments. VAT then produces a reordering that can be found in the array $P = P(1), \ldots, P(N)$ (Hathaway, Bezdek, & Huband, 2006).

The improved VAT algorithm as proposed by (L. Wang et al., 2010) uses a path based distance measure. This distance is defined by:

$$D_{ij} = \min_{p \in P_{ij}} \max_{1 \leq h < |p|} D_{p[h], p[h+1]}.$$  

Where $p \in P_{ij}$ is the acyclic path between vertex $i(O_i)$ and vertex $j(O_j)$, $p[h]$ is the index of the $h$th vertex along path $p$. Finally, $|p|$ is the number of vertices along the specific path. Therefore, $D_{p[h], p[h+1]}$ is the weight of $h$th edge along path $p$. VAT and iVAT are algorithms that transform a dissimilarity matrix into a graph-theoretic distance matrix. Therefore, iVAT can be interpreted as a feature extraction technique.

2.5.4 Validation indices

To validate the (fuzzy) partitioning, cluster correlation validity (CCV) measures can be used (Zhang, Zhang, Sicotte, & Yang, 2009). Pearson correlation based validity measures are especially useful for large datasets. Furthermore, the cluster analysis can also be validated using the Xie-Beni index. This index measures the compactness of clusters and separation between clusters in one index. This index is given by:

$$S = \frac{\sum_{i=1}^{C} \sum_{j=1}^{N} \mu_{ij}^{n} \|v_i - v_j\|^2}{N \min_{i \neq j} \|v_i - v_j\|^2}.$$  

Where a value closer to one implies a better fit. In the numerator, the compactness of the fuzzy partition is indicated while the denominator is an indication of the strength of the separation between the clusters (W. Wang & Yunjie, 2007).
2.6 Case-based reasoning

Case-Based reasoning has some advantages over other AI methods. First, it has less computational cost so it can learn non-linear separable categories and continuous functions. Second, it does not rely on statistical assumptions. Finally, it does not require an explicit domain model (C.-H. Liu & Chen, 2012). CBR has been applied to many different fields and recently to predict energy consumption estimation (Monfet et al., 2014) (Koo & Hong, 2015).

Records in this technique are called cases and can be used to support comparable decisions in the future. This is referred to as case-based reasoning (CBR) and allows the computer to use solutions from old problems to solve new ones. Next to case-based reasoning is inductive learning. This technique implies that the computer uses historical cases to generate rules for solving new problems. These rules can then be used for automating decision support processes (Turban et al., 1997).

New problems are often similar to already solved cases, which allows to interpret the obtained solution within the context of the unsolved problem. Watson (1999) describes the CBR-cycle as follows:

1. Retrieve similar cases to the problem description.
2. Reuse a solution suggested by a similar case.
3. Revise or adopt that solution to better fit the new problem if necessary.
4. Retain the new solution once it has been confirmed or validated.

Note that this cycle is a methodology and does not describe techniques. The CBR-cycle is displayed in Figure 2.10.

![Figure 2.10: The CBR cycle, obtained from Watson (1999)](image)

The CBR must retrieve a case from a case-library and assess the similarity of the cases in the library compared to the problem that needs to be solved. Then, the CBR must reuse the
solution of the retrieved case, with or without revision. The last step is that the CBR should extend its knowledge by retaining new cases.

The most common technology in CBR is nearest neighbour. The similarity of the case to another case in the case-library is determined for each attribute. Sometimes, a weighting factor is added. The similarity is represented by the sum of the similarity of the attributes. The following equation represents the similarity [Watson, 1999]:

\begin{equation}
\text{Similarity}(T, S) = \sum_{i=1}^{n} f(T_i, S_i) \ast w_i
\end{equation}

In which \(T\) denotes the target case, \(S\) is the source case and \(n\) the number of attributes in each case. Furthermore, \(f\) is a similarity function and \(w\) the weight of attribute \(i\). The calculation is repeated for all the cases in the library to rank the cases by similarity. Most of the time, similarities are between 0 and 1 or as a percentage. This technique was used by (Juan, 2009) to find similarities in houses that need refurbishment and by (Monfet et al., 2014) to predict total energy demand.

Another CBR technology that is commonly used is induction. These algorithms built decision trees from case histories. The heuristic then partitions the cases into clusters, based on patterns that are identified. The algorithms can then cluster the cases, where it is assumed that cases with similar problem descriptions refer to similar problems with similar solutions. This technique was successful in predicting the due date assignment of jobs (C.-H. Liu & Chen, 2012).

Fuzzy logic can also be used as CBR technique. Fuzzy logic uses linguistic terms that describe differences in an attribute that describes a feature of the case. When fuzzy logic is used in CBR, a fuzzy preference function is used to define the similarity of a case with the attribute of the target case [Watson, 1999]. A fuzzy preference function results in a vector, which is described as the fuzzy preference vector, which contains a fuzzy preference value for each attribute. Through weighted aggregation, these values can be combined and result in a similarity value. Then, a fuzzy preference function can be used to compare an attribute with the qualitative description of other attributes. Fuzzy logic within CBR can be used for travel planning, help desk services and many other applications (Chaudhury, Tuhina, & Goswami, 2004).

Finally, database technology can be used in case based reasoning. Well-formed queries can retrieve cases with high similarity. A drawback of using database technology is that databases retrieve cases using exact matches of the queries. The use of wild cards can make a query more applicable, but cannot be used to measure similarity.

### 2.7 Neural Networks

Artificial neural networks are inspired on the functioning of the brain, where billions of neurons are interconnected, which implies that the neurons can communicate with their neighbours. This is done by small electric or chemical signals. Recently, there have been many improvements realized in neural networks, including new activation functions, adaptation and other learning functions (J. Yang, Rivard, & Zmeureanu, 2005). Therefore, neural networks
are able to produce good results for the prediction of building’s energy consumption and are a popular method (Hippert, Pedreira, & Souza, 2001), (Tso & Yau, 2007).

Neural networks consist of several processing elements, which are the neurons and can be organized and create the structure of the network. An example of a neuron is displayed in Figure 2.11.

![Image of an artificial neuron](image)

**Figure 2.11: An example of an artificial neuron**

A neuron receives information through its input nodes $X_i$ (Hippert et al., 2001), which is processed. Then, the neuron puts out a response. Mostly, this is done in two stages. In the first stage, the input weights $w_i$ are combined linearly and the result is used in an activation function. All the weights and the constant $\theta$ (bias term) are used. Note that the activation function must be nondecreasing and differentiable. The reason that the function must be differentiable is that for the backpropagation algorithm a steepest-descent technique is used. This technique is based on the computation of the gradient of the loss function with respect to the network parameters (Hippert et al., 2001).

Consider a simple model that predicts gas consumption of buildings on a 15 minute time interval. This model only uses Temperature ($X_1$) and Gross floor area ($X_2$) as inputs. As has been described in Section 2.7, a neural network consists of neurons (nodes) that process values. This simple model is displayed in Figure 2.12.

![Image of a perceptron](image)

**Figure 2.12: A perceptron**

This node is based on the feed-forward model, which implies that inputs are that are received by the node are processed and produce and output. For this example, a temperature of five degrees Celsius and a gross floor area of 100$m^2$ are inputs. As can be obtained from Figure
the node receives *weighted* inputs through weights $w_{ij}$ which are the weights from input $i$ to node $j$. When a node is initialized, random weights are assigned. Let’s say that $w_{11} = 0.01$ while $w_{21} = -0.05$ (gas consumption tends to increase with the area of a building while it decreases with higher temperatures). Now, the inputs are multiplied by the weights and summed. For this example, that would be $100 \times 0.01 + 5 \times -0.05 = 0.75$\(^3\). The sum of the weighted inputs is then passed through an *activation function*. Consider an activation function that describes output just as the sum of weighted inputs $w_{11} \cdot X_1 + w_{21} \cdot X_2$. In this case, output of the model would be $0.75\text{m}^3$ of gas consumption. However, if temperature would be 30 degrees Celsius, the model would predict $100 \times 0.01 + 30 \times -0.05 = -0.5 \text{m}^3$, which is impossible. To overcome this, a *constant bias* with weight $\theta$ can be added to the model. This constant bias has always the value of one. Adding the constant bias is displayed in Figure 2.13. If the weight $\theta$ is 0.6, output of the model would be $0.6 + 100 \times 0.01 + 30 \times -0.05 = 0.1 \text{m}^3$.

![Figure 2.13: A perceptron with constant bias](image)

Suppose that data for one building (GFA = 100 m\(^2\)) is available. The observations are taken on different dates resulting in different temperature inputs $X_2 = [5, 20, 30, 10, 0]$. Using the parameters from Figure 2.13 the model predicts outcomes $\hat{y}_i$. For these five measurements, outcome values are known as $y_t$. Now, the mean squared error between the predicted values and the observed values can be calculated. Table 2.3 gives the results for this model.

<table>
<thead>
<tr>
<th>$X_2$</th>
<th>$\hat{y}_t$</th>
<th>$y_t$</th>
<th>Squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.35</td>
<td>1.4</td>
<td>0.0025</td>
</tr>
<tr>
<td>20</td>
<td>0.6</td>
<td>0.8</td>
<td>0.04</td>
</tr>
<tr>
<td>30</td>
<td>0.1</td>
<td>0.4</td>
<td>0.09</td>
</tr>
<tr>
<td>10</td>
<td>1.1</td>
<td>1.2</td>
<td>0.01</td>
</tr>
<tr>
<td>0</td>
<td>1.6</td>
<td>1.6</td>
<td>0</td>
</tr>
</tbody>
</table>

A common structure for a neural network is a layered structure. An example of a layered neural network is shown in Figure 2.14.

\(^3\)For this example, it is assumed that temperature only takes positive values. In reality, this problem is overcome by normalizing the data.
In this structure, a collection of neurons is grouped into layers. The network is feed-forward because the outputs of one layer are used as inputs for the next layer. In some layered neural networks, a hidden layer is added. A hidden layer takes its input from the previous layer and processes it to the next layer. Sometimes, multiple hidden layers can be present, although the use of one hidden layer is the most common one. To determine the number of hidden layers and nodes in each layer, several approaches exist currently ([Stathakis, 2009]), being trial and error, heuristic search, exhaustive search and pruning and constructive algorithms.

Deep neural networks (DNN’s) are multilayer perceptrons with many hidden layers. In general, networks with more than three hidden layers are considered as deep. These NN’s are known to be hard to train, basically because adjusting the input weights often results that the solution gets stuck in a local minimum ([Larochelle, Bengio, Louradour, & Lamblin, 2009]). For complex problems, deep architectures can be more effective than shallow architectures although this is not necessarily the case for simple problems. Another condition for good performance of deep neural networks is that enough data is available. Recently, deep networks have emerged to solve complex problems such as speech recognition ([Siniscalchi, Yu, Deng, & Lee, 2013]) and visual pattern recognition ([Schmidhuber, 2015]). So far, deep learning has not been used to forecast future energy consumption in buildings or energy consumption forecasting on a macroeconomic scale.

Neural networks can transfer different types of information, both categorical and numerical. Within one layer, the neurons share the same information from the inputs but are not connected to each other. Inputs of the neural network are regarded as attributes. The connection weights $w_{ij}$ are the most important elements of ANN’s. If the network is trained, the connection weights adjust to produce more accurate outputs.

### 2.7.1 Learning

To train a neural network, a learning algorithm is required. These learning algorithms determine the process of how the neural network learns the relation between the inputs and outputs. The process can also aim at only discovering the relation between the inputs. In that case, the learning algorithm is unsupervised. In case of unsupervised learning, the network is self organizing, which implies that each hidden process element responds to a set of inputs. The amount of categories can be fixed by setting model parameters.
The most common technique for supervised learning is backpropagation (Hippert et al., 2001), which implies minimizing an error function. Errors are measured as the difference between actual output of the network and the output as specified in the training set (Turban et al., 1997). With backpropagation, the connection weights are adjusted starting at the output nodes and then propagated back through the layers of the network. The connection weights can be set at random or via predefined rules. Backpropagation basically implies the repeated application of a chain rule, computing the influence of each weight in the network (Riedmiller & Braun, 1993). This is done with respect to the error function, which can be described as:

\[ \frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial s_i} \frac{\partial s_i}{\partial \text{net}_i} \frac{\partial \text{net}_i}{\partial w_{ij}} \]

In which \( w_{ij} \) is the weight from neuron \( j \) to neuron \( i \), \( s_i \) is the output and \( \text{net}_i \) is the weighted sum of all the inputs of neuron \( i \). If all the partial derivatives are known, the error function is minimized by performing a gradient descent:

\[ w_{ij}(t+1) = w_{ij}(t) - \epsilon \frac{\partial E}{\partial w_{ij}}(t) \]

Where the learning rate \( \epsilon \) is important considering the time until convergence is obtained. If \( \epsilon \) is too small, many steps are needed, greatly increasing the learning time. If it is set too large, it will prevent the error from falling below a certain value (Riedmiller & Braun, 1993).

Let’s get back to the example from Section 2.7. In supervised learning, the model is optimized based on the outcomes of the data. For this example, the Mean squared error is 0.0285. Now, the next step called training occurs. This is done by updating the weights to reduce the error based on learning rate \( \epsilon \). The new weight can be described by:

\[ \text{New weight} = \text{weight} + \Delta \text{weight} \times \epsilon \]

Where \( \Delta \) is based on the error and the inputs. Suppose that only \( w_{21} \) is updated and that based on the error and on \( \epsilon w_{21} \) is equal to 0.045. Now, the model can again predict values for the given inputs. The results are given in Table 2.4.

<table>
<thead>
<tr>
<th>( X_2 )</th>
<th>( \hat{y}_t )</th>
<th>( y_t )</th>
<th>Squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.375</td>
<td>1.4</td>
<td>0.000625</td>
</tr>
<tr>
<td>20</td>
<td>0.7</td>
<td>0.8</td>
<td>0.01</td>
</tr>
<tr>
<td>30</td>
<td>0.25</td>
<td>0.4</td>
<td>0.0225</td>
</tr>
<tr>
<td>10</td>
<td>1.15</td>
<td>1.2</td>
<td>0.0025</td>
</tr>
<tr>
<td>0</td>
<td>1.6</td>
<td>1.6</td>
<td>0</td>
</tr>
</tbody>
</table>

With these outcomes, the new mean squared error (MSE) is equal to 0.007125. This is lower which implies a better model fit. As the MSE is still above 0, there are possibilities for improvement. However, too much training iterations can result in overfitting, where the model is too much adapted to the training data. Therefore, the model is tested on the validation set.

40
If the error of the model regarding the validation set increases, the training algorithm stops and obtains the final weights. Figure 2.15 illustrates this process.

![Figure 2.15: Relation between errors on the training and validation set](image)

To avoid getting stuck in a local minimum, the model always performs validation checks. When an increase in the error regarding the validation set is increasing, five more training iterations are performed. If all these five iterations further increase the error, the model assumes that the global minimum is found and the training process is frozen, obtaining the final model. If one of these iterations results in a lower error, the process continues until the global minimum is found.

Deep neural networks require different training techniques because the error-function often gets stuck in a local minimum. (Larochelle et al., 2009) proposed the following principles for training deep neural networks:

- Pre-training one layer at a time in a greedy way
- Using unsupervised learning at each layer in a way that preserves information from input and disentangles factors of variation
- Fine tuning the whole network with respect to the ultimate criterion of interest

### 2.8 Summary

In this chapter, relevant literature has been described. First, it has been shown that ZBS can solve problems that are present in energy investment projects nowadays. Risk aversion, lack of information transparency and inadequate calculations are the main problems. ZBS can help to overcome these problems and help to invest in profitable energy investment projects. Furthermore, previous research projects in energy consumption estimation have been addressed. From these studies, applicable techniques are obtained. The concept of machine learning is described, and three techniques (clustering, case-based reasoning and artificial neural networks) are presented in more detail.

Chapter 3 describes the methodology that is adopted. Furthermore, Chapter 4 gives a description of the dataset. In this chapter, it is shown why the described techniques are applicable to the dataset.
Chapter 3

Methodology

Adopting the right methodology is crucial for the success of a research project. As was shown in Section 1.6, CRISP-DM is the most used and preferred methodology within the data mining field. This chapter describes the current calculation module and the CRISP-DM methodology in more detail. First, an overview of the framework is given, followed by a breakdown into four levels.

3.1 The current calculation module

For calculating the expected savings of the proposed measures, Ploos currently uses an accredited mathematical module designed for assigning EPA-U labels to buildings. An EPA-U label is a measure for energy efficiency of houses and other buildings, where energy efficiency is measured in an energy-index. For assigning an energy label, an experienced analyst inspects the building and makes use of building specifications and documentation. The lower the energy index, the better the label that the building receives. Ploos is certified to assign energy labels to buildings, so it has access to the required software. If a customer requests Ploos to make an analysis on energy saving possibilities, a start is made with analyzing the current situation. This analysis is presented in a diagram which shows energy consumption for all different consumption posts. An example of this analysis is presented in Figure 3.1.

![Consumption shares](image)

Figure 3.1: Example of an energy consumption analysis

http://www.rijksoverheid.nl/onderwerpen/huurwoning/puntensysteem-huurwoning/puntensysteem-en-energielabel
Then, several energy saving measures are proposed to the customer. These measures are presenting including the implementation cost, their net present value, payback time, energy label\textsuperscript{2} cost savings per year and CO\textsubscript{2} reduction per year, measured in percentage. However, it turns out that some measures reduce gas or heat consumption while electricity consumption is increased. For example, heat pumps have this property. Of course, the implementation of specific measures is only suggested if total energy consumption (gas and electricity combined) is reduced.

The current calculation module is mostly based on estimations. Experience has shown that the difference between calculated energy consumption and consumption in practice is approximately 5 percent, based on an analysis made for one year. An example of a calculation that is made with the current module can be described as follows: light bulbs that use 50 Watts of power are replaced by LED lights that consume only 20 Watts. Then, the savings on energy consumption are 30 Watts per light that is replaced multiplied by the estimated number of hours the light is switched on each year. This calculation does not take into account that light bulbs produce mostly heat and not light. So, the 30 Watts savings on electricity will probably incur an increase in heating costs. Also, during hot days it is expected that cooling load is decreased because of the LED light. This example clearly describes the problem that is faced with the current calculation module. As a result, customers have not enough confidence in the calculations to convince them to use the ZBS service.

3.2 CRISP-DM

As was described in Section 1.6, CRISP-DM is an industry wide methodology for approaching data mining projects. This section describes the six main phases of the methodology and also the breakdown of the steps which is shown in Figure 3.2.

The Business Understanding step describes the managerial need for the data mining study. It aims at answering specific questions. Also, a project plan is made in this phase and an estimation of the budget is given.

The second step, Data Understanding, identifies the relevant data from the available databases. In order to understand the data, graphical tools such as histograms, box-plots and other statistical summaries can be presented.

In the Data Preparation (sometimes called data preprocessing) step, the selected data is prepared for analysis by data mining methods. In many projects, this is the most time consuming step. Possible reasons for this are that data is often lacking values, noisy or inconsistent. Data Preprocessing has four main steps; being Data Consolidation, Data Cleaning, Data Transformation and Data Reduction.

Step four is called Modeling Building. Here, multiple modeling techniques are used to address the business need as identified in the Business Understanding step. Also, a comparative analysis is made between the different models that are used.

In the fifth step, Testing and Evaluating, the models are assessed and evaluated for their accuracy and generality. In this phase, it is important to test if the selected model meets the business objectives and to what extent the models need to be developed and assessed. Also in this step, the data is visualized.

\textsuperscript{2}A hypothetical energy label is displayed in case that only that specific measure is implemented
The final step is *Deployment*. The deployment step ranges in complexity from presenting a report to implementing a repeatable data mining process across the enterprise. Maintenance of the models can also be a part of the Deployment step, as well as monitoring.

Besides the general process, the CRISP-DM process can be split up further. Figure 3.2 shows the breakdown of the CRISP-DM methodology.

![Four level breakdown of the CRISP-DM methodology](https://example.com/image)

Figure 3.2: Four level breakdown of the CRISP-DM methodology, obtained from [Smart Vision Europe 2013](https://example.com)

The top level is called *phases* and each phase has several second level tasks. The phases are the steps of the CRISP-DM process that was showed in Figure 1.6.

The second level tasks are described as *generic tasks*. This step is intended to cover all possible data mining situations. This level has two aims to be as complete as possible and to be as stable as possible. Complete implies that the whole data mining process is covered but also all applications. That the model should be valid for developments in the process is meant by *stable*.

The third level is called *specialized tasks*. It describes how actions in the *generic tasks* level should be performed under specific situations.

Finally, the lowest level describes the *process instances*, these record what actually happened in a particular engagement.

Some of the phases do not need to be specified completely. However, the *data preparation* phase is important and as described before often requires a great amount of time to complete. Therefore, this phase will be specified in Section 3.1.
3.2.1 Data Preparation

Figure 3.3: Breakdown of the data preparation step

Figure 3.3 shows the breakdown of the Data Preparation step. As can be obtained from the figure, four generic tasks are identified; which are data consolidation, data cleaning, data transformation and data reduction. It was described that these generic tasks are meant to be stable and complete. It is clear that real world data needs to be prepared before it can be used in a model. These four steps describe the complete set of operations that is needed to prepare the data.

The first specialized task in the generic task data consolidation is data collection. For this thesis, data will be collected from three different sources; Ploos Energieverlening has a personal database which stores all consumption data of their customers. Secondly, the database of the KNMI (Royal Dutch Meteorological Institute) contains data from weather measurements. Also, data from Ploos’ customers will be used. This data contains information about the customers’ organization and building. Relevant variables will be selected based on expert knowledge.
The second specialized task is the selection of the data. In this task, the focus will be on the time span of the selected data. More specifically, the time span around the implementation of the implemented energy saving measures has to be determined. For this specialized task, it is important that the date of implementation is known and then the right period to analyze has to be determined.

Finally, the integration of the selected data will be performed as a specialized task. Here, the data will be organized based on the buildings and time spans. More specific, the right consumption values have to be matched with the corresponding climate parameters for the building for each time interval.

The next generic task is data cleaning. This tasks has three specialized tasks; impute missing values, reduction of noise in the data and eliminate inconsistencies. These generic tasks will be described in more detail in Chapter 4.

Furthermore, data transformation is considered as the third generic task. It consists of three specialized tasks named normalize data, discretize/aggregate data and construct new attributes. These tasks will also be described into more detail in Chapter 4.

The final generic tasks is data reduction. Datasets are often very large containing millions or billions of records. It might be necessary to reduce the number of records or some of the variables but this heavily depends on the available dataset.

The specific steps that have been used for this thesis can be found in Chapter 4. The data preparation breakdown steps are described in more detail in Section 4.2.
Chapter 4

The dataset

This chapter describes the dataset. The dataset consists of commercial buildings and residential buildings which are split up in two datasets. First, the data is explored. This is covered for both the internal data and the external data in the first section. Then, the data is prepared for use, which is done in the second section of this chapter.

4.1 Data understanding

This section describes the data that is used to build the models. The first subsection presents an overview of the external data, which in this case is climate data obtained from the KNMI. The internal data comes mostly from the ploos database itself. Also, this dataset contains both office and apartment buildings.

4.1.1 External data

External data in this thesis consists of climate data and buildings’ occupation data. Climate data in this thesis is extracted from the KNMI\(^1\). The data is extracted from 37 different weather stations, spread across the Netherlands, which record one measurement per hour. Figure 4.1 shows a map with the locations of the weather stations indicated.

\(1\)Royal Dutch Meteorological Institute, available via http://www.knmi.nl/klimatologie/uurgegevens/selectie.cgi

Figure 4.1: Locations of the weather stations
Note that the numbers in Figure 4.1 represent the location numbers of all the weather stations. As can be obtained from Figure 4.1, the weather stations are spread relatively equally across the country. This makes it more easy to link a certain climate pattern to a specific building, giving it more accurate values. According to the KNMI website, the weather stations record 22 different weather parameters. However, some weather stations record only a few of them. Others have started at different dates with the recording of different values, resulting in an inconsistent dataset. As a result of that, it has been decided to use only eight parameters in the Ploos database. For the dataset of this thesis, only three variables are relevant for energy consumption prediction. The chosen parameters are displayed in Table 4.1.

<table>
<thead>
<tr>
<th>ID</th>
<th>Symbol</th>
<th>Name</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>69</td>
<td>T</td>
<td>Temperature</td>
<td>0.1 Degrees Celcius at 1.5m height</td>
</tr>
<tr>
<td>70</td>
<td>Q</td>
<td>Global radiation</td>
<td>Joule per cm²</td>
</tr>
<tr>
<td>72</td>
<td>U</td>
<td>Relative humidity</td>
<td>Percentage at 1.5m height</td>
</tr>
</tbody>
</table>

Other external data is corresponded directly to Ploos by customers. For their offices, the daily number of employees that is in the building is recorded. Together with the FTEs that are in the Ploos database for all buildings, the occupation rate of the building can be calculated, as will be shown in Subsection 4.1.2. A disadvantage is that not all employees are fulltime employed and might only spend a few hours per day in the building (for example cleaners). This results in an occupation rate that exceeds 1. An example for this is given in Figure 6.5 in which it is shown that the occupation rate can be above 1. Occupation rate is only available for buildings of one particular customer and only for the first five months of the year 2015.

4.1.2 Internal data

Within the internal dataset, extracted from the Ploos database, difference can be made between real-time data streams and other data.

---

2 Fulltime Equivalent
**Real-time data**

Many buildings have smart meters that send a consumption measurement each time interval. Three types of consumption meters are used: electricity, gas and heat meters. In this dataset, buildings have a gas or a heat connection, not both. Also, some buildings do not have a gas or a heat connection at all. This can be due to the fact that no gas or heat is supplied to the building, but also because the gas meter is not a smart meter. Table 4.2 shows the energy measurements.

Electricity meters send a measurement each 15 minutes. Also, some meters have two consumption types, being peak and low tariff measurements. Peak tariff is between 7 A.M and 11 P.M. on weekdays, while low tariff is between 11:00 P.M. and 7:00 A.M. each weekday and on weekend days, so no overlap between the tariffs exists. Therefore, total electricity consumption can be regarded as the sum of peak and low tariff consumption. Electricity consumption is measured in kiloWatt-hour (kWh).

Regarding gas and heat meters, consumption measurements are taken each hour. Compared to electricity meters, gas and heat meters do not have a peak and low tariff. Additionally, gas meters measure consumption in cubic meters (m³) while heat meters measure consumption in Gigajoule (GJ). For electricity, gas and heat meters, different types of meters have been found in the dataset. Some meters have an accuracy of two or three decimal numbers and some have an accuracy of just whole numbers, this has been found for electricity and gas meters.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Unit</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electricity</td>
<td>kWh</td>
<td>Each 15 minutes</td>
</tr>
<tr>
<td>Gas</td>
<td>m³</td>
<td>Each hour</td>
</tr>
<tr>
<td>Heat</td>
<td>GJ</td>
<td>Each hour</td>
</tr>
</tbody>
</table>

**Non real-time data**

Non real-time data can be described as information that does not change frequently. Examples are the energy saving measures (which are at least constant for the payback time of the investment) and the gross floor area of the building (which is assumed to be constant for as long as the building remains in use). An overview of the non real-time data is provided in Table 4.3.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy saving measures</td>
<td>Classes</td>
</tr>
<tr>
<td>Gross floor area</td>
<td>m²</td>
</tr>
<tr>
<td>Energy label</td>
<td>Classes</td>
</tr>
<tr>
<td>Fulltime equivalent</td>
<td>Natural numbers</td>
</tr>
<tr>
<td>Customer type</td>
<td>Classes</td>
</tr>
</tbody>
</table>

This data consists of the energy saving measures as described in Subsection 4.2.1, the type of customer (in this dataset only office, datacenter and residence are present) and the energy label.
4.2 Data preparation

This section describes the data preparation step of the CRISP-DM framework.

4.2.1 Data consolidation

As was described in Section 4.1, data from internal and external sources is used. The external data is obtained from the KNMI website. Here, a query can be executed in which the user can select the weather stations, the variables and the time span of the data. Then, a text file containing the data can be downloaded. For this thesis, it was requested to integrate the consumption and weather data. As a result of that, weather conditions are connected to consumption values for each time span. The other external data, being the occupation rate of one particular customer’s buildings is only supplied on request. For this thesis, occupation rates are available for the year 2015. Internal data consists of energy consumption data and building characteristics. This information can be downloaded using the Ploos online database. The extraction of the categorical data as described in Subsection 4.1.2 has also been integrated into the Ploos database.

Based on expert knowledge and the technical report by SenterNovem from 2007 only three climate variables are considered as useful for energy consumption prediction, being temperature, global radiation and relative humidity.

Consumption data has been extracted for all buildings where energy saving measures have been implemented. Since communication from customers regarding implementation of these measures is often insufficient it has been decided to use consumption data from 01-01-2014 to 31-07-2015. These dates have been chosen because for most of the measures, it is assured that on 01-01-2014 the measures were implemented and active.

4.2.2 Data cleaning

This subsection describes the process of data cleaning, which includes missing value analysis, outlier detection and elimination of inconsistencies.

Missing value analysis

First, a missing value analysis is performed. It turned out that for some buildings, all gas measurements are missing. Since there is not a single value known for these measurements, there is no possibility for imputation. It was also obtained that for some specific time slots, consumption data is missing. This can be a whole day of measurements but also values for just an hour. It was found that when time switches to summer time schedule (time switches from 2:00 A.M. to 3:00 A.M.) four time slots are lacking values.

As was described in Subsection 4.1.1 some of the weather stations lack measurements for specific parameters or time spans. In case of missing values, it was decided to replace the missing values by values of weather station De Bilt for that particular time. One could suggest to use measurements of the nearest station in case of missing values. However, it turned out that in many cases that weather station is missing particular measurements as well. Also De Bilt is the main weather station of the KNMI and is located on a central position in the Netherlands.
For this dataset, no missing values in climate data were present so all values in the dataset are taken from the closest weather station.

**Outlier detection**

An outlier is considered as an observation that is inconsistent with the majority of observations in a data set (Seem, 2007). Several methods exist for detecting outliers in energy data, including an Energy Consumption Index (ECI) (Dodier & Kreider, 1999), a Generalized Extreme Studentized Deviate (GESD) procedure and the modified $z$-score (Seem, 2007). These methods are mostly for large datasets containing many different buildings. For this dataset however, it is possible to have a closer look at the data. First, an outlier filter using the well known $\mu - 3\sigma$ and $\mu + 3\sigma$ has been applied. Several types of outliers have been identified using this approach. More about the types of outliers and examples can be found in Appendix A.3. In general, most values that appear to be outliers could be realistic in practice. Therefore, for commercial buildings only one type of outlier has been removed from the dataset. More specifically, cumulative measurements have been corrected. Problems can best be solved at the source. Therefore, these values have been corrected in the Ploos database. Considering the residential buildings, connections of $3 \times 25$ amper are present. It can be calculated using the voltage of 230 in the Netherlands that each of these connections can use at most 4.3125 kWh per 15 minutes. Therefore, values above this number are removed from the dataset. For each climate variable the minimum and the maximum in the dataset have been calculated. No odd values have been found so no corrections have been made.

**Elimination of inconsistencies**

As has been described in Section 4.1, electricity consumption data is measured over a time span of 15 minutes while gas consumption is measured over the time span of an hour. Therefore, gas consumption per hour has been devided by four as the consumption for each interval of 15 minutes within the hour. It could be argumented to summarize the electricity consumption per hour as the sum of the four 15-minute intervals. However, it has specifically been chosen not to do this because the valuable detailed electricity consumption data would be lost. For the same reason, climate data has been transformed in four values per hour as well. Note that climate data is measured each round hour (03:00 P.M., 04:00 P.M. etc.) and consumption data is measured per time interval. To overcome this inconsistency, climate data has been transformed into time intervals as well where the value of the closest hour is taken as value for that interval. For example, for the time interval from 03:15 P.M. to 03:30 P.M. the climate parameter values measured on 03:00 P.M. are taken. The same holds for the time interval from 03:45 P.M. to 4:00 P.M. Based on expert knowledge this decision is taken. It is expected that the deviation of time does not influence the model a lot since buildings tend to store heat in their walls, roofs and other surfaces. Since all consumption intervals are measured on intervals containing quarters of an hour and all climate parameter values are measured on whole hours this completely removes all inconsistencies.

Another inconsistency that is present is the measurement on the meters. Consumption meters for small volumes have measurements with two or three digits while the meters for large volumes have round numbers. No changes have been made on these measurements. A reason for this is that some errors as described in Subsection 2.4.1 are measured in percentages. If
small meters would be rounded to the nearest number, errors would increase too much. Instead of that, consumption volumes for large volume meters are treated just as numbers with digits of zero.

### 4.2.3 Data transformation

This subsection describes the next specialized task of the data preparation phase. The data is normalized. Also, aggregation and discretization steps have been performed, followed by the construction of new attributes.

#### Data normalization

Data normalization has been performed in Matlab using `mapminmax` command. It normalizes all values to numbers between zero and one.

#### Aggregation and discretization of the data

Aggregation has been performed using the building’s postal code and the locations of the weather stations. Each building has a postal code, that can be linked to the nearest weather station. Using these, the climate data can be connected to the building. This combined with the energy consumption data and the building characteristics completes the dataset.

#### Attribute construction

The Ploos database used to classify measures obtained from EPA-U certificates in the following classes: Not defined, lighting, construction, comfort, devices, other. A disadvantage of this classification is that it does not make distinction between the many possible measures that appear in the system. An example of the classes in the database is given in Figure 4.3.

![](old_classification.png)

**Figure 4.3: Old classification of measures**

As a result of the poor classification, effort has been spent to make improvements in the classes. To obtain this improvement, an interview has been conducted. Participants in this interview are the experts Paul Vermeulen, Johan Aarts and IT supervisor Martijn Ploos van Amstel, all of them are Ploos employees. The transcription of the interview is given in Appendix [A.1](#). The interview was conducted in Dutch.
The interview has resulted in a more detailed classification. The classes lighting, comfort and construction have been split up in multiple options. In general, it can be said that some classes have been created because the descriptions of the energy saving measures were not detailed enough. Also, the effect of the measures on energy consumption have been taken into account. The new classification tree is displayed in Figure 4.4.

As can be seen in Figure 4.4, the description of the energy saving measures is much more detailed. The new classification has also been implemented in the Ploos database which makes it more easy to extract data based on queries. For this dataset, the energy saving measures are added as binary variables where each building is assigned a value of 1 if the measure has been implemented and a value of 0 if not. The amount that each measure is present in the dataset is displayed in Table A.2 in Appendix A.3.2.

A similar approach has been applied to the energy labels that have been attached to the buildings. Energy labels range from A (energy efficient) to G (energy inefficient). The results are summarized in Table A.3 in Appendix A.3.2.

The types of buildings have been added as binary variable as well. The results are given in
Table A.4 in Appendix A.3.2

Considering the output of the model, an attribute is constructed using the energy values from electricity and natural gas. One kWh of electricity contains $3.6 \times 10^6$ Joule while one m$^3$ of natural gas contains $31.7 \times 10^6$ Joule (SenterNovem 2007). The sum of these values then represents total energy consumption of a building.

As has been described in Section 2.5, clustering algorithms can be applied to the dataset for improvement. A fuzzy cluster model has been used to create load-climate profiles of buildings. More about this can be read in Section 5.1.

Finally, while building the model the need for control on date and time has emerged. As a result of that dummies have been created that represent specific time slots during the day or during the year. The process of creating these time dummies is described in Chapter 5.

### 4.2.4 Data reduction

Data reduction consists of three specialized tasks. However, for this thesis, only a limited amount of variables has been selected, based on expert knowledge. This removes the need to reduce the number of variables. Also, a select number of cases has been added to the dataset so no cases have to be deleted. Finally, since only a limited amount of buildings is added to the dataset and for all buildings the same time span of measurements has been selected, no balancing operations have been performed.

The only exception is when consumption data is missing. As will be shown in Chapter 5, the dataset is large which resulted in long computation times. Therefore the records that lack consumption values have been removed from the dataset rather than imputing values.
Chapter 5

Model building

This chapter describes the models that have been used for this thesis including their properties. For the models, the Matlab software package is used. Then, the clustering model is described, followed by the neural network. During the model building phase, it was decided to concentrate completely on the neural network and not to finish the CBR model. The most important reason for this is that it was found that adding more data to the model can improve the results. Therefore, effort was spent on gathering additional data and on implementing the model in the company. The neural network is tested for two datasets, being commercial buildings and apartment buildings.

5.1 Software package

For building the neural network, Matlab 2015b has been used including the Neural Network and Fuzzy Logic toolboxes. The Neural Network toolbox has many predefined functions that make it easier to create a model as in Section 5.3. The Fuzzy Logic toolbox is needed to execute certain functions regarding the clustering model of Section 5.2. The advantage of Matlab over other software packages is that TU/e students and employees have experience using it which speeds up the process of model building. Matlab also offers the possibility to fully automate the execution of the model. Finally, Ploos has an executable version of Matlab which enables them to use the model for energy consumption prediction even after this research project is finished.

5.2 Clustering

For improving the dataset that is used in the Neural Network and the CBR model, clustering data can be a useful technique. Therefore, a clustering model has been created using averages and summations of office hours (7 A.M. to 7 P.M. on weekdays) and non office hours (7 P.M. to 7 A.M. on weekdays and both time slots on weekend days). Variables included in this analysis are electricity and gas consumption and the three climate variables described in section 4.1.

For calculating the distances between objects, the summations of electricity and gas consumption have been created, as well as for global radiation. For temperature and relative humidity, the averages are taken. For all these calculations, a three dimensional matrix has been created resulting in a row in the third dimension containing exactly 48 measurements. This matrix is then reduced back to two dimensions giving the sum or average for each 12 hour time slot.

The iVAT script that has been used for this thesis will display a figure. This can then be interpreted by the user to determine the number of clusters for the Neural Network and the CBR model. It was first tried to obtain the number of clusters using normalized data. However, it turned out that normalization made it more difficult to determine the number of clusters.
Figure 5.1 shows the results for the iVAT analysis using normalized and non-normalized data. Note that the pattern from the normalized data is still present in the non-normalized data.

Figure 5.1: iVAT images with normalization (left) and without normalization (right)

As can be seen from Figure 5.1, the suggested number of clusters is two or three. So now the fuzzy clustering procedure can be started using the pre-defined number of two and three clusters. The resulting membership grades can then be used as inputs for the Neural Network and the CBR model.

To validate the number of clusters obtained from the iVAT procedure, a Pearson’s correlation cluster validity index (CCV) and a Spearman’s CCV are performed. The results are shown in Section 6.1. These analyses are performed ranging from two to nine clusters. Finally, a Xie-Beni index is calculated for validation purposes using the results of the two and three cluster analyses as input.

5.3 Neural Network

For predicting energy consumption of buildings, a neural network as described in Section 2.7 is created. This section describes the properties of the neural network.

In this thesis, a feedforward layered neural network with backpropagation is used. The Matlab Neural Network Toolbox has different predefined functions that can be used to execute a neural network such as patternnet in case of a [0,1] classification problem. When regression is used, the Matlab function fitnet from the Neural Network toolbox creates the model. Since the process has to be fully automated and should be able to apply on any dataset that is loaded into the model, the model is build such that it can handle different lengths of inputs, as well as different input variables. Each NN that has been tested has one output; energy consumption (Joule), electricity consumption (kWh) or gas consumption (m$^3$) on a 15 minute interval. Two datasamples have been used as inputs, being exclusively residential buildings and commercial buildings.

For selecting the number of nodes in the hidden layers, the heuristic Number of nodes = (Inputs + Outputs) / 2 is used as a baseline. To derive the optimal result of the network, tests
are performed for the number of nodes ranging from the outcome of the heuristic -4 to +4 in each hidden layer. The network always tests for the number of nodes since the optimum might differ among datasets.

In scientific literature, most neural networks use one, two or three hidden layers for energy consumption prediction. Since computation time increases by adding more hidden layers, the network is tested only for one and two hidden layers. As with the number of nodes, in the optimal solution, the one or two layered model is selected. Note that the optimal solution is based on the mean squared error (MSE). The model is trained for each combination of nodes and hidden layers making it nine training sessions for the one layered model and 81 for the two layered model. The models have been tested using 5-fold cross validation to obtain the optimal number of nodes and 10-fold cross validation on the final model to obtain the performance indicators. For each iteration of $K$, the validation set is selected randomly from the data, being equal to 20 percent. Then, a list is displayed which shows the actual consumption values as well as the forecasted values. From this list, the performance indicators as described in subsection 2.4.1 are calculated.

The model has also been tested on random datasamples of different sizes using different numbers of $K$ on the final model to show that the results hold under different circumstances. For these training sessions, the data is randomly split into groups where the training set is 70 percent of the data and both the validation and testset consist of 15 percent of the data.

Different possibilities exist for optimizing a neural network. The Neural Network Toolbox in Matlab offers functions for determining these properties. The bias weights can be set to be updated or to remain constant. Both approaches have been tried. The results show that updating the biases results in significant better predictions than constant weights. Furthermore, different transfer functions can be used. For a fitnet, a linear combination that sets all negative values to zero (poslin) has been tried in the output layer to prevent negative predictions for gas consumption. However, this worsened the results. Therefore, the output layer uses a linear combination (purlin) where each value is just the processes value the layer receives. For the hidden layers, the matlab function tansig gives the best result.

### 5.3.1 Electricity

A Neural Network was as described in Section 5.3 was created and tested for two datasets, being commercial or apartment buildings. Each of the subsections describes the process of creating the model and optimizing the dataset to obtain the best possible results for the two datasets.

#### Commercial buildings

The commercial building dataset consists of 11 buildings. Figure 5.2 shows one of these buildings.
The Neural network has first been tested for electricity alone, resulting in acceptable results. Therefore, it was decided to first put effort in estimating gas consumption. In a later stadium, the results found for gas consumption prediction have also been used for the electricity model.

**Apartment buildings**

The dataset of apartments consists of five apartment blocks. Block 1, 3 and 5 house nine private rented apartments while block 2 and 4 have three private rented apartments and eight apartments that are rented by a foundation that takes care of disabled people. This foundation rents two floors of block 2 and 4. These two floors share one consumption meter. For apartment buildings, the basic model has been applied (without time dummies). Results show that without control for residents’ behavior, electricity consumption is hard to predict.

Each of the five apartment blocks is equipped with solar panels on the roof. A satellite image of the buildings is displayed in Figure 5.3.
These panels supply energy that is used for public rooms in the building (elevator, lights in the hallway etc.). Figure A.11 in Appendix A.5 shows the structure of the consumption meters. When the energy supply of the solar panels (meter CM3 in Figure A.11) exceeds the consumption of public rooms, the energy is supplied to the network. These earnings are divided over the apartments of each block. In other words, each apartment receives an equal share of the solar panel earnings, based on meter CM2 in Figure A.11. It was therefore decided to include meter CM2 into the dataset.

Using a semi-automated method (the information that is supplied by the Ploos database does not provide information about the block each apartment belongs to) the consumption values of meter CM2 are subtracted from the consumption meters of the apartments, taking the share of the earnings into consideration. Then, the neural network was run using the new consumption values. Block 2 lacks consumption values for meter CM2 and was excluded from the dataset.

5.3.2 Gas

In general, it can be said that gas consumption on a 15 minute time interval is less predictable than electricity consumption. The main reasons for that are that buildings tend to keep heat in their walls, roofs and other properties. Also, most buildings have central heating that is turned on when temperature drops below a certain number and is turned off as soon as temperature has reached the desired temperature. As a result of that, a drop in outside temperature does not directly lead to an increase in gas consumption for keeping the temperature inside at a constant level. Finally, many values of 0 gas consumption have been discovered in the dataset that cannot be a result of high temperatures outside (see Appendix A.3.1).

First, the basic model has been applied to a random sample of the dataset containing 100000 records. As was expected, the results for gas consumption prediction are less accurate compared to electricity. Figure 5.4 shows the results for the basic gas model.

![Figure 5.4: Regression results for the basic gas model](image-url)
As can be obtained from Figure 5.4, quite some mispredictions are made. Especially the pattern at the bottom of the cloud drew attention. Those dots represent datapoints that have been predicted too low compared to the reality. Commercial buildings tend to have high gas consumption at the beginning of the workday to heat the building to the desired temperature. To capture this, time dummies for each time interval (96 in total) have been added to the dataset and were tested on a sample of 10000 but worsen the performance significantly. Also, computation time increased because of the extra variables included. Based on expert knowledge, dummies for office hours, the start of the workday (office hours between 7 A.M. and 10 A.M.), heating season and cooling season were added. Adding a dummy variable to a neural network can be regarded as putting an extra input to the model that takes the value 0 if not present and the value of 1 if present. Figure 5.5 shows a simple model that predicts gas consumption using only three inputs, Tempreature, floor area and a location.

![Figure 5.5: Model without time dummies](image)

Suppose that results for this model are not very satisfying. Figure 6.6 shows a typical pattern for gas consumption during the heating season (meter 8716880000000**** on 01-10-2014).

![Figure 5.6: Typical pattern of gas consumption during the heating season](image)

The extra gas consumption that is used to heat the building during the start of a workday is clearly visible, as well as the relative stable pattern during the rest of the office hours and
non-office hours. Therefore, a binary variable that takes the value 1 if the time of the day is between 7:00 AM and 10:00 AM is added. Figure 5.7 shows what the model looks like if time dummies are added.

![Figure 5.7: Model with time dummies](image)

Note that the heating season is the time of the year between October and March while cooling season is between June and August. All these dummies have been tested separately and gradually increased performance of the model. Then, the final model was also tested on a dataset containing 10000 samples. Figure 5.8 shows the regression results for this model.

![Figure 5.8: Regression results for the gas model including dummies](image)

Performance increased but still too many samples containing zero actual consumption appear in Figure 5.8. Another approach was to change the transfer function of the output layer into the poslin type which is a function that returns 0 if the function receives a negative value.
This could increase the model fit since gas cannot be supplied to the network, ruling out the possibility of negative values. Unfortunately this resulted in a decrease of the model fit. As a final approach, the suggestion was made that using daily summations of gas consumption can increase accuracy. This has been implemented using time slots of 12 hours (each containing 48 records). Accuracy measures increased including a CV-RMSE of about 20 percent (more about this in Chapter 6). This removed most of the unpredictable zero values from the results although a few still remain. For the daily summations, the results from the cluster analysis as described in Section 5.1 have also been added to the dataset. Based on the available data all possibilities for further improvements have been exhausted which implies that the final model was obtained.

### 5.3.3 Energy

Another option to calculate energy consumption is to summarize gas and electricity into energy. As was described in Subsection 4.2.3, one m$^3$ of gas contains 31.7 MegaJoule of energy while one KWh of electricity contains 3.6 MegaJoule. The sum of electricity and gas consumption for each time interval is taken as output for this model. It has been tried both for energy prediction on a 12 hours base. As with the model for gas consumption, results from the cluster analysis have been added to the dataset.
Chapter 6

Evaluation

This chapter gives the results of the models. First, the results for the fuzzy clustering model are given including validation measures, followed by the results of the neural network. The results are summarized in tables. Also, a sensitivity analysis for the neural network is presented.

6.1 Clustering

To validate the cluster analysis from Section 5.2, a Pearson’s CCV and a Spearman’s CCV are performed. Table 6.1 gives the correlation coefficients for different numbers of clusters.

Table 6.1: Results for the Pearson’s and Spearman’s CCV using non-normalized and normalized data

<table>
<thead>
<tr>
<th>Number of clusters</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson non-normalized</td>
<td>0.9543</td>
<td>0.8171</td>
<td>0.7066</td>
<td>0.7221</td>
<td>0.658</td>
<td>0.5971</td>
<td>0.6014</td>
<td>0.5409</td>
</tr>
<tr>
<td>Pearson normalized</td>
<td>0.7001</td>
<td>0.8183</td>
<td>0.784</td>
<td>0.7751</td>
<td>0.7543</td>
<td>0.7324</td>
<td>0.7607</td>
<td>0.7307</td>
</tr>
<tr>
<td>Spearman non-normalized</td>
<td>0.3035</td>
<td>0.3092</td>
<td>0.317</td>
<td>0.3172</td>
<td>0.3172</td>
<td>0.3199</td>
<td>0.3148</td>
<td>0.3208</td>
</tr>
<tr>
<td>Spearman normalized</td>
<td>0.6129</td>
<td>0.6182</td>
<td>0.6461</td>
<td>0.6458</td>
<td>0.6461</td>
<td>0.6547</td>
<td>0.6543</td>
<td>0.6597</td>
</tr>
</tbody>
</table>

The results from the Pearson’s CCV are consistent with the results from the iVAT and show that the highest correlations are found for analyses using two clusters for non-normalized data and three clusters for normalized data. For the Spearman’s CCV, the results are less obvious. It shows low correlations for any number of cluster for the non-normalized dataset. For normalized data, it shows medium high correlations for all numbers of clusters that have been tested. Based on the Pearson’s CCV and the iVAT analysis, the number of clusters included in the model is two and three.

To validate the analysis itself, the Xie-Bani index for the two and three cluster models are calculated. Table 6.2 displays the results. It shows that the two-cluster model has a higher validity than the three-cluster model, which is consistent with the other validity measures.

Table 6.2: Xie-Beni index values

<table>
<thead>
<tr>
<th>Number of clusters</th>
<th>Xie-Beni index</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.3174</td>
</tr>
<tr>
<td>3</td>
<td>0.1344</td>
</tr>
</tbody>
</table>

Based on this analysis, the appropriate number of clusters is two. Since both Xie-Beni scores are below one (Xie-Beni can take values above one), the outcomes suggest that the clus-
ter analysis scores better on compactness relative to separation. Different fuzzyfiers \( m \) have been tried ranging from 1.25 to 3 but the results for the Xie-Beni index remained within the range 0.07-0.19 for the three cluster model. Therefore it was decided to keep the default value \( m = 2 \). Another reason to keep the default value is that tests with different fuzzy partitions \( U \) in the neural network are time consuming while model performance improvements are only minor.

6.2 Neural Network

This section describes the results for the neural network. First, the results for the commercial dataset are given. This subsection describes electricity, gas and energy. Then, the results for the residential buildings are presented. Note that the mean absolute percentage error can not be calculated because of outcomes that take the value 0 appear in the dataset.

6.2.1 Commercial buildings

In this subsection, results for electricity, gas and energy models are described. The models have been tested on different datasamples. The results are summarized in tables. First, it can be noted that testing for different configurations of layers and nodes does pay off. For example, in the basic electricity model the difference between the lowest and the highest mean squared error is 35.2 percent. By far, most optimal configurations are two-layered models although the difference between the optimum in one and two layered models is smaller.

Electricity

As was written in section 5.2, electricity consumption has been predicted using different inputs. Table 6.3 presents the results. Note that all specifications are optimized in the two-layered model. All specifications have been tested using 5-fold cross validation to obtain the best model and 10-fold cross validation to obtain the performance indicators. For these models, a random sample of 10000 observations has been used. The sample consists of the same observations for all the model specifications.

Table 6.3: Neural network: results for electricity prediction

<table>
<thead>
<tr>
<th>Model</th>
<th>Nodes 1</th>
<th>Nodes 2</th>
<th>RMSE</th>
<th>CV-RMSE</th>
<th>SMAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>No binary</td>
<td>27</td>
<td>28</td>
<td>44.4548</td>
<td>0.3062</td>
<td>0.4428</td>
</tr>
<tr>
<td>Binary workday</td>
<td>22</td>
<td>24</td>
<td>28.5629</td>
<td>0.1958</td>
<td>0.2871</td>
</tr>
<tr>
<td>Binary seasons</td>
<td>23</td>
<td>30</td>
<td>43.8576</td>
<td>0.3023</td>
<td>0.4407</td>
</tr>
<tr>
<td>Binary workday and seasons</td>
<td>27</td>
<td>28</td>
<td>29.262</td>
<td>0.2009</td>
<td>0.3181</td>
</tr>
<tr>
<td>Binary workday, seasons and startwork</td>
<td>25</td>
<td>32</td>
<td>29.0336</td>
<td>0.2004</td>
<td>0.3505</td>
</tr>
</tbody>
</table>

Table 6.4 shows the model specifications, the number of nodes obtained in the two layers of the model and the performance indicators. This table shows that the best model uses workday binary variables only. Adding seasons to the dataset increase accuracy only slightly. A Binary for the start of the workday does not increase accuracy.
To check robustness of the models on different datasets, the models have been run again using different datasamples. For each model, a random sample is selected. To speed up the results, K-fold cross validation has only be used to obtain the performance indicators in the final models. The results are presented in Table 6.3.

Table 6.4: Neural network: results for electricity prediction using different samples

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>Nodes 1</th>
<th>Nodes 2</th>
<th>RMSE</th>
<th>CV-RMSE</th>
<th>SMAPE</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>No binary</td>
<td>10000</td>
<td>29</td>
<td>22</td>
<td>43.5564</td>
<td>0.2982</td>
<td>0.4523</td>
<td>10</td>
</tr>
<tr>
<td>Binary workday</td>
<td>10000</td>
<td>24</td>
<td>29</td>
<td>26.2784</td>
<td>0.1862</td>
<td>0.286</td>
<td>10</td>
</tr>
<tr>
<td>Binary seasons</td>
<td>10000</td>
<td>25</td>
<td>26</td>
<td>24.5235</td>
<td>0.1702</td>
<td>0.3047</td>
<td>10</td>
</tr>
<tr>
<td>Binary workday, seasons</td>
<td>10000</td>
<td>31</td>
<td>24</td>
<td>28.8272</td>
<td>0.2</td>
<td>0.3237</td>
<td>10</td>
</tr>
<tr>
<td>Binary workday, seasons and startwork</td>
<td>100000</td>
<td>24</td>
<td>32</td>
<td>26.399</td>
<td>0.1818</td>
<td>0.2725</td>
<td>4</td>
</tr>
</tbody>
</table>

First, the model specifications are given. The second column displays the sample size. Then, the number of nodes in the optimal solution in layer one and two are presented, followed by the performance indicators. Finally, the number of cross-validations is stated. Table 6.3 shows that each of the binary variables added to the model increase accuracy compared to the basic model. All of these models have a CV-RMSE below 25 percent which implies a good model fit. It can also be obtained that the SMAPE is decreasing as more variables are added. In the final model, this percentage error is 27.25. From the results of the final model can be concluded that electricity consumption can be predicted with high accuracy, even for random samples. This is an important conclusion because it shows that computation times can be reduced without huge losses in accuracy.

Finally, occupation rate was available for three buildings in the first five months of the year 2015. Because occupation rate is measured on a daily base, only observations between 7:00 A.M. and 7:00 P.M. are included. This holds for workdays and weekendays obtaining a dataset with 21600 records. Table 6.5 shows the results. Since the occupation rate can be considered as a replacement of daytime variables, none of them are included in this model (with the exception of the heating season).

Table 6.5: Neural network: results for electricity prediction using occupation rate

<table>
<thead>
<tr>
<th>Model</th>
<th>Nodes 1</th>
<th>Nodes 2</th>
<th>RMSE</th>
<th>CV-RMSE</th>
<th>SMAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>23</td>
<td>23</td>
<td>59.4153</td>
<td>0.1792</td>
<td>0.2543</td>
</tr>
<tr>
<td>Occupation rate</td>
<td>21</td>
<td>22</td>
<td>39.1426</td>
<td>0.118</td>
<td>0.0823</td>
</tr>
</tbody>
</table>

Table 6.5 shows that adding occupation rate to the dataset is by far the best specification. Compared to the basic model, all performance indicators increase resulting in only 8.23 percent errors measured on a symmetric base. From this it can be concluded that even when occupation is measured on a daily base it is the best predictor for electricity consumption. In the absence of this information, time variables can be used to improve the model.
Gas

Table 6.6 shows the results for gas consumption prediction on a 15 minute interval. For this model, the same specifications as for the electricity model have been used.

Table 6.6: Neural network: results for gas prediction

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>Nodes 1</th>
<th>Nodes 2</th>
<th>RMSE</th>
<th>CV-RMSE</th>
<th>SMAPE</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>No binary</td>
<td>100000</td>
<td>21</td>
<td>30</td>
<td>5.8135</td>
<td>0.5986</td>
<td>0.6618</td>
<td>10</td>
</tr>
<tr>
<td>Binary workday</td>
<td>10000</td>
<td>23</td>
<td>25</td>
<td>4.0644</td>
<td>0.4376</td>
<td>0.6349</td>
<td>10</td>
</tr>
<tr>
<td>Binary seasons</td>
<td>20000</td>
<td>26</td>
<td>27</td>
<td>5.3406</td>
<td>0.5735</td>
<td>0.7074</td>
<td>10</td>
</tr>
<tr>
<td>Binary workday and seasons</td>
<td>10000</td>
<td>28</td>
<td>29</td>
<td>4.0425</td>
<td>0.4327</td>
<td>0.6644</td>
<td>10</td>
</tr>
<tr>
<td>Binary workday, seasons and startwork</td>
<td>10000</td>
<td>29</td>
<td>30</td>
<td>4.2679</td>
<td>0.4289</td>
<td>0.6667</td>
<td>10</td>
</tr>
</tbody>
</table>

Compared to electricity consumption prediction, results for gas show higher performance errors. The main reason for this is that many values of 0 occur in the dataset. Although, it can be obtained that adding a workday variable in the dataset did increase the results. This also holds for adding variables for the heating season and the cooling season although the results increase only slightly.

Table 6.7 gives the results for the model where summations of gas consumption have been used. This dataset uses the variables for workdays and seasons obtained from the gas dataset. The distance used for obtaining the cluster centers is Euclidean.

Table 6.7: Neural network: results for gas prediction based on daily summations

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>Nodes 1</th>
<th>Nodes 2</th>
<th>RMSE</th>
<th>CV-RMSE</th>
<th>SMAPE</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>No clusters</td>
<td>6431</td>
<td>23</td>
<td>24</td>
<td>135.1287</td>
<td>0.2991</td>
<td>0.488</td>
<td>10</td>
</tr>
<tr>
<td>2 clusters</td>
<td>6431</td>
<td>25</td>
<td>26</td>
<td>124.2049</td>
<td>0.274</td>
<td>0.4441</td>
<td>10</td>
</tr>
<tr>
<td>3 clusters</td>
<td>6431</td>
<td>32</td>
<td>26</td>
<td>132.9099</td>
<td>0.2929</td>
<td>0.4703</td>
<td>10</td>
</tr>
<tr>
<td>2 clusters, zeros removed</td>
<td>6387</td>
<td>30</td>
<td></td>
<td>110.833</td>
<td>0.243</td>
<td>0.4366</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 6.7 displays that gas consumption on a whole day is more predictable than on a 15 minute interval. This is mainly because of the ‘thermostatic’ nature of gas consumption. Adding clusters to the dataset did improve the model. Especially the two-cluster model is significantly better compared to the basic model. For the best model, the effect of zero-consumption measurements was tested. Therefore the zeros were removed from the dataset. The results indicate that indeed the performance can be improved by removing the zeros. Still gas prediction shows higher errors compared to electricity. It can be concluded that the available data is insufficient to estimate gas consumption even on a daily base.

To test the effect of distance measures on the outcome of the model, cluster analyses based on standardized Euclidean distance, Manhattan distance and Mikowski distance have been used. For the model with the zeros removed, the optimal solution is found in the one layer model.

1For the model with the zeros removed, the optimal solution is found in the one layer model.
used. In the model that uses Mikowski distance the value of $q$ has been set to three. For all these analyses, iVAT images have been made to obtain the best number of clusters. These images are displayed in Appendix A.6. Based on these images the number of clusters has been set to two. The function `nerfcm` has been used to perform the cluster analyses. This algorithm is obtained from Hathaway and Bezdek.

Table 6.8: Neural network: results for gas prediction based on daily summations using different distance measures

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>Nodes 1</th>
<th>Nodes 2</th>
<th>RMSE</th>
<th>CV-RMSE</th>
<th>SMAPE</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized Euclidean</td>
<td>6431</td>
<td>24</td>
<td>29</td>
<td>140.3261</td>
<td>0.3098</td>
<td>0.5</td>
<td>10</td>
</tr>
<tr>
<td>Manhattan</td>
<td>6431</td>
<td>28</td>
<td>25</td>
<td>139.0621</td>
<td>0.3076</td>
<td>0.4741</td>
<td>10</td>
</tr>
<tr>
<td>Mikowski power 3</td>
<td>6431</td>
<td>30</td>
<td>28</td>
<td>134.7254</td>
<td>0.2976</td>
<td>0.4606</td>
<td>10</td>
</tr>
</tbody>
</table>

The results from Table 6.8 show that the distance measures used did not increase performance. The best results are still obtained using Euclidean distance and two clusters.

Finally, occupation rate that has been used for electricity consumption prediction has also been used to predict gas consumption. The results are given in Table 6.9.

Table 6.9: Neural network: results for gas prediction using occupation rate

<table>
<thead>
<tr>
<th>Model</th>
<th>Nodes 1</th>
<th>Nodes 2</th>
<th>RMSE</th>
<th>CV-RMSE</th>
<th>SMAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>26</td>
<td>20</td>
<td>3.9372</td>
<td>0.3187</td>
<td>0.325</td>
</tr>
<tr>
<td>Occupation rate</td>
<td>24</td>
<td>26</td>
<td>2.7421</td>
<td>0.222</td>
<td>0.26</td>
</tr>
</tbody>
</table>

It can be obtained that adding occupation rate to the dataset has a positive effect on accuracy. Note that gas consumption is estimated on a 15 minute interval. As Table 6.9 indicates, the best model fit is achieved on this dataset including occupation rate. This model even outperforms the model based on daily summations of gas consumption!

Energy

In some countries (for example Poland) electricity and gas are charged together. Because of that it can be of use to predict total energy consumption of a building. The energy dataset consists of 6431 observations. For this model, daily summations of total energy consumption are taken. The model is tested without clusters. For improvement, a two or three cluster analysis is added to the dataset for normalized data. Table 6.10 shows the results. Again, 5-fold cross validation is used to obtain the best model and 10-fold cross validation for the performance indicators.

Table 6.10: Results for energy prediction based on daily summations

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>Nodes 1</th>
<th>Nodes 2</th>
<th>RMSE</th>
<th>CV-RMSE</th>
<th>SMAPE</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>No clusters</td>
<td>6431</td>
<td>31</td>
<td>25</td>
<td>6085</td>
<td>0.1183</td>
<td>0.2645</td>
<td>10</td>
</tr>
<tr>
<td>2 clusters</td>
<td>6431</td>
<td>30</td>
<td></td>
<td>5962</td>
<td>0.1166</td>
<td>0.2752</td>
<td>10</td>
</tr>
<tr>
<td>3 clusters</td>
<td>6431</td>
<td>26</td>
<td>25</td>
<td>5958</td>
<td>0.1167</td>
<td>0.2446</td>
<td>10</td>
</tr>
</tbody>
</table>

\(^2\)Note that for the two-cluster model the optimal solution is found in a one-layer configuration
The best specification for the energy dataset is obtained using the three-cluster model. However, the improvement is only minor compared to the model without clusters. Also, the improvement is less compared to the gas consumption model in Table 6.7. Predicting gas consumption on a 12-hour interval using occupation rate was not possible because of insufficient observations.

6.2.2 Residential buildings

For residential buildings, results were far less accurate compared to the commercial building dataset. Table 6.11 shows the results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Sample size</th>
<th>Nodes 1</th>
<th>Nodes 2</th>
<th>RMSE</th>
<th>CV-RMSE</th>
<th>SMAPE</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>No solarpanels</td>
<td>10000</td>
<td>23</td>
<td>15</td>
<td>0.1306</td>
<td>1.8879</td>
<td>1.0118</td>
<td>10</td>
</tr>
<tr>
<td>Solar panels</td>
<td>10000</td>
<td>19</td>
<td>16</td>
<td>0.1955</td>
<td>19.6983</td>
<td>1.1504</td>
<td>10</td>
</tr>
</tbody>
</table>

When solar panel data is included the results suggest a worsened fit. A higher RMSE in this case is due to the added data. The second specification is simply not the same data as the first. The other performance indicators also show a decrease in accuracy. Performance was compared to the regression results of the residential building model. These are shown in Figure 6.1.

![Figure 6.1: Regression results for residential buildings without solar panel data (right) and including solar panel data (left)](image)

When regression results are analyzed, we can definitely see an improvement in predictability. It is suggested that this follows from the global radiation variable that is included in the dataset. The global radiation shows high correlation to the revenues of the solar panels. Note that revenues are measured on meter CM2 as explained in Appendix A.5. Figure 6.2 shows this correlation.
Because of the disappointing results and lack of other information regarding the apartments and their inhabitants no further effort was spent on this dataset.

6.3 Sensitivity analysis

To check the impact of different variables on the outcome, a sensitivity analysis is often used. This can be done using many different techniques. For neural networks, it is common to check the impact of variables in different datapoints. To obtain the results, all variables are normalized and all observations are replaced by the mean value. Then, several points in the spectrum are used as baseline. In this case, all variables are normalized to values between zero and one. Then, arbitrary points are chosen between zero and one. On each of these points, a small fraction $\Delta X$ is added or subtracted. In this analysis, baseline points of [0.1, 0.2, ..., 0.9] have been chosen and values of $\Delta X$ have been set to 0.01 and 0.02. This is illustrated in Figure 6.3.
Figure 6.3 shows an example where the sensitivity of variable $X$ to output $Y$ in point 0.2 is higher than in point 0.9. Each combination of baseline points $\pm \Delta X$ is used as input for each variable while the other variables remain constant at their mean value. For all the combinations, the absolute value of the difference between the baseline outputs and baseline plus deviation outputs is used as impact measure. The sum of all differences is then used as final impact measure for each variable. It enables the variables to be ranked. Important to note is that this analysis does not imply what the effect of each variable is on output. It simply explains how much a small deviation of each input has an effect on the output. The analyses has been done and the results are given in Table 6.12. The inputs that have been used are the observations of the testset. Note that the scores refer to values of $10^7$.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>1.129</td>
<td>1</td>
</tr>
<tr>
<td>BVO</td>
<td>1.109</td>
<td>2</td>
</tr>
<tr>
<td>Dummies LN2</td>
<td>1.105</td>
<td>3</td>
</tr>
<tr>
<td>Heating season</td>
<td>1.105</td>
<td>4</td>
</tr>
<tr>
<td>Devices</td>
<td>1.104</td>
<td>5</td>
</tr>
<tr>
<td>$U$</td>
<td>1.102</td>
<td>6</td>
</tr>
<tr>
<td>Label F</td>
<td>1.102</td>
<td>7</td>
</tr>
<tr>
<td>$Q$</td>
<td>1.102</td>
<td>8</td>
</tr>
<tr>
<td>Comf_settings</td>
<td>1.1</td>
<td>9</td>
</tr>
<tr>
<td>Office</td>
<td>1.099</td>
<td>10</td>
</tr>
<tr>
<td>Dummies LN1</td>
<td>1.095</td>
<td>11</td>
</tr>
<tr>
<td>Datacenter</td>
<td>1.094</td>
<td>12</td>
</tr>
<tr>
<td>Comf_heating</td>
<td>1.089</td>
<td>13</td>
</tr>
<tr>
<td>Label G</td>
<td>1.088</td>
<td>14</td>
</tr>
<tr>
<td>Other</td>
<td>1.088</td>
<td>15</td>
</tr>
<tr>
<td>Occupation</td>
<td>1.087</td>
<td>16</td>
</tr>
<tr>
<td>Dummies LN3</td>
<td>1.086</td>
<td>17</td>
</tr>
</tbody>
</table>

Based on the results of Table 6.12, it can be obtained that temperature is the most volatile factor.

6.3.1 Application: forecast gas consumption as function of temperature

The same analysis as in Section 6.3 has been performed for gas. Based on the small deviations in a function, a graph can be created that shows the derivative of gas consumption relative to temperature. It was found that taking the average value for all variables does not make sense, in that case a hypothetical non existent building would be evaluated. Instead of that, values for building properties can be specified according to what a specific building has. This has been applied for a building where gas consumption is evaluated relative to temperature. The other climate observation values have been set to the average. In this case, not the absolute value of changes has been applied but the real value for 1000 different temperature inputs. Figure 6.4 shows the predicted gas consumption while Figure 6.6 shows the marginal differences.
Although it can be expected that gas consumption is reduced when temperature increases, some abnormalities are present in Figure 6.4. For example, it is strange that when temperature changes from four to five degrees the gas consumption tends to increase rather than decrease. To check the observed values of gas consumption have been divided into classes of one degree Celsius. Then, the average value in that class is calculated. Figure 6.5 shows the results.

Values on the horizontal axis refer to upper limits of the classes. Note that in Figure 6.4 the values for relative humidity, global radiation and occupation rate are constant at the average while this is not the case in Figure 6.5. Still, the pattern is comparable showing an increase of gas consumption at four degrees Celsius. To check the influence of the number of observations in a class on the predicted gas consumption, a frequency table has been created. It was obtained that only the extreme low and extreme high temperature observations are underrepresented. This excludes the possibility where the model is interpolating. Furthermore, the observed gas consumption was compared to the predicted gas consumption of the neural network resulting in a pattern very much comparable to Figure 6.5.
Figure 6.6: Marginal difference in gas consumption for location P.132708.02

Figure 6.4 and Figure 6.6 can be used to find abnormal patterns in gas consumption. For example, it is strange that when temperature changes from four to five degrees the gas consumption tends to increase rather than decrease. But as Figure 6.5 shows this pattern is present in reality.
Chapter 7

Conclusion

The following chapter concludes the thesis. First, a summary is given. Then, the main conclusions are stated including answers to the research questions. Furthermore, the limitations regarding this research project are given, followed by relevant recommendations for future research projects and the company. Finally, the deployment sections describes the developments the company is willing to make in order to implement the model and gather additional data.

7.1 Summary

Zero Budget Sustainability (ZBS) is a service that is offered by Ploos Energieverlening since 2011. In short, in this service Ploos acts as intermediary between financers that want to invest and building owners that want to implement energy saving measures but lack budget. By bringing these two together the owner of the building can reduce its energy consumption while the financer has access to a profitable investment. However, because both customers and financers have insufficient confidence in the projected energy savings only two customers have used the service since 2011. Based on real world consumption data from smart meters Ploos wants to increase confidence in ZBS by showing that energy consumption of buildings can be estimated with high accuracy.

Data mining methods and techniques are applicable to transform low-level data into actionable results. A scientific literature study obtained that an artificial neural network is suitable to estimate the energy consumption of buildings. Furthermore, a cluster analysis can be used to improve the data and obtain a higher accuracy. The CRISP-DM framework was used to perform a structured approach to the research project.

The models were applied to a dataset containing 11 commercial buildings where energy saving measures are implemented. For these buildings, 18 months of consumption data, climate data and building properties were available. The dataset was improved by creating binary time variables. The model was also tested on a dataset containing apartment buildings.

The results show that the energy consumption of buildings can be estimated with high accuracy. Electricity consumption of commercial buildings measured on a 15-minute time interval can be predicted resulting in only 8.23 percent symmetric errors. Over a longer period, the difference between the predicted and observed electricity consumption is 0.3 percent. Gas consumption was found to be less predictable. However, by adding behavioral data gas consumption estimation increased. For the best model, the difference between observed and predicted gas consumption is 4.3 percent. The models have been tested on random datasamples proving that the results are still accurate for different datasets. Lack of behavioral data for apartment
buildings resulted in a low model fit although adding re-supplied electricity from solar panels
did improve the results.

A contribution to the scientific literature is made by adding low-level building properties to
the dataset. So far, only global building characteristics have been used. The model is built
such that it can handle 20 different low-level building variables. Future research objectives for
further improvements have been identified. They are written in Section 7.5.

Following the research project several recommendations are made to the company. These are
described in Section 7.4. Finally, the model also enables the company to extend its services.
More about this can be found in Section 7.6.

7.2 Conclusions

In this section, the relevant conclusions that are obtained from this model are described. This
is done by answering the problem definition from Section 1.3 as well as by answering the
research questions from Section 1.6. Furthermore, the possible usage of the model for the
company is described.

7.2.1 Adressing the problem description

In the literature review (Chapter 2), it has been obtained that ZBS can help to implement
energy saving measures (and save energy) without increasing costs for customers. The main
problems that could not be answered by literature is that the energy consumption of buildings
is predictable under specific circumstances. The model that was built shows that by using the
right inputs, electricity consumption is predictable as well as gas consumption, although to a
lower extend. The highest accuracy was obtained for electricity consumption prediction using
occupation rate. In this model, the symmetric percentage error was only eight percent. More
about how the problems regarding ZBS are solved in practice can be found in Section 7.6.

7.2.2 Answering the research questions

In this subsection, the research question and the subquestions will be answered based on the
results from Chapter 5.

- **Can the energy consumption of buildings be predicted with high accuracy
  within the context of ZBS?**

  To obtain a useful dataset, effort was first spent on creating better classes of energy sav-
ing measures. As a result, the scheme from Figure 4.4 was obtained. The measures have
been classified such that they will have comparable effects on energy consumption of buildings.
The Neural Network resulted in a minimum symmetric percentage error of 8.23 percent using
occupation rate and 28.71 percent without using occupation rate for electricity consumption
prediction on a 15 minute base. In these models, the CV-RMSE is below the desired level of
25 percent which makes the models acceptable in scientific terms. Gas consumption turned
out to be less predictable compared to electricity. The best model for gas prediction on a 15
minute interval still shows 66.67 percent symmetric error. CV-RMSE is 42.89 percent. How-
ever, when occupation rate is included results improved significantly. In the absense of this
information, gas consumption can be predicted on a 12 hour interval resulting in an accuracy around the indicated threshold. Finally, total energy consumption shows high accuracy as well.

The model has also been applied to residential buildings. No information about residents or behavior was available and results were far from satisfying. Re-supplied electricity from solar panels was subtracted from the consumption meters of the apartments in order to increase predictability but accuracy was still insufficient.

From these results, the research question can be answered: Energy consumption of commercial buildings can be predicted within the context of ZBS when the right data is available.

- **Which variables are good predictors for energy consumption based on the available data?**

  First, building properties have been identified as important predictors. Next to general properties such as floor area and other global indicators, low level characteristics are added as well. Although the exact effect could not be tested, it is expected that by adding more data the results will be significant.

  Based on expert knowledge, three climate variables have been selected from the database, being temperature, global radiation and relative humidity. These variables are relevant for the energy saving measures that have been implemented until today. In future, other variables could be relevant as well such as wind speed in case a windmill is built. This all depends on the development of energy saving. Therefore, the climate variables included should be revised from time to time.

  The time of the day turned out to be relevant as well. For electricity consumption, it was found that distinction between office hours and non-office hours is enough to increase performance of the model. For gas, the first three hours of the day are relevant as well. Furthermore, the season the time interval is in has also a small effect on performance. Finally, adding occupation rate to the dataset resulted in a huge increase of performance for both electricity and gas consumption prediction. Occupation can be considered as a behavioral variable. Therefore, it is expected that other behavior inputs can increase accuracy as well.

- **Which models are applicable to the described problem?**

  In the literature study, two models have been identified as relevant for energy consumption prediction. These are artificial neural networks and case based reasoning. From these models, only an artificial neural network was tested. The choice for not building a case based reasoning model has been made because when other data was added to the model, accuracy increased a lot. It was therefore expected that gathering more data can result in higher accuracy than building a second model.

  Furthermore, a cluster analysis on the dataset was made to obtain load and climate profiles for the buildings. The memberships to the clusters were added to the gas and to the energy model. It was shown in the gas model that adding cluster memberships to the dataset can result in a reduction of symmetric errors up to four percentpoint.
7.3 Limitations

This research project has some limitations which are described in this section.

First, the amount of energy saving measures in the dataset is small. Although effort was spent to increase this amount for some measures still only one implementation is present. This removes the possibility to analyze between buildings that have these properties as well. Furthermore, it was not clear when exactly these measures have been implemented. Because of that, the date 01-01-2014 has been chosen as starting date. A limitation of this choice is that no comparison can be made for the building before and after the measures have been implemented.

Another limitation is that very little behavior information was available for both commercial and residential buildings. Occupation rate was available for three buildings but only measured on the whole day. As scientific literature indicates, even for office buildings behavior can improve results.

Finally, complex problems can be solved by using deep neural networks. Although scientific literature does not indicate the use of these models, it is worthwhile to try. Because of time constraints (deep neural networks require more computation time), this has not been done. Moreover, testing deep neural networks for this dataset requires a different experimental setup. For example, pair multiple processors to increase computation speed.

7.4 Recommendations

During the writing of this thesis, several recommendations have been made to the company. Especially those needed for data collection. As has been described in Chapter 4, a tool has been built by the IT department to integrate building characteristics, climate data and consumption data. Also, classification of energy saving measures has been improved. After the results of the models were obtained, the following recommendations are made to the company:

- More customer involvement and better feedback

  The data for occupation rate that was supplied helped to improve accuracy of the model. To obtain this information, many e-mails had to be sent in order to convince the customer of the importance of this data. It became obvious that at this moment customers are not convinced that sharing information can also lead to benefits for themselves. Also, companies view privacy issues as more important than having access to better services supplied by Ploos. Therefore, Ploos should spend more effort to convince customers that their data is in good hands and that sharing information can lead to better services that benefit customers as well. Also, customers might implement energy saving measures themselves without Ploos knowing of it. For better consumption estimation, this information should be available too.

- Integrate more data into the database and update regularly

  Much of the information that Ploos possesses is written on Excel sheets that are hidden somewhere in a folder on a hard disk. Some of this information was used for this thesis, but in
order to obtain it many documents had to be written and manually added to the database. A more efficient approach would be to integrate them in the system automatically. For example, when an EPA-U analysis is made, current consumption shares should be integrated because they could be a good energy consumption predictor. Also, it was found that documents contained implemented energy saving measures that were not present in the database. These have been added manually but this process was time consuming.

• **Standardize documents**

During the data collection phase, it turned out that many different documents are used for data storage. Furthermore, it was found that names used in the database are not consistent to those on the documents itself. As a result, data had to be integrated manually which is time consuming and prevents an automatic analysis. This was for example the case with occupation rate data.

• **Implement more partial meters**

Although adding solar panel data to the residential building dataset did not lead to better results, it is expected that more partial meters can improve accuracy of models for commercial buildings. Installing these meters also enables possibilities for analyses on a smaller scale. This can be useful if a company uses flexible working places but also to test the effect of energy saving measures themselves. It was found that the meters themselves consume very little electricity so that should not be a reason to withhold.

### 7.5 Further research

From this thesis, new ideas have emerged for further research. This section gives an overview about what could be done to improve the research and models.

• **Add more data**

Several possible variables have been identified as good energy consumption predictors. For example, occupation rate did improve accuracy. This could be further improved by adding *real time* occupation rate where one measurement is taken each hour or each 15 minutes. Also, consumption shares could be useful. If for example LED lighting is implemented while the consumption share of lighting is small the effect on the total energy consumption is also expected to be small.

• **Make a distinction between for the year an energy saving measure is implemented**

The quality of building materials and climate systems is increasing constantly. As a result, energy saving measures that have been implemented 10 years ago will not have the same effect on energy consumption as one that is implemented today. This was not relevant for this thesis because all measures have been implemented shortly before 2014. In order to maintain the model and extend it in the future, this information should be added.
• Test the Case-based reasoning model

The Case-based reasoning model has not been tested. Scientific literature indicates the usefulness of this technique. The specifications for the model have been written so it should not be too much work to try and test it. This is especially the case when more data is available.

• Test for more hidden layers

So far, deep neural networks have not been used for energy consumption prediction. It was found that the two-layer model produced on average better results than the one-layer model. Therefore, testing for three or more layers could improve accuracy even more.

• Adding trend analysis for the gas model

Even when behavior variables are added, gas consumption shows to be less predictable than electricity. This is mainly because of the thermostatic effect a gas fired boiler has. Therefore, the last few gas measurements are expected to influence current consumption of gas. This analysis could be supported by adding the absorption capacity of a building’s heating system to the dataset.

7.6 Deployment

The company is satisfied with the results from this thesis and has decided to implement the model in their system. Therefore the Matlab software package including the Neural Network Toolbox and the Fuzzy Logic Toolbox has to be bought. By integrating the model, the company has obtained three possible applications.

• Calculate the Marco Maaskant payback Time (MMT)

If more energy saving measures are included in the database, possibilities to calculate effects arise. In future, the payback time of specific measures can be calculated. This can be done by setting the dummy variable for that specific measure from zero to one. To calculate the energy consumption with that specific measure, a moving window of the ten year average climate values can be used.

From the consumption estimations the energy costs can be calculated. The system has to be adapted for that because the energy costs also depend on the type of connection a building has. The IT department will make these changes soon so that when enough information about energy saving measures is available and the results are satisfactory the improved payback time can be calculated. It is called the Marco Maaskant payback Time (MMT).

• Online tool for customers

Ploos uses the slogan “grip on energy” which implies professional employees, solid systems, intelligent models and online accessible and reliable information. In order to show the last two points following from this slogan, the neural network will be used as an online tool for customers. A regression graph will be showed on the website that displays actual energy consumption compared to forecasted volume. Also the customer can insert specific inputs for climate where the model returns a value for the estimated energy consumption.

1http://www.ploos-energie.nl/nieuws/wijziging-naam-in-ploos-energieverlening/
• **Tool for sales employees**

Sales employees in their turn can use the model while visiting their potential customers. If a company requests a service offered by Ploos, the model can be used to show that Ploos uses accurate models to support their services. It can also be used to convince potential customers that the energy savings proposed are actually realized. This can be done in combination with the projected payback time that also follows from the model. When a regression plot is shown in combination with a list of real outcomes and outcomes from the model, it is expected that customers are more likely convinced of the accuracy of the calculation module.
Appendix A

This section provides additional information regarding several subjects that are discussed in the thesis. First, a transcription of the interview is included. Second, a hypothetical scenario regarding the implications of ZBS is included, followed by a detailed discussion about outlier detection. Also, information about the consumer energy price is given. Furthermore, the scheme from the solar panels of the apartment buildings is presented. The final section shows iVAT images for different distance measures used in the cluster analysis.

A.1 Transcription classification interview (Dutch)

This appendix gives the transcription of the interview that was held to classify the energy saving measures into more relevant categories. Only the relevant sentences are transcribed. The interview was conducted in Dutch. The details of the interview are as follows:

File: 1.13PM, Jun16.mp3
Duration: 31:20
Date: June 16, 2015

Interviewer: Paul Vermeulen, vice-director
Respondent 1: Johan Aarts, senior consultant
Respondent 2: Martijn Ploos van Amstel, IT supervisor
Supervisor: Marco Maaskant, researcher

Paul Vermeulen: Waar het om gaat is van, er is een hele lijst met maatregelen en locaties en besparingsinitiatieven en weet ik het allemaal en waar het nou om gaat is van die willen we classificeren. Nou hebben we in het systeem al een classificatie staan en dat zijn 4 of 5 onderwerpen betreffende verlichting, comfort, constructie, apparaat. Maar dat is niet erg divers. De vraag is: moet je dat verder uitsplitsen? Kunnen we dat verder uitsplitsen zodat de grip op de maatregelen wordt verhoogd? Want als ik het goed heb is dat jouw vraag.

Marco Maaskant: Ja.

Johan Aarts: Hoe bedoel je verder uitsplitsen, want hier staat natuurlijk de naam is aanpassen ledverlichting; dan is het ledverlichting. Maar daar staat ook nog een maatregel voor bijvoorbeeld TLD vervangen door TL5 verlichting. Ik denk dat je wel redelijk kunt clusteren als je dat bedoeld. Want er zijn wel heel veel maatregelen die je op dezelfde manier kunt benoemen. De ene keer staat er dubbel glas toepassen en de andere keer is het enkel glas vervangen door HR++ glas.

Paul Vermeulen: Heeft het toegevoegde waarde om bijvoorbeeld bij de schil aan te geven op welk gebouwdeel het betrekking heeft? Een muur, glas, dak?

Johan Aarts: Nee.
Paul Vermeulen: In mijn beleving ook niet alleen gewoon ook even over jou want ik heb dat de vorige keer ook aangegeven als je naar de nieuwe isolatienormen gaat zie je daar eigenlijk geen verschil want volgens mij maakt het geen donder meer uit want alles is tegenwoordig 5\(^1\) 3.5 is de oude....

Johan Aarts: Veel wel ja.

Paul Vermeulen: Dus dan zie je dat het dak en de wanden is allemaal op een RC waarde van 3.5 of van 5 afgeschoten dus dat is allemaal uniform.

Marco Maaskant: Oke want dat zie ik hier wel af en toe tussen staan: RC 3.5, RC 2.5....

Johan Aarts: Ja maar er staat ook duidelijk voor of het geveldak is dakisolatie of gevelisolatie of vloerisolatie. Dus ik denk wel dat je dat kunt specificeren allemaal. Alleen als je het over constructies hebt zullen de meeste maatregelen niet zijn doorgevoerd.

Paul Vermeulen: Nee maar dat was even niet de vraag. De vraag is wil je classificeren? En wil je dus eh....

Johan Aarts: Clusteren bedoel je?

Paul Vermeulen: Je kunt, onafhankelijk van hoe wij het op de website doen, uiteindelijk wil ik mogelijk met Martijn\(^2\) gaan bespreken of we die classificatie verder gaan toepassen maar we willen eerst weten wat de toegevoegde waarde is. Vervolgens komt, wij hebben het vermoeden van de professor en van Marco dat een betere classificatie toegevoegde waarde kan hebben. En nou is de vraag gesteld door Marco eigenlijk van hoe zouden we dat beter kunnen classificeren? Nou toen heb ik gezegd van er is maar één manier om daar achter te komen dat is een gesprek met ons drieën, dat wij samen even er naar kijken want wij zijn degenen die weten hoe die technische maatregelen redelijkerwijs vertaald kunnen worden.

Johan Aarts: Ja.

Paul Vermeulen: Dat is de reden dat we hier zitten dus wij moeten er even samen over brainstormen want het kan best zijn dat we zelf zeggen: ja maar een constructie, kijk ik zie hier heel veel glas, trippel glas, glas en isolatie, vloerisolatie, dubbel glas vervangen door HR++ glas. Daar staat gewoon heel veel in en heeft het zin om dat apart te maken? Wat ik zie is dat een warmtepomp is een apparaat, maar dat is natuurlijk niet zo, dat is een comfortinstallatie.

Johan Aarts: Dat geloof ik ja.

Paul Vermeulen: Dus er staan ook fouten in. Verhogen temperatuur van de serviceruimte, het is ook maar de vraag of je dat als apparaat moet zien.

Johan Aerts: Dat moet je wel als apparaat invoeren.

Marco Maaskant: Het is een instelling.

Johan Aarts: Maar dat moet je wel als apparaat doen.

Paul Vermeulen: Ja maar is dat voor de classificatie van maatregelen, is dat een apparaatmaatregel of een comfortmaatregel?

Johan Aarts: Misschien comfort, maar daar kan je een discussie over hebben.

Paul Vermeulen: Het gaat niet om goed of fout.

Johan Aarts: Het gaat niet om mensen, als je comfort mensen bedoeld dan is het een apparaat.

Paul Vermeulen: Het is gewoon een vraag, hoe maak je helderheid in de enorme massa van data die er staat. En apparaten is eigenlijk iets... elimineer apparaten! Apparaten is zo divers Marco.

Johan Aarts: Nou, apparaten komt zo weinig voor.

\(^1\)This implies an RC-coefficient of 5

\(^2\)This refers to Martijn Ploos van Amstel, IT supervisor
Paul Vermeulen: Comfort wel.
Johan Aarts: Comfort wel, die moet je dan gaan specificeren voor verwarming, verlichting, koeling?
Paul Vermeulen: Ja, daar zou je een verdiepingsslag kunnen maken.
Johan Aarts: Dan zou ik het bij constructie ook doen. Dat zijn er ook niet veel, dat is: gevel, dak, vloer, ramen (glas). Klaar.
Paul Vermeulen: Oke
Johan Aarts: Dan heb je dat. Dan heb je comfortinstallaties, dan heb je verwarming, koeling, luchtbehandeling, stoombevochtiging, warm tapwater en zonnepanelen, dat valt dan onder comfort.
Paul Vermeulen: Ik zou dat "duurzaam" maken, stel je voor dat we een keer een windmolen tegenkomen. Wat dat betekend is dat je een lijstje krijgt.
Johan Aarts: Ja.
Paul Vermeulen: Wat dat betekend is dat je een lijstje krijgt, we hebben verlichting, we hebben comfort, constructies.
Johan Aarts: Sectoren zou je nog kunnen noemen.
Paul Vermeulen: Nee sectoren vind ik moeilijk, dat zou ik niet doen. We hebben apparaten, comfort, constructie, overig en niet gedefinieerd.
Johan Aarts: Niet gedefinieerd is fout, dat is altijd overig.
Paul Vermeulen: Overig en apparaten dat gaat hem nooit niet worden. Dat is het vervangen van een koffiezet apparaat, ja als er een warmtepomp in staat dan is dat een warmte ding maar dat zijn ze niet.
Johan Aarts: Nee die staat dan gewoon verkeerd. Daar staat dan gewoon een foutje in.
Paul Vermeulen: Comfort: LBK, vocht, verwarmen, koelen. Mis ik er een Johan?
Johan Aarts: Ja duurzaam, warm tapwater, instellingen
Paul Vermeulen: Is dat bij comfort?
Johan Aarts: Nee die staat dan gewoon verkeerd. Daar staat dan gewoon een foutje in.
Paul Vermeulen: Comfort: LBK, vocht, verwarmen, koelen. Mis ik er een Johan?
Johan Aarts: Ja duurzaam, warm tapwater, instellingen
Paul Vermeulen: Wat dat betekend is dat je een lijstje krijgt, we hebben verlichting, we hebben comfort, constructies.
Johan Aarts: Sectoren zou je nog kunnen noemen.
Paul Vermeulen: Nee sectoren vind ik moeilijk, dat zou ik niet doen. We hebben apparaten, comfort, constructie, overig en niet gedefinieerd.
Johan Aarts: Niet gedefinieerd is fout, dat is altijd overig.
Paul Vermeulen: Overig en apparaten dat gaat hem nooit niet worden. Dat is het vervangen van een koffiezet apparaat, ja als er een warmtepomp in staat dan is dat een warmte ding maar dat zijn ze niet.
Johan Aarts: Nee die staat dan gewoon verkeerd. Daar staat dan gewoon een foutje in.
Paul Vermeulen: Comfort: LBK, vocht, verwarmen, koelen. Mis ik er een Johan?
Johan Aarts: Ja duurzaam, warm tapwater, instellingen
Paul Vermeulen: Wat dat betekend is dat je een lijstje krijgt, we hebben verlichting, we hebben comfort, constructies.
Johan Aarts: Sectoren zou je nog kunnen noemen.
Paul Vermeulen: Nee sectoren vind ik moeilijk, dat zou ik niet doen. We hebben apparaten, comfort, constructie, overig en niet gedefinieerd.
Johan Aarts: Niet gedefinieerd is fout, dat is altijd overig.
Paul Vermeulen: Overig en apparaten dat gaat hem nooit niet worden. Dat is het vervangen van een koffiezet apparaat, ja als er een warmtepomp in staat dan is dat een warmte ding maar dat zijn ze niet.
Johan Aarts: Nee die staat dan gewoon verkeerd. Daar staat dan gewoon een foutje in.
Paul Vermeulen: Comfort: LBK, vocht, verwarmen, koelen. Mis ik er een Johan?
Johan Aarts: Ja duurzaam, warm tapwater, instellingen
Paul Vermeulen: Wat dat betekend is dat je een lijstje krijgt, we hebben verlichting, we hebben comfort, constructies.
Johan Aarts: Sectoren zou je nog kunnen noemen.
Paul Vermeulen: Nee sectoren vind ik moeilijk, dat zou ik niet doen. We hebben apparaten, comfort, constructie, overig en niet gedefinieerd.
Johan Aarts: Niet gedefinieerd is fout, dat is altijd overig.
Paul Vermeulen: Overig en apparaten dat gaat hem nooit niet worden. Dat is het vervangen van een koffiezet apparaat, ja als er een warmtepomp in staat dan is dat een warmte ding maar dat zijn ze niet.
Johan Aarts: Nee die staat dan gewoon verkeerd. Daar staat dan gewoon een foutje in.
Paul Vermeulen: Comfort: LBK, vocht, verwarmen, koelen. Mis ik er een Johan?
Johan Aarts: Ja duurzaam, warm tapwater, instellingen
Paul Vermeulen: Wat dat betekend is dat je een lijstje krijgt, we hebben verlichting, we hebben comfort, constructies.
Johan Aarts: Sectoren zou je nog kunnen noemen.
Paul Vermeulen: Nee sectoren vind ik moeilijk, dat zou ik niet doen. We hebben apparaten, comfort, constructie, overig en niet gedefinieerd.
Johan Aarts: Niet gedefinieerd is fout, dat is altijd overig.
Paul Vermeulen: Overig en apparaten dat gaat hem nooit niet worden. Dat is het vervangen van een koffiezet apparaat, ja als er een warmtepomp in staat dan is dat een warmte ding maar dat zijn ze niet.
Johan Aarts: Nee die staat dan gewoon verkeerd. Daar staat dan gewoon een foutje in.
Paul Vermeulen: Comfort: LBK, vocht, verwarmen, koelen. Mis ik er een Johan?
Johan Aarts: Ja duurzaam, warm tapwater, instellingen
Paul Vermeulen: Wat dat betekend is dat je een lijstje krijgt, we hebben verlichting, we hebben comfort, constructies.
Johan Aarts: Sectoren zou je nog kunnen noemen.
Paul Vermeulen: Nee sectoren vind ik moeilijk, dat zou ik niet doen. We hebben apparaten, comfort, constructie, overig en niet gedefinieerd.
Johan Aarts: Niet gedefinieerd is fout, dat is altijd overig.
Paul Vermeulen: Overig en apparaten dat gaat hem nooit niet worden. Dat is het vervangen van een koffiezet apparaat, ja als er een warmtepomp in staat dan is dat een warmte ding maar dat zijn ze niet.
Paul Vermeulen: Zo heb je gevelplaten, dat is allemaal wand. En uiteindelijk gaat het er over wat je met die wand gaat doen. En als wij zeggen er gaat een verbetering plaats vinden, dan is dat of terwijl isolatie, of het is een vervanging van de wand, maar dat zou wel heel rigoureus zijn.

Marco Maaskant: Maar dan komt het allemaal op hetzelfde neer, het resultaat zal zijn dat hij RC 5 zal zijn.

Paul Vermeulen: Dat is altijd het gevolg.

Marco Maaskant: Daar kan je wel wat mee doen, je kan bijvoorbeeld zeggen RC 2.5, RC 3.5, RC 5. Dat is misschien wel nog interessant om te doen.

Paul Vermeulen: Ja maar dan moet je er 2 bij gaan houden, de beginwaarde en de eindwaarde. Daar heb je gelijk in maar dat betekend echt dat je een beginwaarde en een eindwaarde moet pakken. Want dan zeg je eigenlijk ik heb hier een wand....

Johan Aarts: Die was 0.5.

Paul Vermeulen: Deze was 0.5 en we stellen voor om die naar 2.5 te brengen. Dat is dan wat je krijgt. Maar dan moet je dus beide waarden registreren.

Marco Maaskant: Nee dat hoeft denk ik niet, RC is, je gaat gewoon kijken naar wat gaan we verbruiken in de nieuwe situatie. In de nieuwe situatie is RC 0.5 niet meer van toepassing.

Paul Vermeulen: Maar hoe weet jij dan wat de verbeterslag is?

Marco Maaskant: Je hebt het verbruik voor en na.

Paul Vermeulen: Oh het oude verbruik, oke.

Johan Aarts: Wat wil je met die classificaties uiteindelijk gaan doen?

Marco Maaskant: Nou je gaat kijken of je dus op basis van historische gegevens ga je kijken of je je nieuwe gebouw kan classificeren in een bepaalde groep.

Paul Vermeulen: Wat er eigenlijk gebeurd is je hebt de maatregelen aangegeven met een terugverdientijd van 3 jaar. En die maatregelen zijn: je brengt een deel van de wand, of de hele wand, op een RC waarde van 2.5. Daar heb jij een energetische besparing bijgezet, en die energetische besparing wordt getoetst. En over de langere loopijd gaan we kijken of we met ZBS, want dat is eigenlijk het onderzoek, of je dan op basis van die geclassificeerde maatregelen of dat ZBS voldoende houvast biedt om te garanderen dat we die besparing gaan realiseren. Dat is wat dit onderzoek Big Data beoogt. Kunnen wij ZBS zodanig stevig in de markt zetten dat het echt een garantie biedt. Ja? Kan je met ZBS waarmaken wat wij met zijn allen roepen. Nu zeggen we: de uitgangspunten zijn bepalend, en de wijzigingen die in de uitgangspunten worden aangebracht. En dat is heel divers, dus daar heeft hij ook nog een hele studie mee. Dat is nog een aardige uitdaging, welke uitgangspunten liggen vast en welke niet? Ik snap wat je zegt, uiteindelijk zeg je: alleen de nieuwe situatie is wenselijk.

Marco Maaskant: Ja

Paul Vermeulen: Dan heb je bij verlichting, het enige wat ik me daar bij kan voorstellen dat zijn de drie die de dingen ook kent, dat is binnen, buiten en sfeer. Maar kunnen we dat classificeren, kunnen we dat makkelijker herkennen?

Johan Aarts: Op basis van de tekst bij de maatregelen niet.

Paul Vermeulen: Dus dat onderscheid is op zich wel te maken maar het is moeilijk te herkennen.

Johan Aarts: Bijna alles is binnen. En bij sommigen staat het er ook bij, buitenverlichting. Ik denk dat verlichting ook moeilijk is, maar dat zijn wel de meeste maatregelen.

Marco Maaskant: Maar je hebt wel heel veel dingen, het aantal watts staat er bij. Dus daar zou je misschien ook iets mee kunnen.

Paul Vermeulen: Dat is wel heel moeilijk.
Johan Aarts: Je zou ook nog kunnen doen: LED en overig.

Paul Vermeulen: Eigenlijk is LED en overig gewoon een datum van de maatregel, na een bepaalde datum is het alleen nog maar LED. We hebben een overgangsgebied gehad waarin we geen LED deden.

Johan Aarts: Nou ik denk nog steeds wel.

Marco Maaskant: Ik heb ook nog wel eens gezien dat er halogeen werd geplaatst maar dan met minder vermogen. In plaats van een andere. Dus dat is dan waarschijnlijk sfeerverlichting.

Paul Vermeulen: Dat kan.

Marco Maaskant: En wat is PL?

Paul Vermeulen: PL zijn die dingetjes die we hier in het trappgat hebben, die mensen in hun keuken hebben, dat zijn de ronde lampen die in het plafond zitten waar je van die staflampjes inzet. En verlichting, dat is de grootste groep.

Johan Aarts: Dan hebben we nog schakeling.

Paul Vermeulen: Maar dat kan je er allemaal niet uithalen. Het probleem is: je krijgt het er allemaal niet uitgehaald. Je kan uit de tekst bij de maatregelen niet de diversificatie die nu eigenlijk vraagt er uithalen. We kunnen dat naar de toekomst er wel inbrengen, want dat is gewoon een afspraak maken.

Johan Aarts: Maar je kunt zo wel eens kijken, want de grootste verlichtingsmaatregelen zijn: LED verlichting toepassen, in welke vorm dan ook dus dat kan zijn halogeen verlichting vervangen door LED spots, downlights vervangen door LED spots.

Paul Vermeulen: Dat is niet relevant, want hij zei net ik heb alleen maar de werksituatie nodig. De nieuwe situatie, dus dat is gewoon LED.

Johan Aarts: Ja maar dan heb ik alleen maar de maatregelpakketten nodig.

Paul Vermeulen: Nee want je weet niet welke maatregelen zijn doorgevoerd. Wij maken wel maatregelpakketten maar dat zijn niet altijd de maatregelpakketten die zijn doorgevoerd.

Johan Aarts: Nee dat klopt.

Paul Vermeulen: Als het zeker is dan kan je dat doen maar dat kunnen we niet. Want de bulkdata die er nu is is zo divers dat een goed onderzoek daarop gewoon te moeilijk is. Dat is eigenlijk het probleem. Dus je wil daar een diversificatie op loslaten. We hebben met constructies en comfort zinnige dingen gedaan. En uiteindelijk gaat het er om: kunnen we daar bij verlichting ook iets zinnigs mee doen?

Marco Maaskant: Ja dat is het

Johan Aarts: Ja dan kom je bij wat ik net zei, LED, geen LED, buitenverlichting. Dat is dan schakelklok ofzo. Dat is ook wel redelijk goed in de tekst te zien denk ik. LED schakeling.

Paul Vermeulen: Dan zou je nog kunnen kijken, doe jij nog gewone TL? TL5 hè?

Johan Aarts: LED, TL5, schakeling, dat kan zowel binnen als buiten zijn, en overig. Maar buiten, waarom zou je buiten apart doen?

Paul Vermeulen: Dat weet ik niet.

Johan Aarts: Het heeft geen invloed op het label, dus dat is het enige.

Paul Vermeulen: Dat hoeft ook niet, het gaat hier niet om invloed op het label.

Marco Maaskant: Invloed op het verbruik.


Paul Vermeulen: Oke, dat is prima. Ik denk dat we nu een aardig lijstje hebben. En nou is natuurlijk de vraag hoe gaat we dat doen?

Marco Maaskant: In het systeem bedoel je?

Paul Vermeulen: Nou in het systeem dat ga ik met Martijn overleggen maar jij hebt nu een
lijst en jij wil dat die lijst straks vertaald wordt en dat dit er in staat.

**Marco Maaskant:** Dat is handwerk.

**Paul Vermeulen:** Voor een groot deel wel ja.

### A.2 Scenario

To indicate the current problems within ZBS, a hypothetical scenario has been written. This scenario also shows solutions that this thesis aims to propose.

Car dealership and garage "Krikkebak" is a company that sells used cars and provides affordable maintenance service for cars. Figure A.1 gives an impression of the company.

The company has currently 19 employees (2 managers, 4 salesmen, 7 mechanics, 2 administrators, a secretary, 2 buyers and a cleaner) of which some are part-time employed. The company owns a 1960's building that houses the showroom, the garage and a few offices. Recently, a new financial manager has been employed. Profit margins have been dropping for years and the manager has decided that the company has to lower its expenditures in order to survive in the near future.

After considering the opportunities to reduce costs, the manager finds out that the monthly energy bill of the company is one of the major expenditures the company is facing. This bill is paid from an energy budget that has been rising by approximately 5 percent each year. The manager decides to compare energy performances of the company and uses the *Benchmark service* offered by Ploos Energieverlening. He quickly finds out that the company is performing far below the average of comparable car dealerships. He found his first opportunity for big savings.

The manager decides to hire Ploos Energieverlening for advice. It turns out that relatively easy opportunities exist for energy saving. The Ploos analyst explores the building and browses
through the administration of the company. The problems become obvious.
The garage is opened from Monday to Saturday from 8:00-18:00. The showroom is open on the
same days but with different times, 9:00-21:00, due to the fact that most customers visit the
showroom in the evening. During the opening hours, the showroom has to be heated or cooled
in order to be comfortable for potential car buyers. The garage is heated during winter days
but is not cooled on warm days. The doors of the garage are opened frequently during the day,
which results in peaks in energy consumption. Also, the showroom is faced southwards and
has a facade that is almost entirely made of glass. As a result of that, the climate system of the
showroom is consuming a lot of energy during summer days for keeping the showroom at the
right temperature. This effect is increased because the facade is made up from single glass!
The total surface of the building is 3500 square meter.

The analyst completes his report and hands it over to the manager. The proposed measures
consist of replacing the climate system and replace the glass by double glass. Also, the south-
dward direction of the showroom makes it ideal for placing solar panels which would reduce
electricity consumption significantly.

By itself, it is proposed that the measures would save the company more than half of its original
energy budget. However, the investment costs are significant and the payback time is calcu-
lated to be multiple years. And there is another problem; the car dealership lacks financial
resources to invest.

Luckily, the analyst informs the manager about another service that is provided by Ploos En-
ergieverlening. It is called Zero Budget Sustainability and allows investing in energy saving
measures within the current energy budget. In this construction, Ploos will control the energy
budget and use it to pay energy suppliers. The budget can be kept at its original level, but can
also be lowered immediately to 80 percent until the energy saving measures have earned itself
back through lower energy consumption. After that, the budget is estimated to be 50 percent
of the original. So in this construction, the car dealer has advantage right after the start of the
ZBS project!

The manager is enthusiastic and points out that he is interested in applying ZBS to the com-
pany’s energy budget. He also reveals that the current financial position is such that external
financers are required to implement the proposed measures. Now, the company and Ploos have
to look for an external financer.

After the Ploos analyst has approached multiple financers, it turns out that an investment
company is interested to finance the energy saving measures as proposed. However, to con-
vince the management of the investment company, a detailed risk analysis is needed. Recent
research done by TU/e has showed that energy savings can be calculated with high accuracy
by applying data mining techniques. In this research project, it was shown that internal vari-
ables such as organization- and building characteristics combined with external variables, be-
ing weather measurements, are good predictors for energy consumption savings. If the exact
savings are known, the risk, interest and maturity of the construction can be calculated. A fur-
ther risk analysis can then be made, allowing the financer to judge the investment opportunity.

The techniques based on the TU/e research project are applied to the situation of the car dealer.
It turns out that the proposed energy saving measures are indeed a decent investment op-
portunity for the investment company. Soon after that, the contract involving the car dealer, Ploos
Energieverlening and the investment company is signed. Now, the process for implementing
the proposed energy consumption savings measures begins....

A.3 Data preparation

This section provides more insight in the process of data preparation. Several steps of data
preparation are described and examples of data transitions are given. First, outlier detection
is described.

A.3.1 Outlier detection

During the process of outlier detection, several different types of outliers were detected for elec-
tricity and gas consumption. They are described in this subsection.

Outliers are detected using the $\mu - 3\sigma$ and $\mu + 3\sigma$ rule. When this filter is applied to the data,
several different patterns of outliers were found. For electricity consumption, they are described
below.

Electricity consumption outliers

Electricity shows a less periodical pattern than gas consumption, which makes it more easy
to detect outliers using traditional methods such as $\mu - 3\sigma$. In Microsoft Excel, a filter was
applied to the dataset and several outliers were detected.

First, it became known that some of the consumption data have negative values. This is im-
possible given the regular consumption of the buildings. However, this particular building
is known for its spare capacity in case of a power supply breakdown. In case this system is
started, the power it produces can be more than the actual consumption of the building. As
a result of that, electricity is supplied back to the network, resulting in negative consumption
values. An example of negative values in the dataset is given in figure A.2.

Figure A.2: Electricity consumption for meter 8716879100003**** from 06-16-2015 to 06-19-
2015
Note that besides the negative values also some other consumption values appear that are below the threshold. The outlier detection rules are represented by the min and max lines; values below or above these lines are considered as outliers. Even if the absolute value of the negative consumption values would be taken, they can still be considered as outliers. No correction to these outliers has been applied in the dataset because it is plausible that the values displayed in figure A.2 are realistic.

Other outliers that have been found take the value of 0. A value of 0 does not have to be an outlier but in specific cases it can be considered as such. An example is given in figure A.3.

![Figure A.3: Electricity consumption for meter 8716879100003**** on 04-15-2014](image)

For most of the 0 values that have been found, the pattern is similar. First, a sharp drop in consumption followed by some values of 0, then some values above 0 but still far below the threshold. After some peaks, the consumption returns to normal values. A possible explanation is that the building is equipped with a flywheel, which results in the peaks as shown in figure A.3. The pattern does not lead to the necessity to remove the outlier values.

Sometimes, a drop in electricity consumption values is present on more than one day. In that case, a pattern is shown that repeats itself on multiple days that follow each other. A possible explanation is that a test has been executed on the spare system. This test is then repeated a few more times on consecutive days. An example is shown in figure A.4.
Finally, due to mismeasurement of meters on a particular day, the cumulative consumption for that day is given. An example is displayed in figure A.5. This is the only type of outlier in the dataset that has been corrected.

This can be corrected easily by taking the difference between two measurements as the consumption for that time period. The consumption values are then realistic. Some authors propose to delete outlier values and replace them by the median of the variable or a moving window (H. Liu, Shah, & Wei, 2004). This is not considered as a good solution for replacing the outliers. It is nice to illustrate what happens if this methodology would be applied, figure A.6 shows the result. Compare figure A.6 to figure A.4.
Gas consumption outliers

Extreme values for gas consumption are harder to detect because gas consumption shows a seasonal pattern. During the winter, high peaks in gas consumption are common just as longer periods without any gas consumption at all during summer days. It has been tried to apply a filter to gas consumption but this has not lead to satisfactory results. A regular pattern for gas consumption is displayed in figure A.7.

Note that the value of 0 is not an outlier. As can be obtained from figure A.7 many of the values that are considered as outliers might not be irregular values at all due to seasonal influences on gas consumption. The one exception that this meter has is displayed in figure A.8.
The figure shows some values of 0 for gas consumption and suddenly a peak which is strange for a day in the summer. However, this might be due to maintenance on the heating installation. In that case, temperature drops during the day and when the heating is turned on again, a peak in consumption might appear. There is no objective method to deal with these irregularities and therefore the outlier values are left in the dataset. Climate variables for the same day show that it was relatively cold, especially during the night. This might explain gas consumption. Values for climate parameters are shown in Table A.1.

Table A.1: Climate parameters for Eindhoven on 18-06-2014

<table>
<thead>
<tr>
<th>Time</th>
<th>0:00</th>
<th>2:00</th>
<th>4:00</th>
<th>6:00</th>
<th>8:00</th>
<th>10:00</th>
<th>12:00</th>
<th>14:00</th>
<th>16:00</th>
<th>18:00</th>
<th>20:00</th>
<th>22:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>11.4</td>
<td>9.7</td>
<td>7.4</td>
<td>13.5</td>
<td>15.5</td>
<td>18.3</td>
<td>18.3</td>
<td>19.8</td>
<td>20.9</td>
<td>18.8</td>
<td>15.6</td>
<td>12.6</td>
</tr>
<tr>
<td>Global radiation</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>78</td>
<td>96</td>
<td>189</td>
<td>133</td>
<td>213</td>
<td>175</td>
<td>64</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Relative humidity</td>
<td>89</td>
<td>92</td>
<td>98</td>
<td>82</td>
<td>77</td>
<td>68</td>
<td>70</td>
<td>66</td>
<td>59</td>
<td>65</td>
<td>75</td>
<td>85</td>
</tr>
</tbody>
</table>

Another example that illustrates how difficult it is to obtain outlier values for gas consumption is shown in Figure A.9. As a result of these difficulties, no corrections in gas consumption outliers have been made.
A.3.2 Binary and dummy variables

Table A.2 summarizes the energy saving measures that are implemented in the buildings of this dataset.

<table>
<thead>
<tr>
<th>Class</th>
<th>Subclass</th>
<th>Count</th>
<th>Buildings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lighting</td>
<td>LED</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Circuits</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Comfort</td>
<td>LBK</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Heating</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Settings</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Construction</td>
<td>Glass</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Devices</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table A.3 summarizes the energy labels of the buildings in the dataset.

<table>
<thead>
<tr>
<th>Energy label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
</tr>
<tr>
<td>Unknown</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.4 summarizes the building types in the dataset.

<table>
<thead>
<tr>
<th>Building type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datacenter</td>
<td>1</td>
</tr>
<tr>
<td>Office</td>
<td>9</td>
</tr>
<tr>
<td>Residence</td>
<td>1</td>
</tr>
</tbody>
</table>

A.4 The energy price

The energy price depends on multiple factors, some of them being influencable for the customer. A schematic overview of the structure of the energy price in the Netherlands is given in figure A.10.
As can be obtained from figure A.10, three main components make up the total energy price. The first part is supply. It contains the standing charge (a fixed cost to be connected to the network), supply cost and energy tax. This energy tax is decreasing for higher supply volumes. Network costs make up the second part of the energy price. It contains a standing charge, a capacity tariff and connection charges. Finally, there is a VAT (value added tax) over the total sum of energy costs.

A.5 Partial consumption meters

Figure A.11: Consumption meter structure of the apartment blocks
A.6 iVAT images distance measures

Figure A.12: iVAT images without normalization (left) and with normalization (right) for the standardized Euclidean distance measure

Figure A.13: iVAT images without normalization (left) and with normalization (right) for the manhattan distance measure

Figure A.14: iVAT images without normalization (left) and with normalization (right) for the Mikowski distance measure
References


