Master Degree Project

UNIVERSITY OF SKÖVDE

USING CASE-BASED REASONING FOR PREDICTING ENERGY USAGE

Master Degree Project in Information Fusion
Two years Level ECTS
Autumn term/Spring term 2012-2013

Johan Bjurén

Supervisor: Gunnar Mathiason
Examiner: Ronnie Johansson

Acknowledgement

As a master student on the computer science program at the University of Skövde, I have focused to gain deeper knowledge in the field of information fusion and prediction methods. This area has caught my focus as it is a quite broad field and deals with different uncertainty factors. The study was made on this subject because of my interest and as a personal challenge, as it was something that I had not done before. The study made me break new grounds in order to solve the problem in the study. As the problem of this study was researched I found a strong connection to the concepts of smart grid. This led to deeper research within the smart grid domain and how problems like the one in this study have been dealt with before.

I would like to thank Gunnar Mathiason, the supervisor of this study, for even though been busy; take the time to provide feedback, help to find new ways to look at the problem when things seemed to have no solutions and sharing knowledge on the subject when needed. Ronnie Johansson introducing the concepts of information fusion to me, Alexander Karlsson for the help with how to interpret the evaluation formulas and Anders Dahlbom for input regarding the method. I would also like to thank my girlfriend and family for the understanding during the process. I would also like to thank my friends for providing an oasis when I needed to take a break and think of anything else than just this work you know who you are. I would like to send a special thanks to Björn, my dad, for always backing me up. Henric Carlsson, which has been working on a master thesis in the same domain, but from the user perspective, for feedback, the endless discussions regarding the studies and everything else. It has been a real pleasure going through this time with you mate.

Abstract

In this study, the inability to in a future meet the electricity demand and the urge to change the consumption behavior considered. In a smart grid context there are several possible ways to do this. Means include ways to increase the consumer's awareness, add energy storages or build smarter homes which can control the appliances. To be able to implement these, indications on how the future consumption will be could be useful. Therefore we look further into how a framework for short-term consumption predictions can be created using electricity consumption data in relation to external factors. To do this a literature study is made to see what kind of methods that are relevant and which qualities is interesting to look at in order to choose a good prediction method. Case Based Reasoning seemed to be able to be suitable method. This method was examined further and built using relational databases. After this the method was tested and evaluated using datasets and evaluation methods CV, MBE and MAPE, which have previously been used in the domain of consumption prediction. The result was compared to the results of the winning methods in the ASHRAE competition. The CBR method was expected to perform better than what it did, and still not as good as the winning methods from the ASHRAE competition. The result showed that the CBR method can be used as a predictor and has potential to make good energy consumption predictions. and there is room for improvement in future studies.

Table of Content

1 Introduction 1
2 Background 3
2.1 Trends in the electrical distribution domain 3
2.2 What affects the consumption 4
2.3 The Smart Grid 5
2.4 Predicting consumption 6
3 Problem definition 9
3.1 Problem statement 10
3.2 Aim 11
3.3 Objectives 11
4 Method 13
4.1 Finding a prediction method 13
4.2 Develop the framework 14
4.3 Evaluating the accuracy of the framework 15
5 Results and analysis 16
5.1 Finding a prediction method 16
5.1.1 In the electrical domain 16
5.1.2 Outside of the electrical domain 17
5.1.3 Qualities that can contribute to a prediction framework 18
5.1.4 The method of Case-Based Resoning 18
5.1.5 Choice of prediction method 18
5.2 Approach 19
5.2.1 How does CBR work? 19
5.2.2 Measuring similarity in Case-Based Reasoning 20
5.2.3 Data preparation 21
5.2.4 Software 22
5.2.5 Implementation 22
5.3 Evaluation of using CBR for predicting electric consumption 28
5.3.1 Experiment 1 29
5.3.2 Experiment 2 31

5.4 Analysis....32
5.5 Summary of result....34
6 Discussion....37
7 Conclusion....40
8 Future work....43

1 Introduction

There is an emerging need of changing the consumption of electricity in the world. Only in India energy consumption is forecasted to have tripled by the year 2050 (Costanzo, Kheir & Zhu 2011). In Sweden 2011 the electrical consumption had a total of 129.2 TWh (losses not included). This was a drop in the energy consumption according to Anners and Enmalm (2012). This drop is primarily due to households lowering their consumption. 2011 the households in Sweden accounted for over 21% (33.6TWh) of the total energy consumption. An objective to decrease the energy usage by 20% has been set up (The European Commission 2008). They point out that "Main obstacles to energy efficiency improvements are the poor implementation of existing legislation, the lack of consumer awareness and the absence of adequate structures to trigger essential investments in and market uptake of energy efficient buildings, products and services." (The European Commission 2008).

A new law, 1997:857 (Sveriges riksdag 2012) was enacted in Sweden first of October 2012 which made it possible for all consumers to buy electricity per hour. This law increases the consumer's influence on consumption which can reduce their electricity costs by adjusting their habits to the current electrical price (Sveriges riksdag 2012).

The price of electricity is set by different parameters. The demand varies over the day and differ trough the different seasons (Zahedi 2011). One impact factor on the price is how much energy is produced and how big the demand is. If there is a low demand of energy and much produced the price drops, and if there is a high demand and little energy produced the price rises.

Smart grid is a concept which includes everyone from the producers to the consumers to address problems within the electrical domain to make the grid more intelligent, future-proof and safe. One branch in smart grids focuses on helping electrical consumers to be more aware of their consumption and what they can do to influence it. A suggested way is to help the users to be more aware of their consumption and stimulate them to shift their consumption away from the periods of peak demand, which means that the consumer becomes more environmentally friendly and can save money. On the other hand does Litos Strategic Communication (2008) state that normal users only want to just put a few hours per year to set their energy preferences. To enable a system like that the homes need to be smarter and be able to adjust consumption to timespans where the electricity price is lower. These smart homes will therefore be able to decide by itself when specific electric appliances should start. The problem is that it is not possible to develop a single model of how consumption works in every house since every household has different set ups of family members, insulation and other things that affects the consumption according to Riihimäki and Koponen (2012).

This study intends to create a framework which will be predicting electricity consumption. The framework uses external sources and maps these to the consumption to create predictions of the consumption. By utilizing information about the consumers' habits, in relation to external inputs such as price of the electricity or the weather, indications on how the consumption and peak demands can be forecasted. These electricity forecasts can be used to help consumers to take decisions on how they should consume electricity and, in a future, for scheduling appliances in an automated system.

In the first chapter, introduction, a short summation of the study and the reasons why the problem is investigated is presented. In chapter two, Background, the source of the problem is explained, what context the problem appears in and research around it. In chapter three the problem is defined. To solve the problem a number of objectives is set up. Chapter four explains which method is used and how it is used to find a solution to the problem, and the objectives to the problem. In chapter five the results of the study is presented and tested. The results of the tests are there after analyzed in order to see how high accuracy the method had. Chapter six brings up some related work where one can see the differences and similarities between this study and the studies presented. In chapter seven the study is discussed, how the method performed, what could have been done different and if the results are satisfying. Conclusions are presented in chapter 9. This chapter are first presenting a short summary of the work and there after presenting the findings in the study. The final chapter also brings up the future work that could be investigated.

2 Background

There is an emerging need of change of how electricity is consumed in the world. Only in India the consumption is forecasted to double by 2030 Prajapati, Ghadiali and Vora (2012) and according to Costanzo, Kheir and Zhu (2011) energy consumption is forecasted to have tripled by the year 2050. As stated in these reports the load on the grid increases and will, in a not that distant future, not be able to provide the consumers with their electrical demands. Several goals have been set up internationally to be able to meet these new requirements on the grid. In America the U.S. Department of Energy set up a goal for 2030 to handle the problem to supply their consumers with electricity (U.S. Department of Energy Office of Electricity Delivery & Energy Reliability 2010), EU has another goal to solve this problem set to the year 2020 (The European Commission 2008) and in Sweden similar goals have been set to the year 2020 (Energimyndigheten 2012).

2.1 Trends in the electrical distribution domain

Today's grid is based on techniques from when the electrical grid was built. As the load increases the grid will in several parts of the world, soon not be able to provide the consumers with their demands (Costanzo, Kheir & Zhu 2011). This is just one of many problems which exist in the grid today and there are ongoing research for different solutions and techniques to solve these.

Today, in Sweden, the demand often exceeds the production. According to (Energimyndigheten 2012) the demand from the consumers and expected production capacity together with political decisions among other things, factors that affects the price of the electricity. If the resellers have a surplus of agreements on electricity their price fall and the opposite if the reseller does not have enough of them the price will rise. When there is a limited access to electricity, often when the demand is high, creating a so-called peak demand, the prices rise. With higher price, the use of electricity often is reduced (Energimyndigheten 2011). This indicates that producers, suppliers and users of electricity are the ones who set the price via interaction of supply and demand.

In Sweden households can either have an agreement with an electrical reseller, which in their turn buy electricity directly from the electrical producers, or buy electricity per hour via Nord Pool's physical market Elspot[1] (Nordpool Spot) which offers different prices for different hours and is determined through interaction of supply and demand (Energimyndigheten 2011).

The price of electricity is affected by many things and not all of them can be foreseen. External events can also have an impact on the electrical price such as the nuclear disaster in Japan, Germany's decision to stop producing nuclear power or economic crisis in Greece. The price for carbon dioxide based emission rights also affects the price of electricity. The nuclear disaster in Japan led to an increased use of carbon dioxide based power sources such as natural gas and coal; this affected the electricity price to increase in Europe (Energimyndigheten 2011). Another factor is that the price for houses with electric heating has recently dropped while the overall prices for electricity, oil and natural gas has risen the last ten years. This has led to that the prices for carbon dioxide based electricity production has risen, which in the end affect the consumers.

[1] http://www.nordpoolspot.com/

2.2 What affects the consumption

According to Swedish Energy Agency (2011) are there several factors that affect the electrical consumption such as business structures, population changes, outdoor temperatures, economic development, technological development and development of energy prices. In a normal household, the consumer is the one who makes the decisions on which electrical appliances that should be used at what time. All appliances that use electricity in a household contribute to the overall electrical consumption. Some of the appliances consume more electricity than others and some consume electricity only when they are actively used while others also consume electricity passively such as the refrigerator or standby mode on the television.

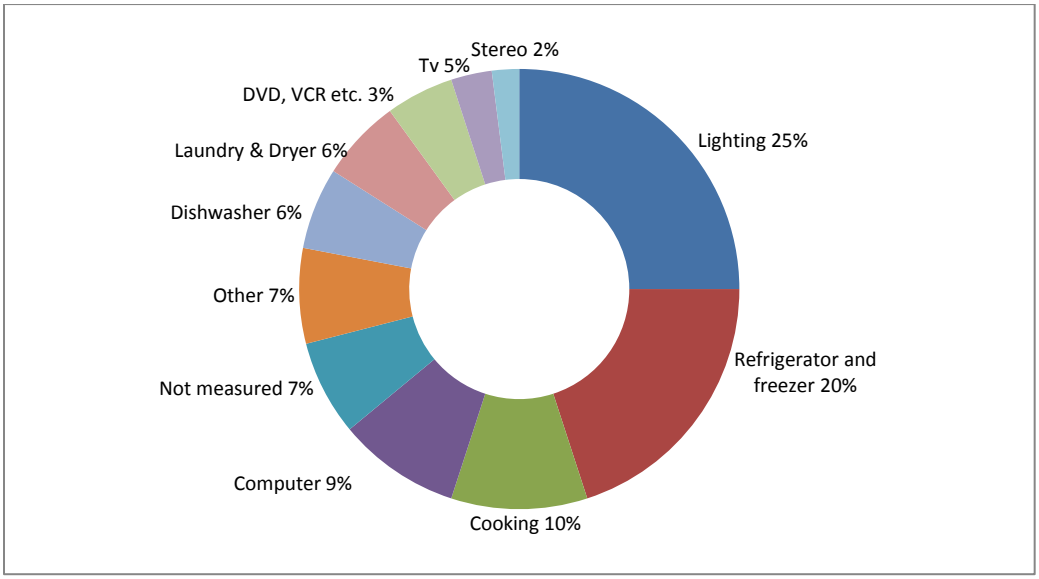

Fig. 1 Energy distribution 2007 (based on Karlsson (2011))

As seen in Fig. 1, in 2007 the largest consuming areas in a common household was Lightning, Refrigerator and freezer. Since then a lot of work has been done in several of these areas such as offering energy efficient LED light bulbs.

There are several things a household can do to affect their electrical expenses, for instance do something about their consumption and peaks in their consumption. To do this a household needs to have knowledge about what can be changed. According to Costanzo, Kheir and Zhu (2011) there are three categories of appliances. The first category is *Long-term scheduled tasks*. This is tasks that you cannot control but still need, like refrigerator, water heater or ventilation. Next category is *Short-term scheduled tasks*, this is things that are controlled by an operator and have a certain run time like dishwasher or washing machines. These appliances are consuming about 6% each of the total consumption in a common household according to Fig. 1 and could have a noticeable impact on the consumption for a normal user. The third category of appliances is *Unscheduled tasks* which are things you do in your everyday life like hair dryer, cooking or working by the computer. As seen here there are two categories that the consumer has control over, and one of them more than the other. This study focuses on the tasks that can be controlled. The Smart Grid is a concept in the electrical domain which deals with problems and innovations in this domain and is looking into this subject among other.

2.3 The Smart Grid

Smart grid, also known as intelligent grid, is a growing concept in the electricity domain. This concept is an "umbrella" term, meaning that it includes all parts that can affect the grid and consumption, from producer to end user. This term has branches that deal with many of the problem the electrical grid has, and will have in the future.

> "The smart grid is the collection of all technologies, concepts, topologies, and approaches that allow the silo hierarchies of generation, transmission, and distribution to be replaced with an end-to-end, organically intelligent, fully integrated environment where the business processes, objectives, and needs of all stakeholders are supported by the efficient exchange of data, services, and transactions" (Farhangi 2010, p. 23).

The European Technology Platform defines a smart grid to be an electrical network that integrates all the users, from consumers to producers in order to create a sustainable, secure and economic platform (European Commission 2005). The US Department of Energy also states that smart a grid should enable participation from consumers and adds that it is a self-healing grid which is resilient towards attacks on the grid from different sources (Litos Strategic Communication 2008). The Australian Government (Department of the Environment, Water, Heritage and the Arts) defines smart grid as a grid that with help of telecommunications and information technology applications consumers in home and business can improve their electricity efficiency (Australian Government: Department of Climate Change and Energy Efficiency 2010).

Smart grids should be able to make more efficient use of existing resources by the use of peak shaving, service quality control and demand-response (Farhangi 2010). By eliminating peak demand, the life of the grid equipment is extended (Prajapati, Ghadiali & Vora 2012) and the need of more power plants is reduced. Even if Smart grids are designed to meet the occasional peak demand and recharges when there is a non-peak, lowering the load on the grid is something that has to be considered further. That is why Smart grid also is focusing on reducing the load on the grid (Zahedi 2011) by helping users to make good decisions for their electricity consumption.

A household can do different things to lower its environment footprint and spend less on the electrical bill. As stated above, when there is a high demand and not that much electricity, i.e. a peak demand, the electrical price tends to go up (Energimyndigheten 2012). According to Samadi, et al. (2010) consumers are more willing to do something about their electrical expenses such as insulate their homes and change their behavior and schedule their electrical consumption to off-peak hours due to the recent increases in price of energy. Changing how the use of their *Unscheduled-* and *Short-term scheduled tasks* will affect the electrical consumption with a result of cheaper electrical bills. Smart grid makes the consumer interaction possible and empowers consumers to reduce their energy costs by adjusting their energy use (Farhangi 2010).

According to Khattak, Khan and Mahmud (2012) the level of awareness must increase in order to improve the consumers' decision making process on how and when they can lower their level of energy consumption. Different approaches have been studied to increase the awareness. In many countries meters for automatically acquire the meters value, also known as automated meter

reading (AMR), has been installed in households (Prajapati, Ghadiali & Vora 2012). These meters were originally used to get the correct amount of energy as foundation by electric suppliers for invoicing, but they also opened up the possibilities to let consumers know how much electricity that was used in their homes. By using these meters consumers could be alerted how they can change their consumption during a price peak to save money. There are indications that in the future, most of the electricity consuming products will be interconnected through an automated system that controls the equipment (ABB n.d.) and there is ongoing research on how energy prediction can be made in this kind of systems (Hawarah, Ploix & Jacomino 2010). Farhangi (2010) state that in the long term only having AMR is not enough, since it does not address the demand-side management issue. There is research directed towards advanced metering infrastructure (AMI), also known as *smart meters*, a two-way communication system. This system provides utilities with the ability to modify the service-level parameters. By using an AMI system there can be extended load management, instantaneous retrieval of information about the demand and the consumption (Farhangi 2010). By making good predictions on the energy usage the electrical suppliers could get a good indication on how much electricity they need to obtain to meet the need and how energy needs to be distributed. The development of automatic home appliances is an ongoing research area (Prajapati, Ghadiali & Vora 2012). Future appliances will have the feature to be set to run when specific conditions are fulfilled, like when the electricity cost is below a specified value, e.g. when the electric car should recharge its batteries (Ferreira, Pereira & Filipe 2011)

This means that it could be several benefits for a normal family. A family could cooperate and interact with the grid, and by utilizing the smart grid technologies changing how and when the electricity is consumed to save money by using it when the price is right (Logothetis 2012). A family could also produce electricity (by using solar panels or windmills etc.) and sell the surplus. According to Tang and Ciuciu (2012), after a consumer has used the energy needed of their produced electricity they could supply the grid with the leftovers and save between 5 and 15% on their monthly bill.

There are indications that consumers are ready to engage with the smart grid. Litos Strategic Communication (2008, p. 20) states "… what they [the consumers] will do is spend two hours per year to set their comfort, price and environmental preferences". What it means is that a consumer is not interested spending each day to update their preferences so they have some demands in return on the system, it should be simple, accessible, should not restrict the way they live their lives and letting them maintain a certain level of comfort. This indicates that the future home needs to be either automated and react on the user's specified settings, or being able to provide other support for simplification for the consumer, on how the home should consume electricity.

2.4 Predicting consumption

There is ongoing research on how to predict electric consumption. Most of the research on this topic has been made with industry and commercial buildings in focus where agreements based on hourly prices have been present (Edwards, New & Parker 2012). In these studies, different techniques have been used in order to be able to predict the future electrical consumption with high accuracy.

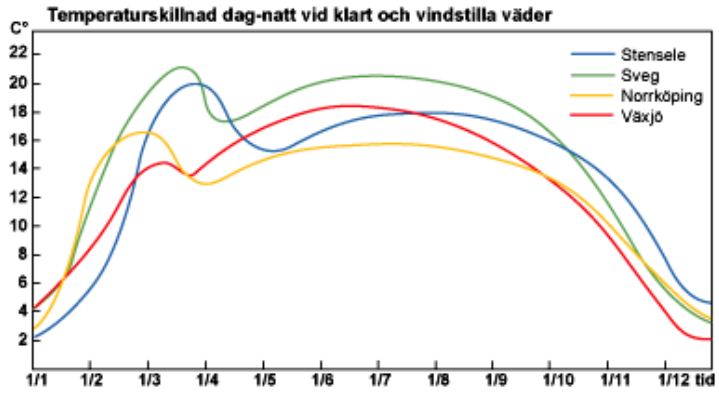

Fig. 2 Differences in temperature (SMHI 2009).

It has been proven in several reports that external factors affect the electrical usage. Riihimäki and Koponen (2012) conclude in their study that there are relations between the electrical consumption and the outdoor temperature. According to Riihimäki and Koponen (2012) the distribution network operators also use the average temperature to predict the energy demand for the next day.

As Swedish Energy Agency (2011) states, there are several factors that affect the electrical consumption like population changes, outdoor temperatures, energy prices among other. These factors do not affect every home in the same manner. Factors such as size of the house, insulation, comfort temperature and heating method also have an impact on the electrical consumption. According to Riihimäki and Koponen (2012), different households can differ in the consumption depending on these structural factors. In Finland where the study was made, the temperature can differ over the day and as the weather and temperatures varies over the year the difference in temperature differs over the year. This is also the case in Sweden as can be seen in Fig. 2, where there is a larger temperature difference between day and night on the summer than the winter. This difference can change depending on the weather. This makes it impossible to have one prediction model that fits every house.

There has been research done in the area of predicting electricity consumption. In these studies has industry and commercial buildings been in focus (Edwards, New & Parker 2012). American Society of Heating, Refrigerating and Air-Conditioning Engineers[2] (ASHRAE) hosted a competition where more than 150 competitors participated to investigate how to predict electrical consumption (Kreider & Haberl 1994).

Several prediction methods, and variations of them, were tested in this competition. A feedforward neural network was used in one of the winning methods by MacKay (1994). MacKay (1994) sorted out irrelevant factors, using a method called automatic relevance determination, and put weights on the different factors to see which factors affected the result most. Figueiredo et al. (2005) made use of techniques like Self-Organizing Maps, cluster compactness and measure of cluster separation to create a system that could be used to predict electricity consumption. Self-Organizing Maps, which is a artificial neural network trained by

[2] https://www.ashrae.org

unsupervised learning. By assign consumers to different “classes” of load profiles, made by using historical data, electrical consumptiopn could be prediced.

Chen, Das and Cook (2010) used the naïve Bayes classifier, Bayes belief networks, artificial neural networks and a Support Vector Machine to learn and recognize complex patterns from sensor-data in order to predict energy consumption.

3 Problem definition

There are several benefits of helping the consumers in the domestic sector on the subject how they use electricity. Smart grids are a concept which includes all technologies regarding producers, the grid and electricity consumers. In the field of Smart grids there are ongoing research in many areas, one area concerns technologies for helping energy users to make decisions that can decrease the environmental footprint and at the same time lower their electricity cost.

Recently, a new law was introduced in Sweden which let all electrical users to buy electricity per hour. The law was introduced to encourage users in the domestic sector to increase the awareness and consume electricity when it is best for the grid and environment. Electricity price is set by how much energy that is produced versus how large the demand of electricity there is.

The information on the electricity cost could be used to benefit the consumer. A future home is expected to have several possibilities to manage the electrical consumption and there are several ongoing projects with homes that can automatically adjust and adapt the consumption to the price. These systems need to be simple for the consumer to use, without the need of micromanagement, and should not restrict the consumer's life.

Many problems in the smart grid domain are still unresolved and different techniques to solve these are researched. Predicting consumption is one of the unresolved issues, which can be used in several areas to increase the effectiveness of energy usage, such as increasing consumer awareness of how the energy consumption will be tomorrow. This study investigates how predictions can be made on electricity consumption. There are several factors which are affecting the consumption such as outdoor temperatures, economic and technological development, development of energy prices etc. This study relates these factors to the consumption to be able to make consumption predictions. These predictions have a great potential for saving electricity and lowering the environmental impact.

This topic is still new and there is still little research in how users' consumption can be related to external factors, such as weather, price etc. There are reports where the relation between the external factors and the consumption is proven such as Riihimäki and Koponen (2012). As these factors can be measured and forecasted with a good accuracy with today's measuring techniques, electricity consumption predictions with a good accuracy can be done using these.

There is ongoing research on how energy consumption could be forecasted. A future home is therefore predicted to be a home which can adjust to certain conditions and adapt to the user's behavior, based on a few overall, user specified, settings.

To be able to deal with the problem of electricity shortage, where the consumption exceeds the production, reports indicate that one solution can be that consumers need to be able to make better decisions on when and how electricity should be consumed. There are several ways a consumer could be helped to deal with this problem. Some reports indicate that it is the users' behavior that needs to change and to accomplish this users awareness must increase. Other reports suggest that the users does not want to change their way of living and that automatically adapt the energy usage should be used instead.

The consumption can differ a lot between households. Factors such as size, insulation, comfort temperature and heating method among others, also have an impact on the electrical consumption. This fact makes it difficult to design a model which is exact enough to capture the usage of a generic consumer, and therefore the need of a prediction method that is unique for each user arises. This study will make use of consumption data gathered on a household level and is based on personal data and therefore is unique for each household. This means that predictions made is personal and will be shaped by the characteristics of the target. This means that the predictions will have little probability to be applicable on any other household. The predictions will be related to other, external, factors that affect or might affect the outcome.

The consumption problem this study is focusing on is therefore the ability to make consumption predictions in the electrical domain that are based on sensor data, such as temperature, wind and the electrical consumption. The consumption is based on a single household and the preferences that come with this household.

A framework, defined in this study as a model with guidelines for supporting a set of procedures, can be created that uses knowledge of how external data relates to electrical consumption in order to forecast electrical consumption. To create such framework a set of procedures, such as input, predicting algorithm, output and other data processing, needs to be examined and defined to acquire a result that produces a satisfying result since there is a need for improved and more accurate energy consumption predictions.

The predictions achieved in this study could be used in several different ways to lower the energy usage, especially when there already is a high load on the grid. The predictions could be used to get a short-term prediction on the electricity consumption in a household. The consumers could use the information to draw conclusions based on the prediction on what the outcome would be if they did something unexpected to the system. To do this the forecasts could be implemented into a recommender system which can present the options in a way that not only expert users can understand.

3.1 Problem statement

Predicting electricity consumption is difficult, especially in the domestic sector where there are several uncertainty factors. There are reports where algorithms and approaches have been tested to fulfill the prediction of electricity consumption. As the predictions have a high uncertainty, there is much to gain on finding more accurate and suitable prediction method for the domestic sector. This study is focusing on finding a method for the domestic sector as this is where the users just have been given the tools to affect their consumption. This study uses external sensor data for the consumption prediction. The focus lies on making short-term predictions. The prediction should be made with a single household in mind since different households have different preconditions.

? How can a framework for short-term consumption predictions be created using electricity consumption data in relation to external factors to aid consumers in the domestic sector with helping energy users to make decisions decrease the environmental footprint and lower their electricity cost in a smart grid context for standalone households?

3.2 Aim

This study aims to create a framework with high accuracy for predicting electricity consumption in relation to real electricity outcome on a short-term basis, using external sensor data as indication sources. The result of the prediction can be used to create meaningful knowledge for electrical consumers. To be able to create predictions within these frames a number of objectives need to be fulfilled.

3.3 Objectives

Edwards, New & Parker (2012) states that that the focus of most of studies is made on consumption prediction in the commercial sector. The domestic sector is still in need of more research on consumption prediction. There are several reasons why this is the case, such as there has not been any benefit for the consumer to adjust consumption to a times when the load on the grid is low. In the commercial sector, several studies have been made on the electricity consumption prediction as they could have agreements based on hourly prices. In Sweden this has recently changed and all electrical consumers can now make use of electricity based on market price (Sveriges riksdag 2012).

To be able to reach the aim, three objectives needs to be fulfilled:

1. Study literature to find relevant and suitable prediction methods and what has been used before to find new promising approaches to the problem. This will be used in the initial stages to understand the domain and the needs of it. Literature will also be studied to gather knowledge of different approaches, if they have been used in similar domains and how they should be performed. This knowledge is beneficial to acquire in an early stage in the process. When it is time to find the algorithm which is used to predict knowledge such as, how these approaches previously have been used and how they can be evaluated, will be valuable. The literature study also provides which external factors that can influence the electric consumption. Further the literature study will be used to formulate the problem and the investigation of related work.

2. Propose a new approach based on the literature findings for the domain. A suitable prediction algorithm is developed with the literature study as a foundation. Investigate how this approach is implemented and if there is any variables that need to be considered. Investigate if there is any existing software that can be utilized that is consistent with the specifications, or can be altered to fit the proposed solution. Furthermore, consider if data preparation is a necessity. If the data has to be prepared in order to use the suggested method this has to be considered and examined.

3. Evaluate the prediction accuracy and compare with alternative methods for such predictions. Measure the accuracy of the proposed approach and compare them with other prediction methods previously developed. The comparison use the same consumption data to be able find if and how much the prediction accuracy differs.

4 Method

This chapter outlines the work process on how the framework is created and how the work on acquiring information on the components that is used in the framework.

The framework will be using a prediction method which can handle input in form of sensor data to create an output in form of a prediction of the energy consumption. The sensor data produces an input that affects, or might affect, the outcome. The chosen inputs should be able to be predicted in forehand. By using the forecasted incoming data, predictions on the outcome can be made using this framework.

4.1 Finding a prediction method

Literature is studied to gain knowledge on prediction methods. This literature study was used to find important concepts and relevant keywords. To find relevant literature, different databases are used to gather information. General search engines for academic literature were utilized to find methods for prediction, which have or have not been applied in the electrical domain. Academic search engines, specialized in information on data management will be used to find information, which has either relevance for the electrical domain or data management in form of prediction techniques, or both. Databases specialized in data management will be used to provide knowledge and alternatives on methods which utilize external sensor data.

In order to find a prediction method for the framework, a few objectives were chosen:

- **Method qualities**
  Study the literature with the aim of finding prediction methods that have been used in the energy usage domain. Find what qualities a method should have to make good predictions and contribute to the system in a desirable way. To achieve this, reports are used to collect information on what kind of system features that are desired for the electrical domain. In particular studying literature on methods used to predict consumption to gather information on what distinguishes the tested method and what makes this predicting method a good solution in the area of consumption prediction.

- **Used methods**
  Find methods that have been used to predict electricity consumption in the electrical domain. Knowledge is acquired on what kind of prediction techniques, which makes use of external sources to make predictions, have been used in the electrical domain or in the smart grid domain. Techniques used in the domestic sector are of special interest. These methods are especially interesting as they might deal with the problem or a related problem specified in this study. In the process of finding prediction methods that have already been tested in the electrical domain, keywords from the literature study are used. Keywords, such as smart grid, electric, energy, awareness and peak demand is used in different inflections and combined with common prediction terms such as predicting, profiling, probability and estimation to get relevant results. Other relevant studies are found by examining the references on the found papers. The result is read through to gain knowledge on what research have been done and provide information on the methods that has been used in the domain.

- **Methods in other domains**
  Find methods that have been used to predict consumption using external sources in other domains. Prediction methods are searched for in order to find if there may be other suitable prediction methods that could be applicable for predicting electricity consumption in the domestic sector. The desired features are used here as a foundation on which methods that may be suitable for predicting electricity outcome. Methods yet untested in the electrical domain is of interest and what they can contribute with. In this part of the study is the domain of information fusion (IF) considered since IF methods are commonly used to make predictions with sensor data as a foundation.

These three objectives are expected to lead to a possible solution to the problem in this study:

- Deeper knowledge on how a suitable prediction method works. By studying literature, knowledge is gained on how the different predicting methods are used and details, such as if there is just one or if it is several underlying algorithms that needs to be considered and how these algorithms work. This will influence how our method is developed.
- Deeper knowledge on how the prediction methods can be evaluated. Study the literature and learn how the methods and algorithms have been evaluated before in this domain in scientific literature.
- Deeper knowledge on the limitations and benefits of the context the methods are used in. The method should be examined to find out how it previously has been implemented and used in order to be able to choose the method and underlying algorithm. Study the literature to gain knowledge on how the methods can utilize sensor data. Is there any need of data preparation?
    - If there is a need of data preparation, literature is studied to gain knowledge how to prepare data to make use of the prediction method. Literature is also studied to gain knowledge on how data has been prepared in this domain before.

The literature study should be able to provide one or two promising prediction methods which are yet untested in the electrical domain. These methods should be able to provide methods that can be used as it is or as inspiration for inventing a method. This or these methods will thereafter be studied further in the development of the framework.

4.2 Develop the framework

Knowledge gathered from the literature study to develop the framework. In this step the chosen prediction methods are studied further in order to be used in the framework.

- Decide how the framework should function. Set up directions for how data input and data output should be handled in the framework. Decide what inputs should be added. How these in- and outputs should be formatted.
- Find out how sensor data can be gathered. Search for sources on how sensor data can be acquired. Check if it is possible to gather sensor data from different sources such as

SMHI[3] or YR[4] which provides weather information. Investigate in what format the data can be delivered and which resolution it can be delivered in.

- Find out if there is existing software or libraries that can be used to test the method. If no software or libraries are found; software is built that is applicable to the method and are able to apply to the methods processes.
- Test the method using representative data to see if the method works with data from the electrical domain and chosen sensor data. Prepare the data in advance if this is needed.

This step is finalized by an implementation of the algorithms.

4.3 Evaluating the accuracy of the framework

The method's accuracy is evaluated by comparing accuracy of the framework with the accuracy of other comparable prediction methods. The following steps describe how this process was executed:

- **Consumption preparation**
  To be able to compare two or more methods they first had to be studied to gain knowledge on how they work and what kind of inputs and outputs could be expected. Second, to get a result that can be compared, both prediction methods need to be using the same dataset. The dataset contain real consumption data from households representing usage in the domestic sector. This means that the methods need to be tested on data used in this work or if this study can utilize the data used in the different studies made using the other prediction methods. Data also might have to be prepared to function with the methods and the underlying algorithms.
    - If other data is used, access must be gained to the dataset used by the other studies. The dataset must also be adjusted to this study.
- **Test the methods or algorithms**
  Use the prepared data together with the method of this study. Test is made by making use of the acquired software, to get the accuracy using this method. If the method utilizes different algorithms, the accuracy is tested by switching algorithms. The accuracy of other methods are tested as well in order to explore their perferences and abilities. The accuracy is calculated by taking the predicted electric consumption and compare with the real electrical outcome by utilizing the evaluation method found in the literature study. This will produce an accuracy of the prediction between 0 and 1. The accuracy of the methods is used in the evaluation. What qualities, in areas such as resource usage and implementation, are also taken in to consideration.
- **Evaluate the methods or algorithms**
  The test of the methods will generate a result between 0 and 1. By comparing the accuracy of the methods a result of which performed best, i.e. that had the highest accuracy, can be acquired.

[3] http://www.smhi.se/
[4] http://www.yr.no/

5 Results and analysis

In this chapter the results are presented from the literature study. Tested and untested methods are analyzed and presented together with what qualities they have which make them a good method. A promising method was found called Case Based Reasoning, which was yet untested in the electrical domain. This method was closely examined to find out how it could be adjusted to fit this problem and what pre-requirements there were. A system was built using databases and the language SQL. Thereafter the method was tested using data from the commercial and the domestic sector.

5.1 Finding a prediction method

The literature study shows that there has been "research" done on predicting energy consumption. Most of the studies have been made with industry and commercial buildings in focus and there is a lack of research done on prediction on hourly or daily basis the domestic sector. In the commercial domain electricity agreements based on hourly prices have been present for a long time (Edwards, New & Parker 2012). The authors discuss in their paper if it is possible to transfer the methods tested in the commercial sector to the domestic sector since the consumption and settings can vary significantly between them.

5.1.1 In the electrical domain

In the competition hosted by ASHRAE there where more than 150 participating competitors who investigated electrical consumption could be predicted (Kreider & Haberl 1994). The competition's goal was to find a method that could make as accurate prediction on the consumption as possible. The project was set up so that every competitor had access to the consumption and a number of external factors. The energy consumption was gathered from a commercial building and the goal was to be able to make predictions on the consumption based on the external factors. One of the popular methods was to make use of feed forward networks to predict the consumption. MacKay (1994) was one of the winners which used a feedforward neural network to predict the electricity consumption. In addition to the feedforward network the author made use of a method to sort out irrelevant factors called automatic relevance determination and put weights on the different factors to see which factors affected the result most. Therefore, it can learn patterns over time and the result is specialized for a specific consumer.

In the ASHRAE competition, the use of Support Vector Machines is also utilized by Thurtel and Feuston (1994) to predict monthly consumption. As seen just from these reports many techniques have been tested and a popular method is to make use of neural networks to predict the consumption.

Chen, Das and Cook (2010) participated in the ASHRAE competition and investigated how the activity in a household affects the energy consumption. They used the naïve Bayes classifier, Bayes belief networks, artificial neural networks and a Support Vector Machine to predict energy consumption. Bayes belief networks performs well in their tests and is one of the top prediction methods in most results. Chen, Das and Cook (2010) found that machine learning techniques can learn and recognize complex patterns from sensors.

In the domain of electricity consumption it is common to use data mining and data modeling methods according to Dent, Aickelin and Rodden (2011) and Do Thi Kim (2011). In Taiwan a system was built to classify different customers depending on which load profile, representing the carasterstics of consumption, they had. They described how Chi-square could be used to identify bad data. This system is used for system planning, load management, maintenance among other things. The load profile predictions are created by using statistical analysis (Chen et al. 1999).

Ramos et al. (2006) used the Knowledge Discovery in Databases (KDD) process, which includes pre-processing, postprocessing and datamining, to create load profiles and classification. In their report they created load profiles with different load curves for each cluster and assigned consumers to these profiles by using classification rules. This algoritm created noticeably different load-profiles and the classification algorithm performed the task with a good overall accuracy.

Figueiredo et al. (2005) created a framework by combining unsupervised and supervised learning techniques. They used techniques like Self-Organizing Maps, cluster compactness and measure of cluster separation. With these techniques they created a set of consumer classes of load profiles by using historical data and assigned consumers to these classes. The characterization was updated as new data was collected. Doing this ensures that the classifications are up to date. Figueiredo et al. (2005) means that these kinds of forecasts are of importance these days due to the competitive retail markets.

Hawarah, Ploix and Jacomino (2010) outlines a system that is using Bayesian networks to predict electricity consumption by trying to predict user behavior. A Bayesian Network is represented by a graphical model, a so called directed acyclic graph, which can be used to model probability relationship between different variables (Hawarah, Ploix & Jacomino 2010). The nodes represent the different variables and the arcs, representing the relationship, provide the probabilistic correlation and can be used to get probabilities for a certain thing to happen given how and if relating factors is set. They concluded the report with that they had belief in this system and what it could achieve when it is going to be created.

5.1.2 Outside of the electrical domain

The literature study shows that there are several techniques that are common outside of the electrical domain but have been used in some form in the domain as well, such as Bayesian Networks. Case Based Reasoning is another method that has not been used in this the electrical domain, according to the findings in this literature study. This is a problem solving methodology (Watson 1999) and has been tested and used as a prediction method before (van Setten et al. 2004). Schiaffino & Amandi (2000) used this method to predict users' behavior. By creating cases from experienced situations and reuse those as knowledge bank to compare the current situation to in order to get an indication on the outcome. The more cases collected the more precise the predicted outcome became. van Setten et al. (2004) also adds that CBR has the advantage of not being in need of expert knowledge to make predictions since the system learns itself over time. This means that the user does not have to spend time to set up and gather required data to set up a new system.

5.1.3 Qualities that can contribute to a prediction framework

There are several aspects and functions of prediction methods that are desirable. From the information in chapter 5.1.1 and 5.1.2, factors were found which seems to have contributing factors to the results in terms of more accurate predictions. The factors which lead to more accurate predictions were factors such as the way Ramos et al. (2006) assign customers to a certain load profile, or as Figueiredo et al. (2005) creates a system that can keep itself updated by adjusting as more data is aqured and how MacKay (1994) create a system which adjusts the relevance of the external factors over time.

5.1.4 The method of Case-Based Resoning

Since there are several types of CBR systems and they all follow a set of rules, Watson (1999) suggests that CBR is a problem solving methodology not a method. The background of CBR was the wish to understand how information was recalled by humans. This wish led to the insight that problems are usually resolved by remembering how a similar previous problem is solved (Watson 1999). CBR uses *cases*, which is knowledge of previously experienced problem situations described using data and information of the situation, to solve a new problem (Schiaffino & Amandi 2000). The case normally contains information about the problem that needs to be solved and a description of the solution. The case is stored and when a new problem occurs, this case or a small set of cases is used. Decisions are being made by comparisons between the new situation, the problem, and old situations, the cases, (Schiaffino & Amandi 2000). CBR offers a basis for intelligent systems since it can both solve problems and adapt to new situations (Slade 1991). CBR is implemented as a machine learning technique which can learn from data and as Chen, Das and Cook (2010) state, machine learning techniques can both learn and recognize complex patterns from sensor data which in this study will be represented by data such as: consumption, price or the weather, which are factors that influence the consumption behavior.

5.1.5 Choice of prediction method

The system can learn over time and set which attributes that are relevant for a specific case by using weights. If an attribute has a strong impact on the result this attribute should have a higher affection on the predicted outcome. This is close to what MacKay (1994) did in the electrical domain where he used a neural network together with an automatic relevance determination system to decide which factors that is relevant. By using CBR with weights, relevance for each attribute can be learnt over time for all cases. By implementing weights, an attribute, that is very crucial for one outcome, can be amplified and an attribute that have not any significance for another outcome could be reduced. The results from a CBR system might not be a perfect match since the outcome has to be an already experienced situation, stored in the CBR system's casebase. Ramos et al. (2006) uses load curves and assigns these to users. Data can be added in to a case based- system in several ways. One way to limit the cases in the casebase is to find cases in the casebase similar to the acquired input. The input is assigned a case if the parameters are within a certain pre-defined threshold. If the parameters exceed the threshold a new case is created. Just as the system created by Figueiredo et al. (2005), this system will also adapt as more cases is collected. The case definitions is updated and new possible results, in form of new cases is created over time by collecting more cases.

In this study, we have found methods, used both in and outside the electrical domain that have provided us with knowledge of different methods and their benefits. The CBR method was found to be both promising and competitive. The method is not unique in its features, but it can comply with several of the desired features that can be found in other prediction methods. The method has not been found in the literature as used as a prediction method in the electrical domain. This Method, together with the desired features, makes an interesting method to test and compare how it performs in relation to other methods.

The method benefit from qualities such as:

- Be able to reuse old data which is specific for the context.
- Can be used with weights to change the importance of the attributes.
- Can adapt to change over time.

This makes this method to an interesting candidate for predicting energy consumption. This study will therefore use CBR for predicting user electrical consumption.

5.2 Approach

In this chapter, we intend to gain deeper knowledge of how the method works in general, how it can be implemented and of and how there are any pre-requirements when this method should be implemented. As (Watson 1999) states, CBR is a problem solving methodology which make use of cases in order to solve new problems (Schiaffino & Amandi 2000). In this chapter, CBR is examined more closely to gain deeper knowledge and find out how this method could be acquired and adapted to fit the consumption prediction problem.

5.2.1 How does CBR work?

Watson (1999) states that CBR is a methodology, not a technology. This methodology must fulfill four different activities, how these activities should be fulfilled is not specified which means that one CBR system can differ from another,

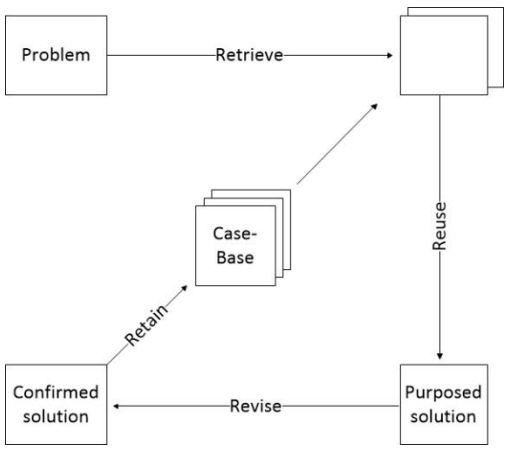

Fig. 3 CBR-cycle (based on Watson (1999))

- *Retrieve* cases that are similar to the problem.

- *Reuse* a solution that has been suggested by measuring the similarity.
- *Revise or adapt* the solution to fit the new problem.
- *Retain* the new solution when it has been confirmed.

Watson (1999) argues that systems that solves problems by reusing old problems in form of saved cases, retrieves these cases by comparing similarity. As soon a problem is solved it is added to the case library for further use as a solution in form of a case is a CBR system. CBR is normally described using the *CBR-cycle*, Fig. 3, which contains four activities. The first activity is to (1) retrieve cases that are similar to the problem; next (2) The solution from the most similar case is suggested. After this, (3) the solution can be revised, if it is necessary, to fit the new problem better. (4) As soon as the new solution is accepted it is saved for the future.

According to Schiaffino & Amandi (2000) are cases used to provide information for detecting patterns. Every case has three main parts, a description of the situation or problem, the solution and the results when the solution is applied.

5.2.2 Measuring similarity in Case-Based Reasoning

There are several techniques to measure the similarity which has been tested and revised in the literature. By comparing the similarity of the attributes in the problem to each representing attribute in the stored cases a similarity can be measured. This similarity measure can be multiplied to a weighting factor and then be summarized to get a measure of similarity between the problem and each case where weights can affect attributes should have more influence than others (Watson 1999). van Setten et al. (2004) states that if a case has high similarity to the measured situation, this case should have a higher importance in determining the outcome for the request and the importance could be used as a weight.

One common way to calculate the similarity between a problem and a case is to calculate the distance between the attributes. *Nearest Neighbor* is commonly used to make a calculation like this. Nearest Neighbor can be calculated in several ways.

van Setten et al. (2004) argue that unweighted Euclidean distance (UED), weighted Euclidean distance (WED) and maximum measure (MMS) also are frequently used as distance measures in CBR.

UED measures the Euclidean distance between two points on a matrix, (x_0, x_1) and (y_0, y_1), where each dimension represents an attribute in the case (Mendes, Watson & Mosley 2002). WED works in the same way as UED but with the use of weights representing the importance of an attribute. Mendes, Watson and Mosley (2002) states that even if the weights were not randomly picked in their work deciding the weights is still a research question which needs to be investigated further. MMS calculates only the attribute that has the highest similarity to the current situation (Mendes, Watson & Mosley 2002).

Mendes, Watson and Mosley (2002) found that using different configurations generated different accuracy when they tested UED, WED and MMS. According to their test, WED was performing best while MMS had the worst prediction accuracy. The authors also tested the performance using different analogies, the number of cases used to make the prediction. They

found that using *closest analogy* worked best on one dataset. By using the mean (the average of the analogies when there is more than one) of the three closest analogies worked best only for a certain type of dataset. In the report, the authors argue that the reason why *closest analogy* is better on one dataset is because that the dataset it performed better on contained more data and the data was more homogenous.

van Setten et al. (2004) is also testing these different techniques, including UED, WED and MMS among others. Their report showed that UED performed best on every test they made on different data and outperformed rule-based prediction strategies that where manually created.

Nearest Neighbor, in UED form, can be represented by Similarity (T, S) where T is representing the target, S the source. One common way to do this equation is just to compare how near one the source is to the target by using the equation:

$$Similarity(T,S) = \sqrt{\sum_{i=1}^{n}(T_i - S_i)^2}$$

In this equation i represents the attribute in the case and source, such as consumption or outdoor temperature. The attribute (i) has a minimum of 1 and can reach to an infinite number.

The attributes can also be mapped using a Bayesian Network (Schiaffino & Amandi 2000). By using the strength of each relationship in a Bayesian Network to map the attributes, the attributes get automatically weighted. These attributes can thereafter be updated like a Bayesian network, this increases gradually as new information is acquired. The more information acquired, the more cases are created and the more precise the system will be.

5.2.3 Data preparation

When using a machine learning technique such as CBR, the input data needs to be discretized. Therefore it is important to find a discretizing method which is fitting for the specified task. According to Liu et al. (2002) is discretization the process of quantizing continuous attributes. Through discretization, data can be reduced and simplified and results are often more accurate, shorter and more compact then continuous data. By discretizing the data you will lose the ability to work with high resolution data and real-time data, which in some cases is needed. According to Chmielewski and Grzymala-Busse (1996) is finding a way to discretize a *continuous-time process* into a *discrete-time process* the first thing to do when working with inductive learning methods. According to Liu et al. (2002) is learning often more accurate and faster with a result that is compact, shorter and more accurate by using discrete data in comparison to continuous data. Liu et al. (2002) also points out that learning algorithms is only able to make use of discrete data. There are many advantages of having discrete data over continuous data, not only that learning algorithms is in need of it, probably a more accurate and a more easily understood result is produced.

Methods	Global/ Local	Supervised/ Unsupervised	Direct /Incremental	Splitting/ Merging	Static/ Dynamic
Equal-Width	Global	Unsupervised	Direct	Splitting	Static
Equal-Frequency	Global	Unsupervised	Direct	Splitting	Static
1R	Global	Supervised	Incremental	Splitting	Static
D2	Local	Supervised	Incremental	Splitting	Static
Entropy (MDLP)	Local	Supervised	Incremental	Splitting	Static
Mantaras	Local	Supervised	Incremental	Splitting	Static
ID3	Local	Supervised	Direct	Splitting	Dynamic
Zeta	Global	Supervised	Direct	Splitting	Static
Accuracy	Global	Supervised	Incremental	Splitting	Static
ChiMerge	Global	Supervised	Incremental	Merging	Static
Chi2	Global	Supervised	Incremental	Merging	Static
ConMerge	Global	Supervised	Incremental	Merging	Static

Table 1. Discretization methods

Different discretization methods have different way of performing the task and there are a few aspects which must be considered before performing a discretization. These aspects which differs the methods in general are specified for each of the discretization method in Table 1. Table 1 is based on the findings from Liu et al. (2002). A discretization method is chosen based on how it work and its properties.

5.2.4 Software

There are several CBR applications described in Watson (1999), some of them cannot be found anymore since the company does not exist any longer and also have others pre-requirements which cannot be met in this study. According to Watson (1999) it is possible to develop a CBR system by using SQL. SQL, Structured Query Language, is a programming language designed for managing data in a relational database. Databases can be used to measure similarity between different cases and attributes in these cases. Watson (1999) also states that databases have an advantage when working with large amounts of data. Working directly with databases also has the advantage that most data is stored using databases.

In a system that predicts electricity usage based on external parameters needs to be able to handle large amounts of data. Using SQL, stored procedures can be created to perform more complex queries.

5.2.5 Implementation

The system: Our CBR system is implemented using MySQL to store and make calculations. By using *stored procedures*, which are functions that can perform one or several calculations, using that, more complex calculations and data processing can be made. Using databases, such as MySQL, also makes it easier to implement our approach and integrate with other systems. To execute the SQL Query's the MySQL Workbench is used.

```
Begin
    add the case label into case-index;
    for each attribute(i)-value(i) pair in the case
        get the list of case labels associated with
        the attribute(i) in indexing-list;
     if the list is empty
        add the attribute(i) into indexing list;
              create a list containing the case label;
        else
              associate the case label into the list of
             case labels
End
```

Fig. 4 Pseudo code example by Maher, Zhang and Balachandran (1995)

Maher, Zhang and Balachandran (1995) outline how CBR queries can be structured and provide pseudo code. This can be altered and implemented using stored procedures in SQL. By following the outline made by Maher, Zhang and Balachandran (1995), as seen in Fig. 3, a CBR system is implemented.

Maher, Zhang and Balachandran (1995) do not address that the results can be weighted. In some reports weights have been used to produce better results (Mendes, Watson & Mosley 2002) and in other not (van Setten et al. 2004). Mendes, Watson and Mosley (2002) address the topic on how to decide the weights and conclude that this is something that is hard to decide and more research within this area is needed in order to create good weights. Since no data is given on how many the different factors affects the result, no weights is set and the system will be built around an unweighted comparison between the problem and the cases which can outperform weighted measures (van Setten et al. 2004).

According to Schiaffino & Amandi (2000) should a case have three main parts. A description of the problem(1), the solution(2) and the results when the solution is applied(3). When cases are added to the case-base, the attributes are automatically created in the system we create. These attributes can be flagged as solution or part of the solution. The rest of the case is describing the setting for the case. This way to structure the case allows flexibility and it fulfills the criteria to have problem, solution and results specified.

- The implementation is using CBR accordingly to the outline by Watson (1999) and Schiaffino & Amandi (2000).
- The CBR system is implemented using SQL, which eases the implementation in existing system where relational databases are used.
- Does not implement weights since it is still unclear how weights should be set and there are reports where unweighted methods have outperformed weighted methods.

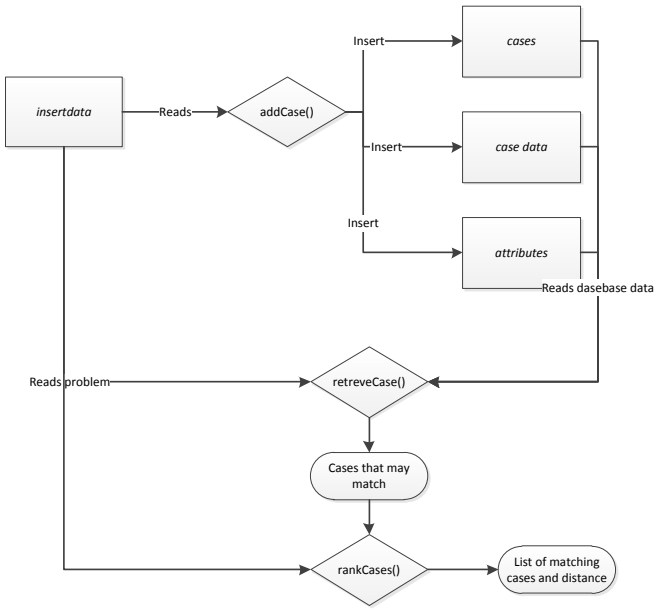

Fig. 5 Database layout

The implementation consists of four database tables. The tables *cases*, *attributes* and *case data* are representing the case and table *insertdata* is where data that is used for adding a case or finding a solution to a problem. The calculations are represented by three stored procedures which are *addCase()*, *retreveCase()* and *rankCases()*. These procedures are based on the code specified in Maher, Zhang and Balachandran (1995). The relations between the tables and the stored procedures are outlined in Fig. 5. addCase() is used to add cases and it uses data from the insertdata table, retrieves cases that have one or more matching attributes and rankCases() takes the result from retreveCase(), calculates a score and presents the cases with a distance measure between 0 and 1. A low distance between the problem and the case means that it is closer and will be ranked higher.

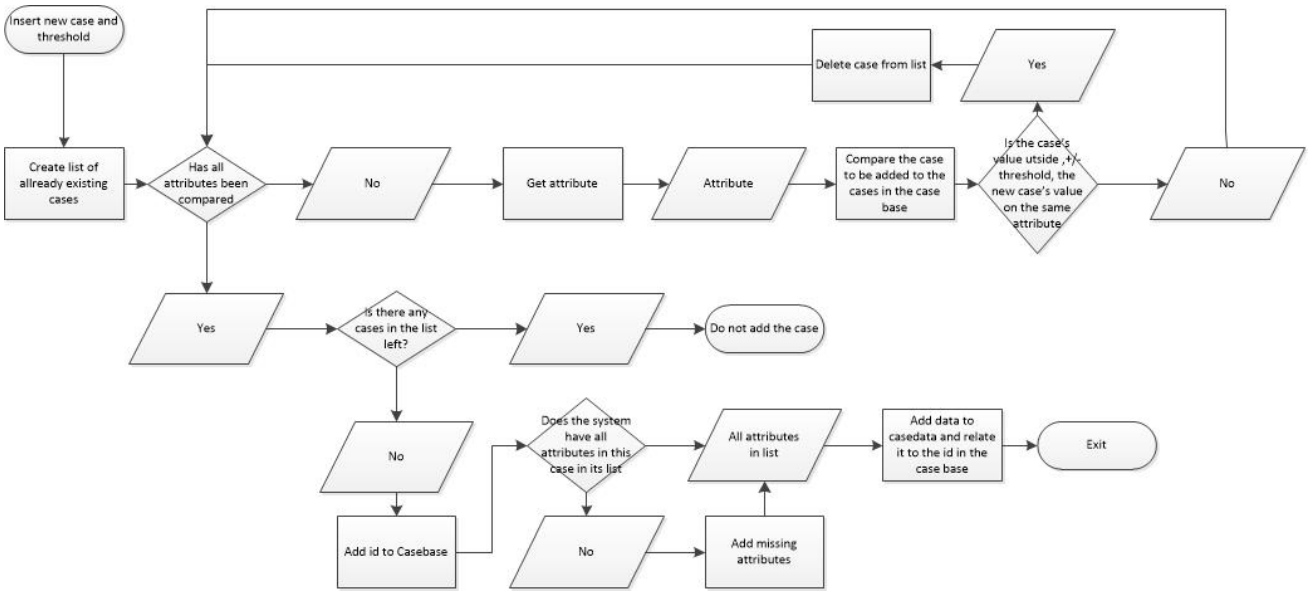

Fig. 6 addCase()

The procedure addCase() takes the new case and compares it to the cases in the casebase. For each case in the casebase, each attribute is tested and if all attributes on the case to be added is inside the range, which is from the value of the attribute minus the set threshold to value plus the threshold, of the case in the casebase, the case is not added. If at least one attribute is outside the range, the new case is added to the casebase. The program's structure is visualized in Fig. 6. When a case is added to the case library a new row is added to the *cases* table containing a unique identifier, date and time. After this the attributes of the case inserted into *case data* table containing references to the unique identifier of the case and type of attribute which is specified in the *attributes* table.

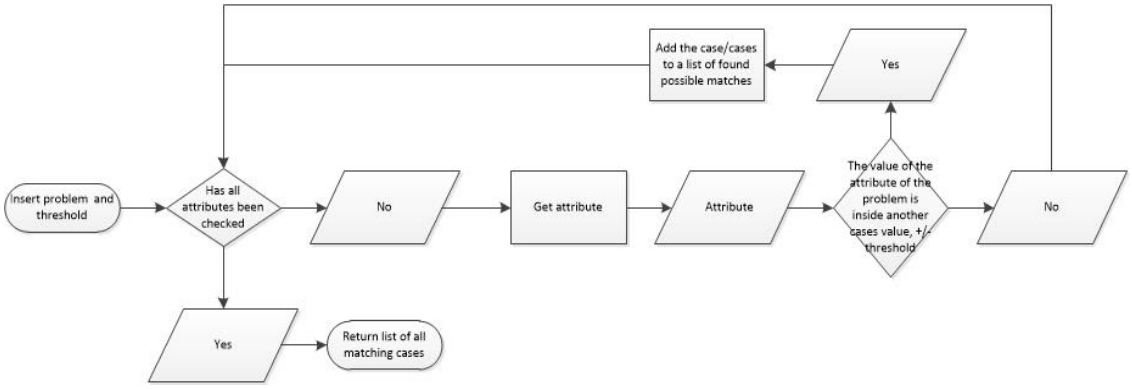

Fig. 7 RetrieveCase()

The procedure retrieveCase(), Fig. 7, checks if there is any case with a matching attribute which is in the range, which is from the problems value adding the threshold to problems value subtracting the threshold, of the problems value on the same attribute. Each case that matches is added to a list of cases. This list is returned as a result when the function is executed.

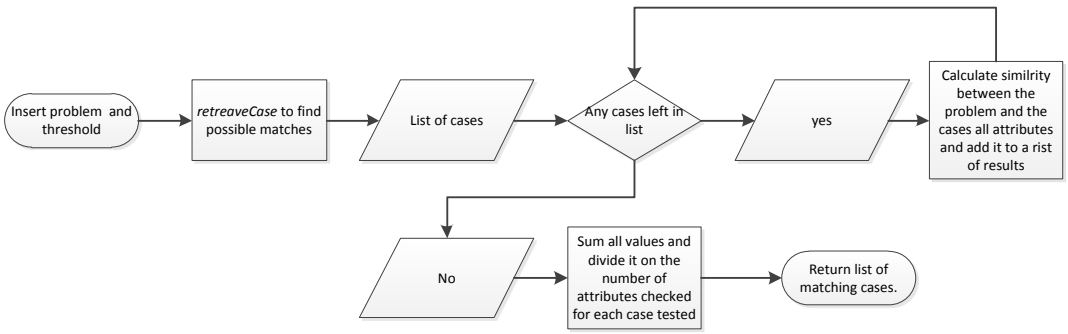

Fig. 8 rankCases()

The procedure rankCases(), Fig. 8, uses the result from retrieveCase() to get a list of cases that might match. For each of the cases in the list the distance of all attributes is calculated between the problems values and the cases values using the Nearest Neighbor UED algorithm presented in 5.2.2. This result is added to a list. When all cases have been calculated, the values are summarized for each case.

Data: Data preparation is an important step of the implementation as stated in 5.2.3 and there are different external factors that affect the consumption. As described previously, the day of the week is one of the factors outlined in the background. Outdoor temperatures and energy prices are two other factors that are directly connected to the energy consumption. An optimal case should therefore at least include information about how much electricity was used in total, price and temperature, preferable in hourly resolution. The case should also include information on day of the week and month to fulfill the main parts, as described for each case. This is done by having the external factors form the description of the situation or problem, the solution formed by the consumption over time and results represented by the total electricity used.

As a machine learning method is chosen, the *continuous data* needs to be discretized. Chen, Das and Cook (2010) argue that Equal-width is commonly used in data exploration, preparation of data, and data mining. This method has previously been used to preprocess continuous-valued attributes. The discretization is made as follows:

- **Discretization of time**
  When it comes to discretization of time intervals, Edwards, New and Parker (2012) states that there is a problem that most research that have been done on electrical consumption in the domestic sector is done on daily, not hourly basis. In this study this problem will be considered and the data will be based on an hourly output. This also fits the newly introduced system in Sweden, where our data comes from, where electricity can be bought on an hourly basis. Based on these preferences time will preferable be discretized in an Equal-width way, with 24 different bins, one for each hour.

- **Discretization of Consumption**
  As this study intends to be a foundation in a unsupervised, self-learning system the only supported discretization methods are equal-width and equal-frequency as described in Table 1. The discretization method equal-frequency divides the range into equally large bins with a specified width and equal-frequency divides the range into equally large bins with equally many results in each. The importance in this study is to get information from a behavior of change through the specified time and therefore is Equal-width a more suitable solution of discretization methods. In this study the consumption is discretized into bins with a span of 100kw/h, were first bin (1) contains 0-99, second (2) 100-199, third (3) 200-299 and so on.

- **Discretization of external sources**
  The external sources, such as weekday or temperature will as far as possible to be discretized to fit the time discretization. This is to be able to make as precise predictions as possible.

Discretization of the weekdays has in other studies been handled different. In the study of Hawarah, Ploix and Jacomino (2010), the authors are using only three values to represent weekdays; weekday, Saturday and Sunday. This is to show if there is a difference in behavior between these time periods. This study will instead save all weekdays for each case to be able to extract information if there are days where the consumers normally spend more electricity.

Data sets: The data used in this study in this study is both from the ASHRAE competition[5] and data gathered from a test site which has been created on a secondary school called Sötåsen in Sweden. The data from the ASHRAE competition "The Great Energy Predictor Shootout" was acquired by Jeff S. Haberl at the Texas A&M University which was one of the supervisors in this project. This data has been used in several other studies and it can be used to compare the results of this study with other prediction tests using a machine learning technique in the same domain.

There were two datasets in the ASHRAE competition. These two datasets contained hourly information on time and date, electricity usage, light and equipment electricity, motor and control center electricity, chilled and hot water for the building, temperature, humidity, global horizontal solar and wind speed.

[5] http://repository.tamu.edu/handle/1969.1/2137

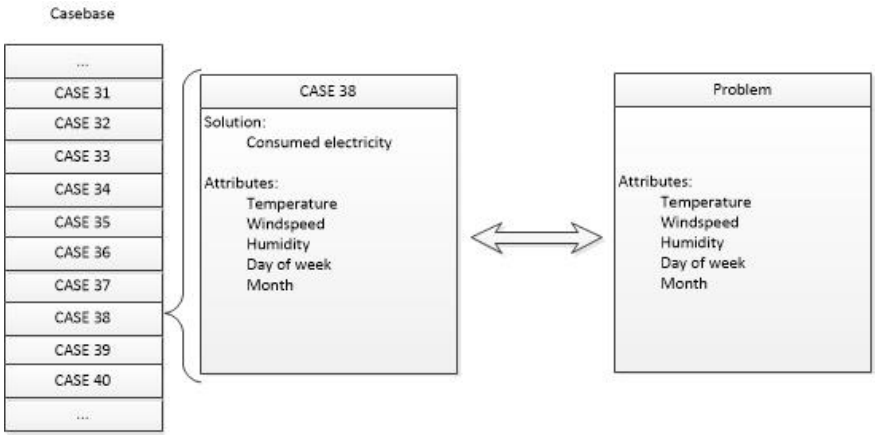

Fig. 9 A specific case, case 38, is on against the problem.

The test using this dataset will include parameters as temperature, humidity, wind speed, time and date to predict the electricity usage which is factors that can affect the consumption different depending on factors such as insulation. This is illustrated in Fig. 9, where the attributes of case 38 is compared to the problem. If the attributes from the case tested is equal or close to the problem, the case will have a short distance. If the case is the one with shortest distance the solution from the case will be used as the answer. These factors based on the theories by Swedish Energy Agency (2011) and Riihimäki and Koponen (2012) on what affects the electrical consumption.

The second dataset contains data from different sources. The source of electricity usage is from the test site Sötåsen where AMR have been installed in most buildings. The electricity consumption is measured with a resolution of about 12 second intervals.

The Sötåsen campus consists of several buildings including School buildings, mess and dorms. This can be seen as a minature version of a society, where people live, eat and work. The first source is from a dorm at Sötåsen where up to three persons can be living in the same "apartment" including one separate kitchen. The weather data is provided by SMHI[6] which is the main provider of weather data in Sweden. The electrical price data is provided by Nordpolspot which handles spot-prices for electricity in the Nordic countries. The data for electricity cost has an hourly resolution. The price is available for the user one day ahead so he or she can adjust and plan the consumption behavior in relation to the price over the day.

5.3 Evaluation of using CBR for predicting electric consumption

Accuracy measurements: To be able to measure how well the system performs, Edwards, New and Parker (2012) and Kreider & Haberl (1994) make use of Coefficient of Variance (CV), Mean Bias Error (MBE) and Mean Absolute Percentage of Error (MAPE). This way of measure the result was used in the ASHRAE competition among others. The CV is used to determine the overall prediction error variation in relation to the mean value. A high value indicates that the model has a high error range and is less trustable. The CV is calculated as follows (Edwards, New & Parker 2012):

[6] http://www.smhi.se/

$$CV = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{\bar{y}} * 100$$

Where $\hat{y}_i$ represents the predicted energy outcome, $\bar{y}$ the average energy consumption, y_i the actual outcome and N is the total number of predictions made.

A MBE calculation shows if the tested model favor a certain trend of the result, in this case if it is over-estimating or under-estimating the predictions. A result close to zero is ideal. This was used to break ties between the different methods tested in the competition. The MBE is calculated, according to Edwards, New & Parker (2012), by the formula:

$$MBE = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)}{\bar{y}} * 100$$

Where $\hat{y}_i$, $\bar{y}$, y_i and N the represents the same values as in the formula for calculating the CV.

MAPE measures the percentage of error per prediction. The variables $\hat{y}_i$, $\bar{y}$, y_i and N the represents the same values as in the formula for calculating the CV. MAPE is calculated by the formula according to Edwards, New & Parker (2012):

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} * 100$$

Where $\hat{y}_i$, $\bar{y}$, y_i and N is representing the same values as in the CV formula.

5.3.1 Experiment 1

The first test was executed using data from the ASHRAE competition. The data from this source was structured into two files, one is which should be used as a training set and the second is used as the test set. The dataset contains data that has been excluded and replaced by the value "*-99*". The CBR system has been adjusted to deal with this and does not include the attribute has a -99 value on either the case or the problem. A value that has a -99 value as a result value is ignored because it cannot be included in the calculation of CV or MBE the difference between the predicted value and the real outcome is not known. Test was made by comparing; using CBR, one situation to the cases in the case base, i.e. each hour was treated as one problem – case comparison. The best matching case is used as the solution. The distance to the result was not added as an factor in order to be able to be used in the evaluation methods.

For the purpose to compare the CBR technique to other techniques, two tests were made on data from ASHRAE competition data set. This testset has been used previously when prediction techniques accuracy have been measured, not only by the competition, but also in Edwards, New and Parker (2012).

	CV (%)	MBE (%)	MAPE (%)
Test1	19.72	3.43	16.08
Test2	51.31	-5.28	44.72
Test3	22.00	-2.29	18.97
Test4	53.18	-28.91	65.03

Table 2: Result of Test1 to 4

Test1: The first test made on the first, "c", dataset from the ASHRAE competition cases was only added from the training set (1990-01-01 to 1990-11-23). Each case consisted of data for one hour. The case contained data about the electrical consumption, temperature, humidity and wind speed. Thereafter the test dataset was used to represent the problems. The test set contained information between 1990-11-28 and 1992-12-31. The problems were compared to the cases in the casebase and the result of the accuracy measures from this test is presented in Table 2.

Test2: The second test made on the dataset, "d", from the ASHRAE competition cases was, as Test1, only added from the training set (1990-12-22 to 1991-07-12). A case consisted of data for one hour and contained data about the electrical consumption, temperature, humidity and wind speed, as Test1 does. The dataset containing the problems stretched from 1991-07-13 to 1992-12-31. The result of CV, MBE and MAPE is presented in Table 2.

Test3: Thereafter a second test was made using the dataset "c" from the ASHRAE competition. In this test case was first added from the training set. The tested problem was added as a case to the casebase after the problem was tested. The cases consisted of data for one hour and contained data about the electrical consumption, temperature, humidity and wind speed. The same training- and testset was used as in Test1. The used calculation for how the method performed using this data is presented in Table 2.

Test4: In the last test using the data from the ASHRAE competition the dataset "d" was tested by first adding cases from the training set. For each of the tested problems a case was added to the casebase. In this test each case consisted of data for one hour and contained data about the electrical consumption, temperature, humidity and wind speed. The same training- and testset was used as in Test2. The results of this test can be found in The used calculation for how the method performed using this data is presented in Table 2.

Edwards, New and Parker (2012) tested the techniques that performed best in the ASHRAE competition in their work. These values can be used to compare with the result of Test1 and Test2 since they are using the same dataset, ASHRAE competition dataset. They are also using the same performance measure algorithms for CV, MBE and MAPE.

	CV (%)	MBE (%)	MAPE (%)
Regression	14.12	7.69	13.41
FFNN	11.29	8.32	9.14
SVR	11.93	8.95	9.63
LS-SVM	13.70	10.32	11.21
HME-REG	14.11	7.66	13.40
HME-FFNN	11.49	2.91	9.73
FCM-FFNN	11.51	8.71	9.45

Table 3: Different techniques tested using the ASHRAE competition data by Edwards, New and Parker (2012)

The results presented in Edwards, New and Parker (2012) using the winning prediction methods from the ASHRAE contest on the ASHRAE contest data were based on a predictor algorithm trained using the training set. The predictor was thereafter used to predict the consumption in the test set.

5.3.2 Experiment 2

Test5: Next test was made using data from the Sötåsen School. This dataset represents the domestic sector. The dorms consumption is used as consumption data to represent the individual's different needs of electricity. This data has a 12 second resolution which needs to be discretized to match the case. The consumption data was gathered under the time period 2012-05-25 to 2012-10-23 (some days were missing). Data was provided from four different houses which have different consumption needs and household settings. On this dataset we have the electrical price which is provided by nordpolspot. Weather data is provided by SMHI. Resolution on weather data is 3 readings for each day. The data was therefore adjusted to fit the resolution on the temperature data which had the lowest resolution. 1/3 of the dataset was used as training set and the rest was used as testing set. In this test the tested problems was not added as cases after being processed.

	CV (%)	MBE (%)	MAPE (%)
House 4	34.98	-13.27	25.49
House 5	55.71	-27.30	51.37
House 7	56.65	-15.71	46.53
House 8	40.47	-8.10	24.27

Table 4 Performance results using Sötåsen consumption data, not adding used data to case base

The results of CV, MBE and MAPE for the predictions made on Sötåsen are presented in Table 4.

Test6: The last test was made using same data as Test5; consumption data from Sötåsen was used. The only thing that differs is that in this test problems were added as cases after being processed.

	CV (%)	MBE (%)	MAPE (%)
House 4	34.98	-13.17	25.49
House 5	55.71	-27.30	51.37
House 7	56.65	-15.71	46.53
House 8	40.42	-8.17	24.13

Table 5: Performance results using Sötåsen consumption data, adding used data to case base

The results of CV, MBE and MAPE, from the prediction test made on Sötåsen, are presented in Table 5.

5.4 Analysis

By first comparing the results from the tests where ASHRAE data set were used one can see that the results differ between the tests. This can be a result of a the prediction method not performing as well as the winning methods from the ASHARE competition but it can also be a product of not using all of the supplied attributes. In this study we only made use of the attributes that has been reported having an impact on the result. By first looking at testset "c" the test where no new cases after a problem had been solved, test1, had a 2.28% better CV than Test3, which used the same testset but was adding cases to the casebase after each problem solved. Comparing the CV from Test2 with Test4, which used dataset "d", shows that Test2, the test that did not add new cases after each test, got a 1.87% better CV result than Test4. The best CV result was Test1 using data from ASHRAE competition. This indicates that the more cases added in the casebase the less accurate the test became.

This is the opposite of what Schiaffino & Amandi (2000) found in their report. They used a Bayesian Network to find which attributes that are relevant. In these tests there are no weights used. If weights would have been used this might have changed the results.

The test used as comparison from Edwards, New and Parker (2012), where the winning methods of the ASHRAE competition using the ASHRAE dataset is presented, is not reusing testset data after each prediction made. To compare the data from the tests in this study, the results where data was not added to the case base after each solved problem will be used. The most appropriate test to compare with in this study is therefore Test1. We can see that the result of Test1 is not as precise as the winning methods from the ASHRAE competition. Comparing to CV, CBR has an accuracy of 5,6% lesser than the method that has the highest CV but still counts as one of the best prediction methods. Looking at MBE, the CBR method has a better result than several of the best methods and MAPE has a value of 2.67% lesser than the method that has the highest MAPE value in the ASHRAE test.

Comparing the results in Table 3 with the results from Test1 and Test3 (Table 2) conclusions can be made that the techniques winning the ASHRAE contest has a higher accuracy in CV, MBE and MAPE than CBR using the configuration used in this study.

Test3, presented in Table 2, provides us with a slightly better result on MBE and MAPE but not CV. Comparing the results from the study from Edwards, New and Parker (2012) to Test2, where

the problems solved are added as cases in to a case library (as a CBR system should work in a real environment), conclusions can be made that the techniques tested in Edwards, New and Parker (2012) provides us with a better result.

MBE shows if the method is over-estimating or under-estimating. A result close to zero is preferred. Test1, 2 and 3 presents a result that is close to, or better, than the best performing methods in the study by Edwards, New and Parker (2012). Test1 and 3 is performing better than Test2 and Test4, this is the test that was not adding new cases to the casebase after a tested problem. This is the same conclusion as the CV analysis. MBE result in Test4 stands out more than the other tests with a MBE at -28,9%.

This could be a result of more than one case fitting the problem. Since no solution is used over another the first fitting solution of the ones that has the same range to the problem is the one picked as solution. Results that are more accurate could be produced by adding weights instead of just picking the first of the fitting solutions.

The MAPE result, which measures the percentage of error per prediction, does not differ from the CV and MBE results. The datasets where cases were not added after each problem checked performed better than the ones adding problems as cases. On the first, "c", dataset the test without adding cases (Test1) after each problem was 2.89% better than the test where cases was added (Test3). On the second dataset the test without adding cases (Test2) was the MAPE result 20.31% better than the problems were added as cases (test4). The results from Edwards, New and Parker (2012) had a better outcome on all MAPE results. Comparing the results in Edwards, New and Parker (2012) to the CBR method used in this study one can find that the CV value and MAPE has approximately the same difference between them in for all methods used, where MAPE is mostly in the range 80-87% of the CV value. The only test that stands out is Test4 where MAPE exceeded the CV value.

Comparing the results from Test5 and Test6 with the results from Test1, 2, 3 and 4 one can see that the result differs between business (Test1, 2, 3 and 4) and domestic sector (Test5, Test6). The business sectors predictions are more accurate than the domestic sector. The results in Edwards, New and Parker (2012) also shows that prediction made using the selected machine learning techniques performed much better on the ASHRAE dataset then a dataset from the domestic sector. The high error rate can be due to that people does not always follow the same procedure every day which affects the energy consumption.

The results did not differ much between the test where cases was added (Test6) and the test where cases were not added after each problem (Test5). The only result that differs is House4 which had a slight accuracy increase in MBE and MAPE and House8 which had a small accuracy increase in CV and MAPE, but an accuracy decrease in MBE. This indicates on that more cases could have changed the accuracy for the better. This is consistent with Schiaffino & Amandi (2000) which has found that as the casebase grows in number the predictions should increase in accuracy.

The data sets in Test5 and Test6, Sötåsen dataset among others, has more things that could affect the prediction result. The weather readings are only 3 readings each day which makes the test use a low resolution on the readings. The temperature reading facility is located in a city

which is more than 20km away. The weather can differ from these places as they have different preconditions such as the town where the temperature facility is located is by a lake and Sötåsen is not. The price factor which is included as a factor that influence the consumption is a relatively new factor in Sweden and a big part of the population has a contract that has a fixed price no matter on what time the electricity is consumed.

Edwards, New and Parker (2012), tested the winning methods of the ASHRAE contest on data representing the domestic sector, which was called Campbell creek house. The methods tested did not perform as good as on the ASHRAE testset. The CV had a result with as best at 20% and at its worst up to 40%. Comparing to the results from the domestic sector in this study the results are within the range of the best predicting methods or just over. The same conclusion can be drawn when comparing MAPE whereas MBE is better on the ASHRAE competition winners. Worth noting is that these two tests, this study and Edwards, New and Parker (2012) does not use the same data set for the domestic sector.

A CBR prediction method, not using weights do define which attributes influence the result more than other attributes, using a nearest neighbor technique, is not performing as well as the winning techniques from the ASHRAE competition. Testing on the same dataset gave a result where the pre-exiting techniques had higher accuracy on CV, MBE and MAPE.

5.5 Summary of result

The CBR method can be used as a predictor, as shown in this study. A summary of the results can be shown below. First the results from within the study are compared.

	CV (%)	MBE (%)	MAPE (%)
Test1	19.72	3.43	16.08
Test3	22.00	-2.29	18.97

Two tests were made on testset c. In Test1, no data were added after each tested case and in Test 2 were tested data added as cases to the casebase after it had been tested. The results were:

- Test1 has 2.28% better CV(%) result than Test3.
- Test3 has a 1.14% better MBE(%) result than Test1, as a result close to zero is preferred.
- Test1 has 2.89% better MAPE(%) result than Test3.

As can be seen here the Test1 had a better CV and MAPE result than Test3. Test3 on the other hand had a better MBE result than Test1. The results did not differ very much in these tests, where the same method was compared and the only thing that did differ was: adding and not adding processed cases to the case library.

	CV (%)	MBE (%)	MAPE (%)
Test2	51.31	-5.28	44.72
Test4	53.18	-28.91	65.03

Two tests were made on testset d. In these tests were Test2 the test where no case were added after each tested case. In test 4 the tests were added to the casebase after each tested case. The results were:

- Test2 has 1.87% better CV(%) result than Test4.
- Test2 has a 23.63% better MBE(%) result than Test4.
- Test2 has 20.31% better MAPE(%) result than Test4.

In the comparison between Test2 and Test4 one can see that Test2 produces a more accurate CV, MBE and MAPE answer. Here we can see that the results are much better in Test2 in MBE and MAPE.

As CV, MBE and MAPE was calculated in the same way as other studies, the result can be compared to result in other studies. The results in this study are compared to the winning methods in the ASHRAE competition Table 3.

	Type		Test1	Diff	Test3	Diff
Regression	CV(%)	14.12	19.72	5.6	22.00	7.88
	MBE (%)	7.69	3.43	4.26	-2.29	5.4
	MAPE (%)	13.41	16.08	2.67	18.97	5.56
FFNN	CV(%)	11.29	19.72	8.43	22.00	10.71
	MBE (%)	8.32	3.43	4.89	-2.29	6.03
	MAPE (%)	9.14	16.08	6.94	18.97	9.83
SVR	CV(%)	11.93	19.72	7.79	22.00	10.07
	MBE (%)	8.95	3.43	5.52	-2.29	6.66
	MAPE (%)	9.63	16.08	6.45	18.97	9.34
LS-SVM	CV(%)	13.70	19.72	6.02	22.00	8.3
	MBE (%)	10.32	3.43	6.89	-2.29	8.03
	MAPE (%)	11.21	16.08	4.87	18.97	7.76
HME-REG	CV(%)	14.11	19.72	5.61	22.00	7.89
	MBE (%)	7.66	3.43	4.23	-2.29	5.37
	MAPE (%)	13.40	16.08	2.68	18.97	5.57
HME-FFNN	CV(%)	11.49	19.72	8.23	22.00	10.51
	MBE (%)	2.91	3.43	0.52	-2.29	0.62
	MAPE (%)	9.73	16.08	6.35	18.97	9.24
FCM-FFNN	CV(%)	11.51	19.72	8.21	22.00	10.49
	MBE (%)	8.71	3.43	5.28	-2.29	6.42
	MAPE (%)	9.45	16.08	6.63	18.97	9.52

Table 6: Winning ASHRAE methods in relation to CBR using same dataset.

The results of Test1 and Test3 in comparison to the winning methods of the ASHRAE competition, using the same dataset, are presented in Table 6. As can be seen here the CBR method was not as good as the winning methods in the ASHRAE competition on CV and MAPE. The CBR method on the other hand had a better MBE than all other methods except HME-FFNN.

Looking at the tests made on the data from Sötåsen among others, presented in Table 4 and Table 5, one can see that it does not differ much in the test where data has been used as cases after being tested versus not using tested data as cases.

	CV (%)	MBE (%)	MAPE (%)
House 4.1	34.98	-13.27	25.49
House 4.2	34.98	-13.17	25.49

	CV (%)	MBE (%)	MAPE (%)
House 5.1	55.71	-27.30	51.37
House 5.2	55.71	-27.30	51.37

	CV (%)	MBE (%)	MAPE (%)
House 7.1	56.65	-15.71	46.53
House 7.2	56.65	-15.71	46.53

	CV (%)	MBE (%)	MAPE (%)
House 8.1	40.47	-8.10	24.27
House 8.2	40.42	-8.17	24.13

Table 7: Results of electricity consumption prediction at the dorms on Sötåsen using CBR.

In Table 7 the results with name ".1" is representing the results where data was not added to the casebase after each test and ".2" is representing results were tested data were used as cases after being processed. As can be seen here on House 4.2 has a better MBE result than House4.1 and House 8.1 has a better MBE result than House 8.2, but 8.2 has a better CV and MAPE. Looking at this it seems like the tests where cases were added after being tested produces a better result but it is very small differences and to draw that conclusion more tests should be made.

6 Discussion

In this study, the method CBR has been used. CBR can be seen as a methodology, since it can be implemented in different ways as long as a set of rules is followed (Watson 1999). These rules are Retrieve Reuse Revise and Retain, and are described in chapter 5.2.1. These rules were followed in this study when the CBR system was created.

To evaluate the results, an approach previously used in the electrical consumption domain was used. The results were compared to other methods which had been evaluated using the same evaluation method and tested using the same dataset.

As found in the Result and Analysis chapters, the CBR method used in this project is not performing as well as most of the winning methods from the ASHRAE competition. It was also found that the CBR method used in this project performed worse on the data from the domestic sector than the data from commercial sector, just as the winning methods from the ASHRAE competition and therefore was this result was expected.

There were both unexpected and expected results in this study. Accuracy result from the accuracy of the CBR method that was not far away from the results from the winning methods from the ASHRAE competition. The result was expected to be near, but not as accurate as the methods tested in the ASHRAE competition. The results were also expected to become more accurate as more cases were collected. The test in this study did not live up to that expectation. The reason to why the result is not as accurate as the ASHRAE competition and what could have been better is partially discussed in the analysis chapter. One factor could be that the results, predicted by the CBR method, could not be new yet un-experienced situations. In a new situation, a prediction will give the closest case as the answer and a distance calculation. If the distance is not considered the answer could be interpreted as a perfect match. In this work, to compare the results with other techniques, the results are treated as a perfect match. The biggest factor that could have changed the outcome is probably by adding weights. This could have changed the predictions in several ways such as if there are several really close cases where one has a common outcome and the other prediction is rarely correct.

The knowledge of predicted electrical consumption in this study has focused on end users, and how the predictions could help them. The result of predicted consumption could also be used in the domain of electrical distribution. If the electrical distributor can get a prediction of how the energy consumption will be distributed, they can plan for how much energy is needed, on which location and when it is needed. This can be done by summarize prediction data for different distribution zones. The benefit could be both economical, as they don't need to buy more energy than needed in their turn, and ease the load on the grid as they could plan the distribution in a better way.

If energy storage is added to the system, either at the consumer or the grid, it could be configured using predictions, as the ones produced by the CBR system, on the consumption behavior. A consumer could use the energy storage to store energy when the price is low and use it when he or she need to either use a large amount of energy or when the price is high. This can be automated as well using the predictions. By using the energy storage in the grid, the predictions can help to decide when the load has its peaks and when it is low load on the grid.

The energy storage could be charged when the load is low and when the load is high and during peaks the energy storage could help by supplying energy back to the grid and normalizing the load.

As this implementation makes use of a relational database, which is commonly used in different systems on the market, it can be implemented with a few tweaks in to existing solutions. The CBR system used in this study can easily be altered to add a weighting factor to the result. But as no appropriate approach to weigh has been found this implementation has not been tested. If an external weighting system was used the CBR system could, with just a few additions, make use of these.

In this study, there were some limitations that might have affected the result. In the results using data from Sötåsen among others, the weather data had a low resolution. This limited the number of predictions to only three each day. By only being able to do three predictions each day gave a prediction that was very limited. The weather data provider could also only provide the temperature, not humidity, windspeed and other factors that could have influenced the consumption. Also since the closest weather station is located in the town Mariestad, which is located by a large lake about 20km west of Sötåsen, the conditions for the weather settings is different from Sötåsen.

If weather data could have been gathered form a resource, closer to the energy measurement point, in a higher resolution the result would probably been better. If more factors could have been provided that influences the electricity consumption could have been gathered the accuracy of the results could have been better.

In the domain of electricity consumption, the measured energy usage can vary, due to climate and season or weather, and also due to consumer's preferences and habits. To have updated weights to these factors, the weights should preferably be dynamic. One should have in mind that the consumer may choose to change his or her consumption behavior and the former weights may not be accurate anymore.

By not using weights, the case that has an outcome that rarely fits the actual outcome, can be selected by the order of the fitting results. An interesting thing to research further is where to add weights, and what values they should have. According to Mendes, Watson and Mosley (2002) it can be hard to find good weights.

Another thing that is important to think of when adding weights to a system like this is where the weights should be applied. The weights could be applied on the result to give extra weight to a result that are more common if the attributes are in a certain state where more than one result has the same, or close to same, values. Another scenario could be that the weights were used to control which attributes that had most impact on the attributes. This would make it possible to control that e.g. temperature had a bigger impact on predictions in the winter while the price of electricity was the factor that had most impact during the summer.

To initiate weights in a system it had to either be trained using old data, or if one could create a small set of cases that would give some general consumption rules that would fit for most but also be so small that after a short time of use it could be replaced by the cases collected from that

specific location. Schiaffino & Amandi (2000) used a Bayesian Network to create weights between the attributes. As more data was processed the relations were updated and more accurate. Their test was made in a school where certain behavior was predicted. Using this method in the electrical domain would require some work on how the Bayesian Network was set up. MacKay (1994) used a technique called automatic relevance determination to find out which of the attributes that had the largest impact on the outcome. It may be useful to see how these techniques can affect the results of the CBR method.

There are tweaks that could have affected the result regarding the cases. The tests using ASHRAE dataset showed that accuracy was better when no new cases were added. In the tests where Sötåsen data was used, the accuracy was better as more case were added. This indicates that there is a limit where more cases no longer contribute to a more precise result.

There are a few methods to restrict the number cases in the casebase. First it should be studied if and how different dataset are affected by different sizes of the casebase. Then the method of restricting the size should be examined.

One simple way is to restrict the number of the cases in the casebase. To get similar cases to i.e. a date in July, it is relevant to check cases that are created this month, or the months before and after, to get a promising result; especially when the affecting factors such as weather can differ so much as they do in countries like Sweden, with many climate zones. If the restriction should be set on cases within a certain time, let us say 1 year, i.e. 12 months, the comparison for the month before last year would not been taken into calculation. Another way to restrict the number of cases in the casebase is to check the similarity of the attributes. By adding rules that can restrict cases that has all attributes within a specified range. As an example, the specified range is set to 10%, if a case has all attributes and the result within the set range, 10%, of another case of another case, this new case is not added to the casebase. By doing this you get a casebase that has a wide range between the cases.

7 Conclusion

In this study, we have examined how short-term electricity consumption predictions can be made. The predictions are made a framework which makes use of external factors in the calculation. The focus is to use it in the domestic sector. The domestic sector is interesting because it has potential to make changes and savings and the electricity can be affected by many things such as insulation and the choice of heating method.

A literature study has been performed to suggest a set of properties that have positive impact on the results from a method. In the literature, Case Based Reasoning, a yet untested a method in the domain of electrical predictions, was found that could be matched to the properties. This method was chosen and studied more closely. This method was found to have several different ways of implementation strategies. This was considered and one strategy was chosen and implemented using databases. Thereafter, the method was tested using data from two different sources, one representing the commercial sector and one representing the domestic sector. The dataset representing the commercial sector had been used before, to test efficacy of electrical consumption predictions. This dataset was therefore used for benchmarking the method. As there has been research done in predicting the consumption the same evaluation methods was used as the ones that has been used in this domain earlier ad described in (Edwards, New & Parker 2012).

In the literature study, a number of methods were examined and a few interesting properties were found. These properties were interesting because they seemed to increase the accuracy of the predictions. These properties were:

- Be able to assign customers to a certain load profile.
- Keep the system updated by adjusting as more data is acquired.
- Adjust the relevance of the external factors over time.

CBR was one of the methods that were found in the literature study. This method was found to be promising because it could fulfill, or partially fulfill, the accuracy increasing properties. This method was studied using an SQL implementation and then evaluated through a number of tests. The use of a relational database is beneficial if the solution is to be implemented in to existing solutions. One of the most beneficial attributes of CBR is that it is able to reuse old data which can make the predictions context specific.

The evaluation method that was used was first found in a competition where predicting energy consumption was the goal. The evaluation methods showed how high accuracy the predictions had when compared to the real outcome (Kreider & Haberl 1994). The evaluation method has later been used in similar studies, such as Edwards, New & Parker (2012).

In the study of Edwards, New & Parker (2012) they tested the winning methods from the ASHRAE competition on the ASHRAE dataset. These results were used in this study to compare the results from the CBR method with. The same evaluation method and dataset were also used.

	CV (%)	MBE (%)	MAPE (%)
CBR_1 (Test1:no cases added after initial training)	19.72	3.43	16.08
CBR_2 (Test3: cases added after initial training)	22.00	-2.28	18.97
Regression	14.12	7.69	13.41
FFNN	11.29	8.32	9.14
SVR	11.93	8.95	9.63
LS-SVM	13.70	10.32	11.21
HME-REG	14.11	7.66	13.40
HME-FFNN	11.49	2.91	9.73
FCM-FFNN	11.51	8.71	9.45

Table 8: Results Using CBR in relation to the winning methods in the ASHRAE competition

Table 8 shows the results from this study together with the results from Edwards, New & Parker (2012), where the best performing methods where tested. The addition from this study is the two CBR results, where CBR_1 is presenting the results where the system got trained and made all the predictions, and CBR_2 is presenting results from the test where the system first got trained. After each prediction made the case was added to the case base, in order to extend the possible predictions. The CBR method does not use all of the attributers supplied from the ASHRAE competition we only made use of the attributes that has been reported having an impact on the electricity consumption.

As Table 8 shows, the winning prediction methods from the ASHRAE competition has all a better CV and MAP result then the CBR that was tested in this study. When evaluating MBE from the result the CBR method was in the top 3 best results.

	CV (%)	MBE (%)	MAPE (%)
House 4	34.98	-13.17	25.48
House 5	55.71	-27.3	51.37
House 7	56.65	-15.71	46.53
House 8	40.42	-8.17	24.13

Table 9: Results in the domestic sector

The test using data representing the domestic sector had a high error rate, as can be seen in Table 9. This result was expected, as the domestic sector does not have same level of routine as the commercial sector. This is not something that is specific for this study, Edwards, New & Parker (2012) had the same conclusions in their report when testing the method on data from the domestic sector. Also in the dataset that was used for the domestic sector there were several thing that probably affected the outcome for the worse, such as that the only temperature was provided regarding weather and the closest weather station was located 20km west from the actual building that was used for predictions.

The CBR method was expected to perform better than what it did, but still not as good as the winning methods from the ASHRAE competition. If the CBR method had been tweaked with different properties, such as weights, different ways to calculate similarity and so on, the method might have given a more accurate result.

The method, as it is designed in this study, does not provide a result with a higher accuracy as the winning methods used in the ASHRAE study. As brought up in the discussion there are a few changes that could change the performance of the method, such as considering and implementing weights. This was one of the toughest choices in this study since there were studies that showed that it was beneficial to make use of weights and studies which did not benefit from the use of weights. There were also studies which pointed out that the use of weights in not something that can easily be implemented; it requires a lot of work to implement it in an optimal way.

This study shows that CBR can be implemented using a relational database, and used to make predictions using external sensor data as a foundation. The study also explains that CBR can be implemented in several ways which can affect the outcome. This study has opened several possible paths for future work to utilize.

8 Future work

In this study a framework was created to be used as an aid for households, in order to predict energy consumption. These predictions can be used in a first stage, as an aid to the consumers when they are planning their energy consumption. Planning is needed for demand-side active households because there are times during the day, where energy is cheaper and has a less impact on the environment. Also the consumption prediction can be used to try to avoid peaks in the electricity, by not using all electrical equipment at the same time. In a second stage the predictions could be used by a smart home to automatically operate certain appliances, such as the *Short-term scheduled tasks* and *Uncheduled tasks* on its own to start these when energy is cheap, has a lesser impact on the environment and does not peak the electricity consumption in a home.

The CBR framework can capture consumer's consumption profile and the system adapts to the consumers' habits. Captured CBR data can be used in a recommender system, or similar systems, in order achieve peak shaving and electrical cost reduction by providing consumers with guidance. Using CBR to make electrical predictions has several interesting aspects that could be explored. The CBR method, as it was configured in this study, did not reach the same accuracy level as the winning concepts from the ASHRAE competition. It would be interesting to see, in a future work, which impact these CBR configurations would have.

One improvement could be the use of weights. Weights can be configured in several ways and affect the predictions different. To research how to implement weights for a better result would probably be beneficial for a more accurate result. In this area research could be made on how and where the weights should be implemented, if it is more beneficial to see if one result is more common than another when a certain setting is provided or if a specific attribute in the case is affecting more, if so is it just in a specific setting, such as just winter this attribute is affecting more. Can the weights detect if the consumer changes his behavior and react on it?

Another interesting possible improvement is how similarity is configured. As discussed in the study, there are several ways to calculate similarity. By testing several of these, and make a deeper study of if there are other similarity calculations would be interesting and probably beneficial for the result.

Making use of more, or other, data could change the outcome for the method. Other inputs that would have been interesting to add is data about, if the person, or several people, is located at the area that is predicted. If the user could give input to the system, supplying information if there will be people at home, if so, how many. The seasons is also one attribute which might have an positive impact on the result as the days in the same season tend to have common attributes.

A scenario worth researching in the domain of consumption prediction is if it would be possible to make use of AMI's in smart homes, where the home can decide to start and stop certain appliances. If it would be possible to use the smart home together with a prediction algorithm and use the AMI to communicate, either directly with AMI's nearby or through a server, to help avoiding consumption spikes for an certain area.

Another area that would be interesting to research is how the information of the prediction could be presented to the consumer, so he or she can understand the prediction and react on it. It would also be interesting to investigate if the consumer could interact with the system where predictions can be made. This opens up for the possibility to not only using external data, which the user cannot affect, but also include data on if he or she is at home, is he or she alone, do they plan on doing something at home or are they planning on going out.

In the domain of predicting future energy consumption there are several interesting aspects. The result of aggregated predictions for households could potentially be used by other sectors such as electrical companies where they could plan the energy load and predict when certain areas or groups of consumers' need the extra power.

In countries where there are problems with overload in the grid, the consumption prediction could be helping to foresee where the load is high and help spreading the load by inform the population in the specific area how to react to normalize the load. Using local energy storages to normalize the load when the load is high could also be an option.

Bibliography

ABB, *Så bor vi i framtiden*, viewed 01 November 2012, <http://www.minimedia.se/abb/sa-bor-vi-i-framtiden>.

Anners, C & Enmalm, S 2012, *Electricity supply, district heating and supply of natural and gasworks gas 2011*, Energimyndigheten & SCB, viewed 02 December 2012, <http://www.scb.se/Statistik/EN/EN0105/2011A01/EN0105_2011A01_SM_EN11SM1202.pdf>.

Australian Government: Department of Climate Change and Energy Efficiency 2010, *Smart Grid, Smart City, Facts about the national Energy*, viewed 01 December 2012, <http://www.climatechange.gov.au/~/media/publications/smart-grid-smart-city/smart-grid-factsheet2.ashx>.

Boström, H, Andler, SF, Brohede, M, Johansson, R, Karlsson, A, van Laere, J, Niklasson, L, Nilsson, M, Persson, A & Ziemke, T 2007, 'On the Definition of Information Fusion', Högskolan i Skövde, Institutionen för kommunikation och information, IKI Technical Reports; HS- IKI -TR-07-006, Högskolan i Skövde, Skövde.

Chen, C, Das, B & Cook, D 2010, 'Energy Prediction Based on Resident's Activity', *Knowledge Discovery from Sensor Data*, Washington, DC.

Chen, CS, Kang, MS, Hwang, JC & Huang, CW 1999, 'Implementation of the load survey system in Taipower', *Transmission and Distribution Conference*, IEEE Operations Center, New Orleans, LA, USA.

Chmielewski, MR & Grzymala-Busse, JW 1996, 'Global Discretization of Continuous Attributes as Preprocessing for Machine Learning', *International Journal of Approximate Reasoning*, vol 15, no. 4, pp. 319-331, <Doi: 10.1.1.85.9533>.

Costanzo, GT, Kheir, J & Zhu, G 2011, 'Peak-Load Shaving in Smart Homes via Online Scheduling', *IEEE International Symposium on Industrial Electronics (ISIE)*, Gdansk, Poland, <10.1109/ISIE.2011.5984354>.

Dent, I, Aickelin, U & Rodden, T 2011, 'The Application of a Data Mining Framwork to Energy Usage Profiling in Domestic Residences Using UK Data', *"Buildings Don't Use Energy, People Do?" – Domestic Energy Use and CO2 Emissions in Existing Dwellings*, Buildings Do Not Use Energy, People Do Research Student Conference, Bath, UK, Bath, UK.

Do Thi Kim, Y 2011, 'Data Modelling of Electricity Data in Sweden', Master thesis, Communication and Information, http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-5236, Skövde.

Edwards, RE, New, J & Parker, LE 2012, 'Predicting future hourly residential electrical consumption: A machine learning case study', *Energy and buildings*, vol 49, pp. 591–603.

Energimyndigheten 2011, *Energy in Sweden 2011*, viewed 26 November 2012, <http://webbshop.cm.se/System/TemplateView.aspx?p=Energimyndigheten&view=default&id=3928fa664fb74c2f9b6c2e214c274698>.

Energimyndigheten 2012, 'Energiindikatorer 2012: Uppföljning av Sveriges energipolitiska mål', Statens energimyndighet, Eskilstuna.

Energimyndigheten 2012, *The electricity certificate 2012*, viewed 30 November 2012, <http://webbshop.cm.se/System/TemplateView.aspx?p=Energimyndigheten&view=default&id=806d7f0a547f45f9b499f6749fcc14fa>.

European Commission 2005, *Towards Smart Power Networks: Lessons learned from European research FP5 projects*, viewed 03 December 2012, <http://ec.europa.eu/research/energy/pdf/towards_smartpower_en.pdf>.

Farhangi, H 2010, 'The Path of the Smart Grid', *IEEE power & energy magazine*, vol 8, no. 1, pp. 18-28.

Ferreira, J, Pereira, P & Filipe, P 2011, 'Recommender System for Drivers of Electric Vehicles', *Electronics Computer Technology*, Institute of Electrical and Electronics Engineers, Inc., Kanyakumari, India.

Figueiredo, V, Rodrigues, , Vale, & Borges Gouveia, 2005, 'An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques', *IEEE Transactions on power systems*, vol 20, no. 2, pp. 596-602.

Hawarah, L, Ploix, S & Jacomino, M 2010, 'User Behavior Prediction in Energy Consumption in Housing Using Bayesian Networks', *Artificial intelligence and soft computing: Part I*, Springer Berlin Heidelberg, Zakopane, Poland.

Karlsson , K 2011, *Svensk Energi: El till vardags*, viewed 11 January 2013, <http://www.svenskenergi.se/sv/Om-el/Elanvandning/El-till-vardags/>.

Khattak, AR, Khan, GM & Mahmud, SA 2012, 'The Power to Deliver: Trends in Smart Grid Solutions', *IEEE Power and Energy Magazine*, vol 10, no. 4, pp. 56-64.

Kosiantis, S & Kanellopoulos, D 2006, 'Discretization Techniques: A recent survey', *GESTS International Transactions on Computer Science and Engineering*, vol 32, no. 1, pp. 47-58, <Doi: 10.1.1.109.3084>.

Kreider, JF & Haberl, JS 1994, 'Predicting hourly building energy use: The great predictor shootout - Overview and Dscussion of Results', *ASHRAE Transactions*, vol 100, no. 2, pp. 1104-1118, <http://www.techstreet.com/ashrae/products/1716675>.

Li, X & Chang, S-K 2005, 'A Personalized E-Learning System Based on User Profile Constructed Using Information Fusion', *In Proceedings of the 11th International Conference on Distributed Multimedia Systems*, Banff, Canada.

Litos Strategic Communication 2008, *The Smart Grid: An Introduction*, viewed 18 November 2012, <http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/DOE_SG_Book_Single_Pages%281%29.pdf>.

Liu, H, Hussain, F, Tan, CL & Dash, M 2002, 'Discretization: An enabling technique', *Data Mining and Knowledge Discovery*, vol 6, no. 4, pp. 393-423, <Doi: 10.1023/A:1016304305535>.

Liu, H, Hussain, F, Tan, CL & Dash, M 2002, 'Discretization: An Enabling Technique', *Data Mining and Knowledge Discovery*, vol 6, no. 4, pp. 393–423.

Logothetis, A 2012, *di.se*, viewed 01 November 2012, <http://www.di.se/artiklar/2012/9/26/spara-pa-att-kopa-el-per-timme/>.

MacKay, DJC 1994, 'Bayesian Non-linear Modelling for the Prediction Competition', *ASHRAE Transactions, V.100, Pt.2*, pp. 1053-1062.

Maher, ML, Zhang, DM & Balachandran, MB 1995, *Case-Based Reasoning in Design*, Lawrence Erlbaum Associates, Inc, New Jersy.

Mendes, E, Watson, I & Mosley, N 2002, 'A Comparison of Case-Based Reasoning Approaches to Web Hypermedia Project Cost Estimation', *Proceedings of the 11th international conference on World Wide Web*, ACM New York, NY, USA, Honolulu, HI, USA.

Prajapati, JK, Ghadiali, S & Vora, DR 2012, 'Smart Grid – A Vision For The Future', *International Conference on Advances in Engineering, Science and Management (ICAESM)*, Nagapattinam, Tamil Nadu, India.

Ramos, S, Vale, , Santana, & Rodrigues, 2006, 'Use of Data mining techniques to characterize MV consumers and to support the consumer supplier relationship', *SMO'06 Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization*, World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA ©2006, Lisbon, Portugal.

Riihimäki, H & Koponen, P 2012, 'Prediction of energy consumption from outdoor temperature for houses electrically heated via storage', VTT-R-02882-12, VTT Tecnical Research Centre of Finland, Espoo.

Samadi, P, Mohsenian-Rad, A-H, Schober, R, Wong, VWS & Jatskevich, J 2010, 'Optimal Real-time Pricing Algorithm Based on Utility Maximization for Smart Grid', *First IEEE International Conference on Communications (SmartGridComm)*, Gaithersburg, Maryland, USA.

Schiaffino, SN & Amandi, A 2000, 'User profiling with Case-Based Reasoning and Bayesian Networks', *International Joint Conference, 7th Ibero-American Conference, 15th Brazilian Symposium on AI, IBERAMIA-SBIA 2000, Open Discussion Track Proceedings on AI*, Springer, Atibaia.

Slade, S 1991, 'Case-Based Reasoning: A Research Paradigm', *AI MAGAZINE*, vol 12, no. 1, pp. 42-55.

SMHI 2009, *SMHI: Temperaturskillnader under dygnet*, viewed 07 January 2013, <http://www.smhi.se/kunskapsbanken/meteorologi/temperaturskillnader-under-dygnet-1.7250 >.

Sveriges riksdag 2012, *Betänkande 2011/12:NU21 Timmätning för aktiva elkonsumenter*, viewed 01 November 2012, <http://www.riksdagen.se/sv/Dokument-Lagar/Utskottens-dokument/Betankanden/Arenden/201112/NU21/>.

Tang, Y & Ciuciu, IG 2012, 'Semantic Decision Support Models for Energy Efficiency in Smart-Metered Homes', *IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, Liverpool, United Kingdom.

The European Commission 2008, *Energy efficiency: delivering the 20% target*, Belgium, Brussel, viewed 02 December 2012, <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2008:0772:FIN:EN:PDF>.

Thurtel, J & Feuston, BP 1994, 'Generalized nonlinear regression with ensemble of neural nets: the great energy predictor shootout', *ASHRAE Transactions, V.100, Pt.2*, pp. 1075-1080.

Tsai, C-J, Lee, C-I & Yang, W-P 2008, 'A discretization algorithm based on Class-Attribute Contingency Coefficient', *Information Sciences*, vol 178, no. 3, pp. 714-731, <Doi: 10.1016/j.ins.2007.09.004>.

U.S. Department of Energy Office of Electricity Delivery & Energy Reliability 2010, 'Multi-Year Program Plan', U.S. Department of Energy Office of Electricity Delivery & Energy Reliability.

van Setten, M, Veenstra, M, Nijholt, A & van Dijk, B 2004, 'Case-Based Reasoning as a Prediction Strategyfor Hybrid Recommender Systems', *Proceedings of the atlantic web intellegence*, Springer, Cancun, Mexico, <10.1.1.3.932>.

Watson, ID 1999, 'Case-based reasoning is a methodology not a technology', *Knowledge-Based Systems*, vol 12, no. 5-6, pp. 303-308.

Zahedi, A 2011, 'Developing a system model for future smart grid', *Innovative Smart Grid Technologies Asia (ISGT)*, Perth, Australia, <10.1109/ISGT-Asia.2011.6167326>.