

Recognizing and visualizing energy consumption patterns of buildings using data mining

University of Southern Denmark

Mærsk Mc-Kinney Møller Institut

Software Engineering

Bachelor Thesis



Almir Mehanovic, Exam no. 323335
Emil Sebastian Rømer, Exam no. 324790

01 June 2015

Supervisor: Mikkel Baun Kjærgaard

Supervisor: Sanja Lazarova-Molnar

Recognizing and visualizing energy consumption patterns of buildings using data mining

Bachelor Thesis

Course	SB6-BP-U1
Supervisor	Mikkel Baun Kjærgaard, Associate Professor, PhD Sanja Lazarova-Molnar, Associate Professor, PhD
Authors	Almir Mehanovic, Exam no. 323335 Emil Sebastian Rømer, Exam no. 324790
Project period	02-02-2015 to 01-06-2015
Submitted	01-06-2015
Institute	Mærsk Mc-Kinney Møller Institut Niels Bohrs Allé 1, D115 5230 Odense M

Almir Mehanovic

Emil Sebastian Rømer

Date

Date

Abstract

This report discusses the application of statistical methods and machine learning to better understand and manage energy consumption in buildings. The report is the result of a project which was conducted in association with the Municipality of Odense. The project's focus was on creating a building energy management system which creates value by solving a number of problems for the municipality.

The resulting software system is a website which presents information regarding heat, water and electricity consumption in the municipality's buildings. The website enables energy managers to easier maintain information about buildings, detect faults, avoid leaks and compare buildings in respect to their consumption. The final system is a working prototype written in R, html, JavaScript and Java and developed using a user centered iterative process.

This report will evaluate the statistical methods that were used, and the developed system's performance for detecting faults and finding buildings with similar consumption patterns. Additionally the systems usability will be evaluated on basis of a user test with a representative from the municipality's Energy and Maintenance department.

The evaluation of the system shows, that the system is capable of achieving good results in respect to fault detection and finding similar buildings. Additionally a user test shows that the system extracts useful information about consumption patterns, which makes many tasks for energy managers in the municipality easier, and enables them to make more informed decisions.

Foreword

This paper is the result of a 6th semester Bachelor project in Software Engineering at The University of Southern Denmark. The project was conducted in the period 2/2/2015 – 1/6-2015.

The group wants to give a special thanks to our supervisors and staff at the Municipality of Odense.

- Kim Allan Kristensen, From Odense Municipality's Energy and Maintenance Department
- Mikkel Baun Kjærgaard, Associate Professor, PhD (Supervisor)
- Sanja Lazarova-Molnar, Associate Professor, PhD (Supervisor)

Further a thank to the municipality of Odense for participating and for making the data available, and to Lone Borgersen, Associate Professor, Educational coordinator Software Engineering, for coordination, planning and organization of the bachelor semester

Contents

Abstract	ii
Foreword	iii
Reading guid	v
1 Introduction	1
2 Background	3
2.1 Current limitations	4
3 Requirements	7
3.1 System and data maintenance	7
3.2 Visualization of buildings and consumptions	8
3.3 Fault detection, diagnosis and resolution	8
3.4 Budgeting	9
3.5 Selecting buildings for renovation	9
3.6 Other requirements	9
3.7 Prioritization of requirements	9
3.8 Overall system description	10
4 Related Work and Methods	12
4.1 Decomposing time series	12
4.2 Cluster analysis	13
4.3 Detecting faults	15
5 Design and Implementation	17
5.1 Development approach	17
5.2 Overall architecture	17
5.3 Integration with Energy Key	19
5.4 The central database	21

5.5	Processing data	24
5.6	Creating consumption profiles	25
5.7	Comparing buildings	27
5.8	Detecting faults	28
5.9	Design of the website	30
5.10	Implementation of website	32
6	Test and evaluation	35
6.1	Technical evaluation	35
6.2	Usability evaluation	40
6.3	Utility and efficiency of the product	42
7	Conclusion	45
7.1	System and data maintenance	45
7.2	Visualization of buildings and consumptions	45
7.3	Fault detection, diagnosis and resolution	46
7.4	Budgeting	46
7.5	Selecting buildings for renovation	46
7.6	Project overall	47

Reading guide

This report is written to be read in chronological order, and in some cases additional information is referred to in appendices, which will be mentioned in the text. All appendices can be found on the accompanying CD. Additionally the CD contains all source code for the developed software system. Along the appendices the CD contains the source code for the final product. This is placed in the folder "Source Code". The website can further be tested by opening the Main.html file located in "Source Code"/Website. The website have been tested using FireFox

In this paper OM has been used in short for *Odense municipality*. The developed software system has been named *BEOVulf* as short for Building Efficiency Optimizer and Visualizer. This name will be used throughout the report.

Chapter 1

Introduction

The developed world's dependency on fossil fuels to power growth and prosperity is becoming an increasingly unsustainable and unfeasible situation. The world will eventually run out of fossil fuels, and the dependency on it means that vital infrastructure is depending on politically unstable regions. What is perhaps even more concerning is the damage our environment suffers from the burning of fossil fuels. Energy efficiency is a major step away from the dependency on fossil fuels, and towards a society powered by economically and socially sustainable energy.

In March 2012 a new Energy Agreement was reached in Denmark, with goals to have approximately 50% of electricity consumption supplied by wind power, and more than 35% of final energy consumption supplied from renewable energy sources in 2020. The final goal is to have 100% renewable energy in the energy and transport sectors by 2050¹.

Buildings in Denmark have over the last decades become more energy-efficient as a result of advances in engineering and the ongoing strengthening of requirements for new buildings. However, around 40% of the total energy consumption is used in buildings, making them the largest contributor to energy usage in Denmark². Therefore retrofitting buildings with energy efficient technology is important, if Denmark is to fulfill the energy agreement and reach its goals. The field of Energy Informatics has recently attracted much attention as large energy efficiency improvements remains to be realized, by utilizing the field to better understand, predict and optimize energy consumption efficiency in buildings³. This field is a cornerstone for realizing the visions of smart buildings and smart cities. In southern Denmark, the Municipality of Odense (henceforth referred to as OM) has come one step closer to these visions, by putting data regarding energy consumption of approximately 550 public buildings on the roadmap to becoming Open Data⁴. This data contains hourly measurements of raw consumption for electricity, heating and

¹See Appendix 10 -Energy Efficiency Policies and Measures in Denmark 2012; Page 3.

²<http://www.kebmin.dk/klima-energi-bygningspolitik/dansk-klima-energi-bygningspolitik/byggeriet-danmark/bygningers>

³Energy Informatics -DOI 10.1007/s12599-013-0304-2

⁴<http://odensedataplatform.dk/>

water in a building. The availability of such data is a crucial step, but the raw data is of little use, unless it is transformed into useful information that can be represented in an intuitive manner to various stakeholders.

This project has worked with OM to find potential use cases of the consumption data, and explore what kind of useful information can be extracted from the available data. The aim was to develop a functioning prototype website capable of highlighting various trends, correlations, odd events, odd buildings etc. The scope for this project was not limited to researching and theoretically proving the applicability of various data processing methods, but included applying statistical methods to solve real problems for OM. Thus the challenges faced by this project included those of a software development project, in which a software system suiting a client's needs had to be developed. The project utilized an iterative development approach, to identify OM's needs in respect to building energy management and design and evaluate the final software solution.

Chapter 2

Background

Odense Municipality (OM) is the third greatest in Denmark with an size of 304 km^2 and a population of nearly 200.000 people. In order to reach the goals of the national environmental policy, OM has created the directive “Environmentally friendly construction - requirements and recommendations¹”, which as the directive states will: “*Concretise the environmental policy and objectives, so that future building projects in the municipality to the up-filling of the overall vision of Odense as Denmark’s most sustainable city.*”². This directive sets energy efficiency requirements for newly constructed and renovated buildings.

OM owns around 850 buildings, which are used for different purposes like schooling, administration, care centers and cultural institutions. The municipality is responsible for paying suppliers of water, heat and electricity for the consumption in these buildings. This is done on basis of billing meters which measure the consumption, and are installed in each building. Depending on the size of the building, there might be a number of internal meters (so called distribution meters), which measure the consumption in different parts of the building, but are not used as billing meters.

Each public building has an appointed energy manager who has the responsibility of performing the manual tasks involved in reading and checking the meters. The energy manager logs reading internally in a control book or scheme and reports them on the last weekday of the month through a web portal. If reading deviate more than $\pm 10 \%$ from the budget forecasts, the manager must submit possible causes for this. As the document “Energy efficiency in the municipality of Odense” states, both types of meters must be often checked to avoid energy leakages. In buildings less than 1.000 m^2 , meters must be checked at least once a week and for larger buildings the frequency is at least once a day. OM has started installing meters which automatically report their readings on an hourly basis. Those are not yet installed in all buildings, but OM has plans to do so. Depending on the concrete building and setup there is a delay from around 1 to 7

¹See appendix: Appendix 01 -Environmentally friendly construction

²See appendix: Appendix 01 -Environmentally friendly construction

days before readings from the automatic meters are visible in OM's energy management system 'Energy Key'³.

2.1 Current limitations

Through interviews with a representative from OM's Energy and Maintenance department, the needs for a building energy management system have been investigated. This analysis revealed a number of areas in which improvements could be made. Before 2013 each building's energy consumption was monitored on site by appropriate staff, and energy bills were paid from the budgets of the organization residing in the building. However, building managers on site had no real tools or means of monitoring their consumption, making it difficult to identify high usage or leakages. This function has now been centralized in the department of Energy and Maintenance which owns the buildings and is responsible for monitoring and paying for energy consumption, as well as renovation and maintenance of the buildings. The organizations residing in the buildings will then just rent the building for a fixed price. This centralization should give OM a better control and overview of the municipality's energy consumption, but a number of new issues have arisen. The task has scaled with the number of buildings that now needs to be monitored, and the problem analysis showed that OM's Energy and Maintenance department does not have the needed staff nor the proper tools to complete it effectively. The problem analysis resulted in the identification of 5 main issues, which has been the focus points of this project.

System and data maintenance

Maintaining accurate data about a building's size, age, type, associated meters etc. can seem like a seemingly small effort, but has scaled to become an overwhelming task when considering the amount of buildings. If the primary data is imprecise, erroneous or inconsistent, the value of data analysis and anything that depends on it will decline. OM's current energy management system Energy Key has no way of telling the user when certain data has last been verified and by whom, nor is there any easy way for users to update incorrect data entries. This makes it difficult to work with the data, because the users are often unsure about the validity of it, and it causes redundant work because multiple users need to verify the same data. There are currently no procedures for how often a building's consumption should be checked and validated, nor does Energy Key support such continuous maintenance tasks in any way. Additionally OM struggles with poor data quality and many missing values regarding consumption figures, which according to OM is because the energy suppliers' systems are simply not built for sharing data with clients.

³<http://www.emtnordic.com/da/energykey-energistyringsprogram>

Visualization of buildings and consumptions

Energy Key allows users to view electricity, water and heat consumption data for individual meters, but the possibilities to visualize and represent data in meaningful formats seems limited. The user can view diagrams depicting consumption over time, and histograms that compare a building's reported consumption with the budget. Other than this, the system does not summarize important key features and energy statistics for a given building. It does not analyze the data to provide new insights nor does it provide any tools for users to do so. The lack of meaningful information and visualization of consumption patterns makes it hard to understand a building's energy usage. This makes it more difficult to motivate users and building managers into making better decisions that lead to energy savings, because they do not know how their behaviour influences the consumption.

Fault detection, diagnosis and resolution

Energy Key has an automatic fault detection feature which can raise an alarm when consumption goes above or below a given percentage of what is budgeted. OM's Energy and Maintenance department has tried experimenting with this system for a single building, and the results were that 37 false alarms were received in 2012 and 34 in 2013. When considering all buildings the amount of false alarms becomes overwhelming. The large amount of false alarms combined with the fact that tasks involved in diagnosis and resolution of faults are not supported in Energy Key, has lead to the alarms being completely ignored. Even if a user can verify that an alarm is valid the fault might be neglected, as there are no standard procedures to follow in order to resolve the fault, nor is there any personnel assigned as responsible. Energy Key simply does not support users in coordinating their efforts towards the resolution of alarms.

Budgeting

Currently energy consumption and budgets for each building are estimated by looking at the same building's consumption in previous years. Estimations have monthly granularities, meaning that the energy consumption estimate for a given month depends on the consumption reported the same month previous years. This approach does not allow one to properly predict consumption under changing circumstances (people usage, environment, building changes etc.). A better model for a building's energy consumption would allow for better predictions and more exact budgets. When information about a building is changed in Energy Key, the previous information is lost making it is difficult to use the previous consumption patterns to plan future budgets as there is no historic data to show under which conditions the previous consumption was produced. Because of the issues with maintaining data in Energy Key, information like for instance how many people are using a building is stored and maintained in excel files.

Selecting buildings for renovation

In 2015 OM's Energy and Maintenance department has budgeted 200 million dkk to renovation of public buildings. As the *directive Environmentally friendly construction - requirements and recommendations*⁴ says all renovations must be done with the most energy efficient solutions. The department now faces the task of finding the set of buildings for which renovations would bring down energy usage the most. Currently a taskforce is screening through the available data manually, and more or less takes educated guesses about which buildings are inefficient and can be optimized. The current system is unable to perform any sophisticated benchmarking or ranking of buildings, making it hard to verify that the most energy inefficient buildings have been found. However, identifying energy inefficient buildings is only the starting point, as a vast range of economical and political parameters also have to be considered before a building is finally selected for renovation.

⁴See Appendix 01 -Environmentally friendly construction

Chapter 3

Requirements

Based on the initial problem analysis and identification of OM's needs, a quick low fidelity horizontal prototype system was created¹. This prototype was evaluated through an interview with a representative from OM's Energy and Maintenance department². Based on the prototype evaluation and the problem analysis a final backlog of requirements was specified. The full requirements list can be found in Appendix 05 -Requirement Specification, while this section will give an outline of how the requirements solve the identified issues.

3.1 System and data maintenance

Any client of a new software system would require that the servicing efforts are minimal, but the problem analysis uncovered that for OM this is of critical importance. OM's Energy and Maintenance department has had a large increase in workload without a proportional increase in resources or personnel. This means that the department simply does not have the time or resources to maintain a new system if the required servicing effort is too high. A number of steps to prevent this have been taken, resulting in formal requirements which can be measured and evaluated.

Firstly, in long term, the new system should replace OM's existing system, but changes in an organization's processes, workflow and technology comes with the expense of time and money. In order to minimize the expenses, a gradual adoption of the new system will be in focus. Requirement *R18*, ensures data consistency in such an adoption phase where multiple systems work on the same data. This ensures further, that maintenance in an adoption phase will remain the same as in the existing system.

Secondly, to minimize the effort needed to maintain large amounts of building information (such as type, age size and so forth), requirement *F2 – 06a* and *F2 – 06b* make sure it is possible

¹See Appendix 04 -Website Alpha Prototype.zip

²See Appendix 02 -Meeting with Kim Allan 29-4-15

to easily update and validate these general building informations. The effort associated with maintaining the data is minimized by informing the user when data was last validated and if it is time to re-validate. Thus the user needs not spend time investigating validity of data every time it is needed. Also as opposed to the current system it will be possible to easily update data when a building is viewed anyway, and then offering the user the choice to validate it now. This solution ensures that even building managers for the different buildings will participate in keeping the information correct. This is additionally supported with requirement *F15* and *F17*, which will allow the assignment of maintenance tasks to the building managers thus distributing the maintenance effort.

3.2 Visualization of buildings and consumptions

Requirements *F2* and *F9* together with the respective sub requirements state that it should be possible to view and compare information about buildings and their energy consumption. The requirements formulate a number of concrete aspects of the energy consumption that must be visualized to the user. In contrast to OM's existing system this will give users better information and support better decision making. Requirement *F2 – 02* gives concrete examples of a number of energy features that must be extracted and visualized from the raw consumption numbers. Those are for instance peak loads, standby usage and daily average consumptions.

Besides pure consumption figures, the visualization of consumption patterns can further enhance the understanding of a buildings consumption. The consumption patterns of a building may be viewed in different levels of granularity, like for instance patterns on a micro scale (daily) and patterns on a macro scale (yearly). The three requirements *F2 – 03a* to *F2 – 03c* describe how consumption patterns on a micro scale are to be visualized through so called daily consumption profiles. Requirement *F2 – 04* require that a buildings seasonal variation in consumption is visualized which might be considered a macro scale pattern.

From looking only at the consumption of a single building without a context or reference point, it can be hard to identify faults or even verify that everything is as it should be. Therefore requirement *F9* and it's sub requirements aim to give the user a way to visualize a building in comparison with others, to have a reference point when evaluating a building³.

3.3 Fault detection, diagnosis and resolution

Requirement *F3*, addresses the fault detection, and here the aim is to build a fault detection system which is more accurate and allows for diagnosing the feature which caused the flaw.

Requirements *F3 – 02*, *F3 – 03* and *F12* (handling of alarms) will try to make sure alarms are not ignored or forgotten by establishing a process/workflow for how they should be handled. The

³Requirement *F9* and comparing buildings is further discussed in section 3.5 Selecting buildings for renovation

system will provide an easy overview of alarms, and requirements regarding task management (*F15 – 17*) will contribute to the creation of a “standard” procedure of handling alarms.

3.4 Budgeting

Requirement *F14* states that the system should implement functionality for creating and maintaining a budget. The system must support the user in creating precise budgets, by being able to accurately forecast the expected consumption of a building as stated in requirement *F13*.

3.5 Selecting buildings for renovation

Selecting buildings for renovation is a decision process with many political and economical parameters which would be impossible for one system to fully implement. However the system will be able to support the process by giving users the ability to quickly identify potential candidates. This is done by implementing requirement *F11*, which will make the system come up with various rankings of buildings, making the worst ranked buildings potential candidates for renovation. It will be considered to map the benchmarks of buildings to the national energy labeling scheme⁴.

3.6 Other requirements

The environment in which the application is to be used, spawns quality requirements for the systems portability and scalability which needs to be met in order to develop a successful application.

The users from OM’s Energy and Maintenance department work both from their offices and from various locations when visiting buildings to inspect. This need is met by requirement *F20*, which addresses the portability of the system and states the application must work on a tablet.

This project has focused on the municipality of Odense, but other municipalities in Denmark could potentially have similar requirements for an energy management system. Based on this, there is a requirement that the application is scalable. Another scenario is that the system is expanded to not only public buildings. In that case the number of buildings handled would be an order of magnitude more, and the system should be able to handle this⁵.

3.7 Prioritization of requirements

All requirements have been prioritized using the MoSCoW-prioritization technique. The prioritizations are based on what is within this project’s scope and boundaries. For instance a task

⁴<http://www.ens.dk/forbrug-besparelser/byggeriets-energiforbrug/energimaerkning>

⁵See requirement *Q19*

management system is out of scope for this project and therefore requirements *F16* and *F17* have been prioritized as *WON'T* have. If a finished product was to be made, these requirements are essential for the user and should therefore be implemented.

3.8 Overall system description

Based on the different requirements and on the prioritization of these, the use case diagram seen in figure 3.1 has been developed. The uses cases have been organized into two groups in order to group similar functionality. The diagram shows all the use cases implemented by the developed system. In relation to this, one should note, that the use cases included in this diagram does only extent to the functionality directly linked to the end user. As an example the requirement for integrating the system with the existing system have not been included, even though it will be implemented.

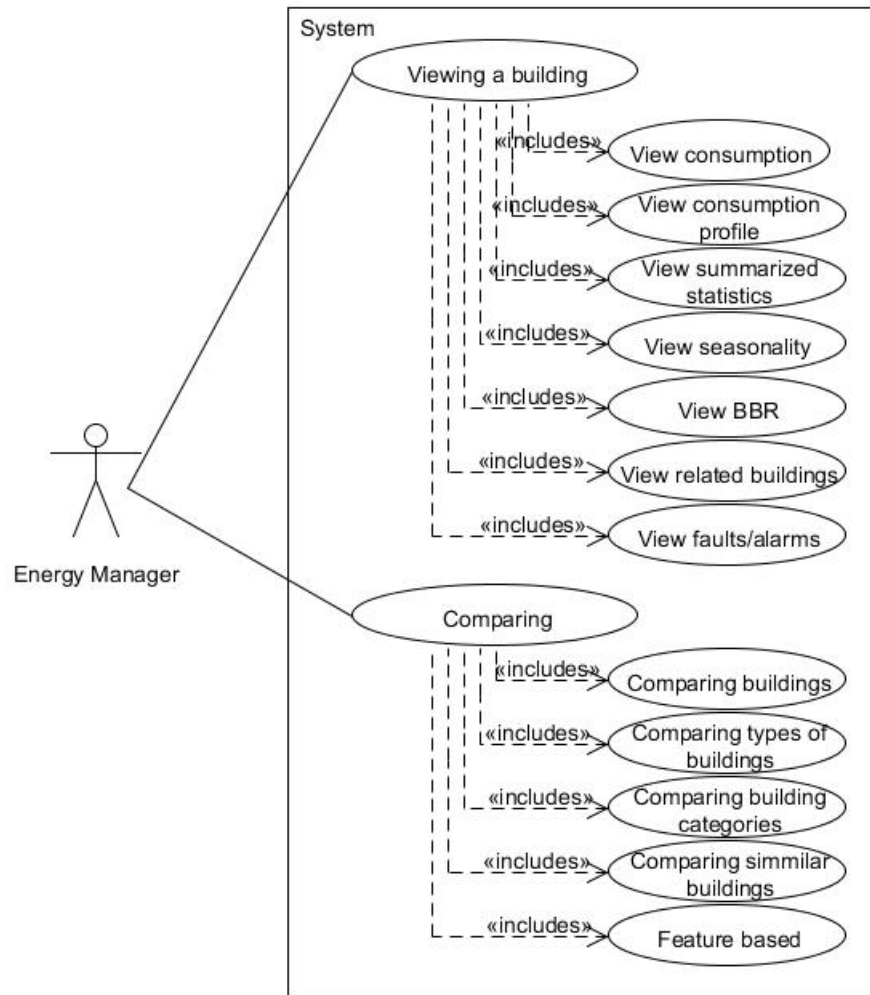


Figure 3.1: Usecase diagram

Chapter 4

Related Work and Methods

Within recent years numerous research papers have emerged showing successful applications of various methods to problems like detecting faults, benchmarking, classifying, predicting or recognizing patterns in the energy consumption of buildings. Most of the applied methods rely on theories rooted in the fields of statistics, machine learning and data mining to analyze and interpret the data. To successfully apply such methods to real word problems one must be familiar with the nature of the data, and have working knowledge of the methods being applied. This section will explore some of these methods and the current literature that applies them to various problems related to building energy management.

4.1 Decomposing time series

The most visible pattern that emerge from time series data of building energy consumption is the daily rise and fall in consumption, which loops every 24 hours following the daily cycles of building users. Also a weekly cycle can be noticed, as different days of the week tend to have different consumption profiles. If looking at several years of data, the yearly cycle following the seasonal changes in environmental conditions can be seen. These cycles form the basis for the components that can be found when decomposing a time serie, and which could be used to reconstruct the original by combining the components. For instance considering the weekly cycles and yearly seasonal change, it would be possible to make a decomposition into a Cyclical Component, and a Seasonal Component¹. Additionally the decomposition can produce a Trend Component and an Irregular Component. The Trend Component reflects the long term progression irrespective of any cycles and seasonality, while the Irregular Component is the noise in the original time serie which could not be explained with the other components. Figure 4.1 shows an example of a decomposition.

¹http://en.wikipedia.org/wiki/Decomposition_of_time_series

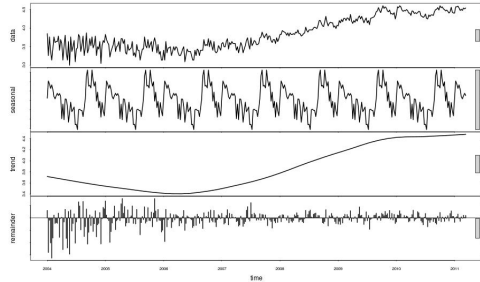


Figure 4.1: Decomposition of time series

The example illustrates how decomposing time series can give more insight into the nature of the data. For instance it can be seen that the consumption is generally going up each year, and that consumption is higher in the winter than in the summer. Decomposition is relevant when comparing consumption between different days, because it might be beneficial to first remove the seasonality component. There are however many ways to adjust for seasonality. For instance [3] suggests subtracting from each day the minimum consumption for that day, while [5] transforms the features of one day by determining the difference between it and a the features of a 1-week average of surrounding days.

The different components of a decomposition can serve as different units of analysis, and are therefore important to consider when extracting features from the data. Whatever forecasting consumption load, detecting abnormal consumption or something else is desired, most approaches agree upon that the first step is to extract interesting features from the data. This might be raw consumption figures (daily average, daily max/min), ratios between features (consumption night/day), temporal properties (time of daily max, first time above 1kW) or statistical properties (variance, correlation between features). [1] brings a list of potentially interesting features when considering energy consumption data, while [7] suggests a range of interesting building properties to consider.

4.2 Cluster analysis

Cluster analysis is widely used within data mining and machine learning, to find underlying structures and patterns in data. [5] proposes a pattern recognition algorithm for determining days of the week based on the raw consumption data, and explains how this information can be used for both forecasting consumption load and detecting abnormal consumption. By creating a cluster for each day of the week and then identifying similar clusters based on various features of daily consumptions, it is possible to come up with a number of ‘day types’ with distinct consumption profiles. The type of day and consumption profile can then be used as input to

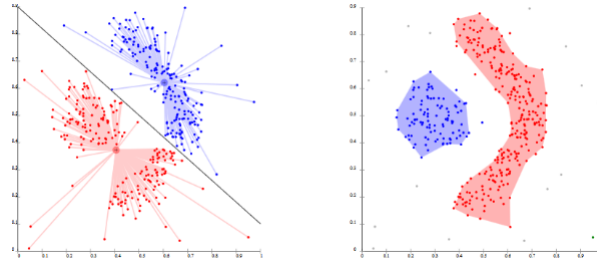


Figure 4.2: *Note these pictures have been taken from Wiki -Cluster analysis*

forecasting algorithms, or as a model of normal consumption from which a large deviation can be considered an anomaly. [3] uses a very similar approach.

There are many different techniques for clustering, and even more algorithms implementing them. Some popular ones are the the centroid-based K-means algorithm and the density- based DBSCAN. The K-means algorithm fixes the number of clusters to k (a parameter given to the algorithm). The algorithm will assign each data point in the dataset to the nearest cluster center, and then seek to find the optimal positions for the k cluster centers, such that the squared distances from each object to the cluster centers are minimal. K-medoids clustering is very similar to k-means, except that while the cluster is represented by a centroid in k-means, it is represented by the data point closest to the center of the cluster in k-medoids. A drawback of K-means and K-medoids clustering is that they do not work well with non convex clusters, which might be present if data point features do not follow normal distributions. This problem is avoided in DBSCAN which creates clusters by connecting points which have a specified minimum number of other points within a specified radius. Based on this density criterion clusters of arbitrary shapes can be formed. Figure 4,2 illustrates how cluster analysis with K-means and DBSCAN can result in very different clusters for the same dataset. DBSCAN additionally has the ability to detect outliers, which would be the data points residing in low density areas not reachable with the density criterion. [4] compares the use of DBSCAN alone with using K-means followed by the Generalized Extreme Studentized Deviate (GESD) outlier detection method, to find outliers and detect faults in building energy consumption. The results however seem somewhat inconclusive.

[1] and [7] demonstrates the use of SOMs (Self Organizing Maps) to perform cluster analysis of building energy consumers. [1] finds clusters of energy consumers by training a SOM using raw energy consumption data, while [7] uses building information as basis for the clustering. The features chosen to use as input for the clustering, are just as important as the choice of clustering method. Too many features might introduce too much noise, so it might be beneficial with some dimensionality reduction before performing cluster analysis. Using too few features however might not properly separate the clusters. A dimensionality reduction can be achieved with a

principal component analysis. By using only the principal components that explain the majority of variation, one would have found the best low-dimensional representation of the variation in a multivariate data set. [3] argues that a better approach to use when classification is the purpose is CVA (canonical variate analysis), also known as linear discriminant analysis (LDA). The purpose of LDA is to find the linear combinations of the original variables that give the best possible separation between the groups. Thus LDA is used with day of week as the grouping variable, to project the original data into new axes which maximize separation between different days².

4.3 Detecting faults

Detecting abnormal energy consumption seems always to follow the basic steps of building some kind of model of normal consumption and then detecting faults with an appropriate outlier detection method [2], [3] and [4]. Another approach would be to use supervised learning to train a classifier that can distinguish faulty consumption from normal. However this would require a proper training set with already identified and labeled errors for each building. The lack of such datasets and the fact that faulty consumption is a fairly rare event, makes the outlier detection approach the favoured one. Apart from using outlier detection methods to detect abnormal consumption, it is used as a preprocessing method to remove data points which are so unusual that they could indicate a measurement error.

An outlier can be defined as data which appear to be inconsistent with the rest of the data set, like for instance a numerical value which is significantly distant from the rest of the data set. There is no mathematically rigid method for defining an observation as an outlier, and many methods to detect outliers both in univariate and multivariate data sets have been proposed. If the data is approximately normally distributed, univariate outliers can easily be identified with a hypothesis test calculating the p-value which is the probability obtaining a result equal to or more extreme than what was actually observed. [3], [4] and [5] all make use of the GESD generalized extreme studentized deviate method for univariate outlier detection. This method works by doing Grubbs' test which is defined for the hypothesis:

H_0 : There are no outliers in the data set

H_A : There is at least one outlier in the data set

If there is an outlier it is removed and the test is performed again, until a specified upper limit is reached or there are no more outliers in the data set. Besides an upper limit on the number of outliers a significance level α must be specified which would translate into the probability of getting a false positive.

²A more in depth explanation of PCA and LDA can be found in: <http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html>

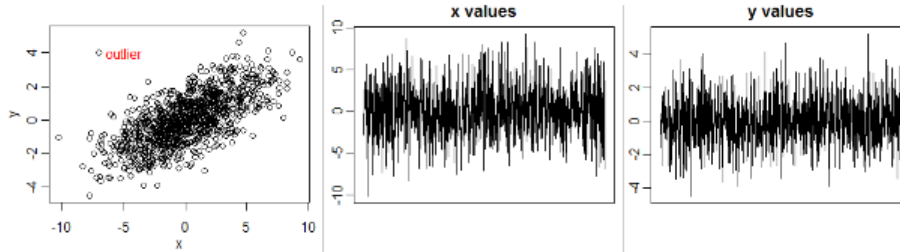


Figure 4.3: Multivariate outliers

In [3] the GESD procedure is applied to each feature of an observation to detect an outlier. However as figure 4,3 illustrates, an outlier in a multidimensional dataset is not always an univariate outlier. Considering the x and y values independently does not obviously reveal the outlier flagged in the first graph. To deal with multivariate outliers one needs a proper method capable of considering the correlation between features. One such method is PCOut [6]. PCOut is an advanced algorithm which utilize inherent properties of principal components decomposition for outlier detection in high dimensional data. It does not require that the data come from any particular distribution, and is computationally fast.

Chapter 5

Design and Implementation

5.1 Development approach

The system was developed using an user centered iterative approach inspired by the development model of Rogers [9], see figure 5.1. This approach puts weight on prototyping and interactions with the client to both identify needs and evaluate prototypes. Thus 3 meetings with a representative from OM was conducted during the two development iterations.

5.2 Overall architecture

In order to develop a system that fulfils both the functional and nonfunctional requirements, it was decided to develop a web-application in the form of a website. Making the system work as a website makes it highly portable as most client systems have browsers. The system thus becomes a distributed system with a very thin client. The thin client distributed pattern has low complexity as only the presentation layer needs to be distributed, while processing and data is centralized at the server. This ensures data consistency, and allows for a simple and efficient

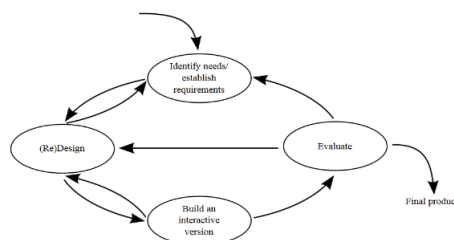
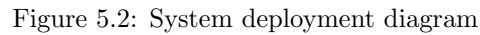


Figure 5.1: Rogers development process



Part 2 consists of a number of R² scripts, which will read the raw meter data from the database and perform a number of operations to convert the data into the kind of information

²<http://www.r-project.org/>

that can fulfill the requirements. This includes extracting various features from the consumption, calculating consumption profiles, detecting faults and leakages, comparing buildings and more. Section *Processing Data* describes the details of this process. The results of these operations are stored back to the database, where they can be accessed by Part 3 of the system.

Part 3 is the website that users will see, which is hosted in an Amazon s3 bucket³. The website is powered purely by html and Javascript and might be considered static⁴. In order to connect the website with the central database, Dreamfactory has been used. It is a framework for automatically creating a RESTful web service which allows front-end applications to easily connect to the backend. Dreamfactory⁵ exposes the database via a REST API, which uses JSON to transfer data to and from the database. The communication with the database is further simplified by making the website use the Javascript client library provided by Dreamfactory. The design and implementation of the website will be discussed in sections *Design of the Website* and *Implementation of the website*.

5.3 Integration with Energy Key

The integration with Energy Key was done in four main steps. First, a number of exports from Energy Key were made to gather data about buildings their attributes and which meters are installed in them. Then in step two a year worth of measurements from the meters were exported. In step three the data was cleaned to sort out inconsistencies. And finally in step four a continues flow of data from Energy Key to BEOVulf was set up.

Step 1 exporting buildings and meters

The following reports were exported from Energy Key and used to map buildings and meters:

- Meter list⁶ (Energy Key → Konfiguration → Målere → Generer rapport)
- Measurement Reading Report⁷ (Energy Key → Aflæsninger → Visning → Generer rapport)

From these reports it was possible to generate a list of all buildings along with a number of attributes such as id, address, name, type (ex. ‘*cultural institution*’) and subtype (ex. ‘*library*’). For each building it was possible to attribute a number of meters which measured either water, electricity or heat. For each meter it was possible to get a installation number, meter number and whatever it was a so called *billing meter* or *distribution meter*. This attribute is important, because the readings from a *distribution meter* can not be classified as part of the consumption.

³<http://aws.amazon.com/s3/>

⁴http://en.wikipedia.org/wiki/Static_web_page

⁵<https://www.dreamfactory.com/>

⁶Translated from: Målerliste

⁷Translated from: Aflæsnings Rapport

-	Heat	Water	Electricity
Total meters	1282	849	709
Billing meters in %	45%	67%	97%

Figure 5.3: Meters and distribution of type

Step 2 - Exporting a year of measurements

The initial export of measurement data was limited to go one year back, as the data quality declined rapidly when going further back. The exported data was in csv files, one for each meter. The files were thoroughly analysed to determine the structure of the data. Then a Java program was created which could read through the files and insert the measurements in the central database, following the structure described in section *The central database*. The parser made sure that data was consistent, it translated between different units, and if some data did not fit with the designed structure, the program stopped with a warning. The error was examined manually and it was decided what to do with it. Then the parser was changed, the database wiped, and a re-run initiated. This last step was repeated until the parser successfully parsed all 3003 csv files with more than 5.5 million measurements.

Step 3 - Detecting errors and cleaning data

When exporting building data from Energy Key there were unfortunately many buildings with missing attributes, which could not be automatically gathered from anywhere since automated access to the BBR registry is prohibited⁸. Buildings where no type and subtype existed were removed along with all meters belonging to them, while other missing attributes were tolerated. The final amount of buildings in the database was 405. Meters which did not have both installation and meter number were removed, along with meters that had no reported data in the whole exported period. The cleaning reduced the number of meters 33% from 4293 to 2840 meters. Figure 5.3 shows a table of the final amount of meters stored in the central database:

Upon further analysis it was detected that the non automatic meters often had large numbers of overlapping or missing intervals. They were not removed as use cases for them might be found, but data from the non automatic meters was simply too unreliable to be used for advanced calculations and analysis of the building consumption. Therefore a view was created in the database which could filter out non automatic meters, by looking at whatever they have hourly readings or not.

⁸<http://bbr.dk/fatibbr>

Step 4 - Continuous meter data collection

After a year's worth of measurements was initially loaded into the database, a daily export job was created in Energy Key. This export job would automatically export csv files containing the daily measurements to BEOVulf's FTP-server located on the Amazon ec2 instance. The Java program used to parse the initial batch of measurement data was extended to be able to listen for changes in the folder used by the FTP-server. Once a csv file is fully uploaded to the folder, the Java program parses it, puts the measurements contained in it in the database and then archives the csv file. Thus observations from buildings' meters would come into the database on a daily basis.

5.4 The central database

This section will describe the design of the central database in two parts. The first part is concerned with how the raw data coming from Energy Key is stored, and the second part is about the processed data.

Raw data

Figure 5,4 shows an UML diagram of the part of the database, where raw consumption and building data is stored. The main criteria for this design was to be able to store all the data retrieved from Energy Key efficiently without losing any information. The database structure was modeled in a way which is generic enough to capture all kinds of observations without needing a table for each observation type. This ensures that all kinds of meter data that comes from Energy Key can be stored, and the design greatly reduces the number of tables needed. It makes sure that the database complies with the first three normal forms of databases, and the structure makes it possible to treat all measurements the same way when processing data.

Buildings and BBR data

The database contains a table called buildings. Each building has a building_id, type and subtype which is taken directly from Energy Key. The BBR-table represents the BBR-data collected for each building. This data contains variables like address, building area, site area, cellar size etc., but for most buildings the majority of the attributes are unfortunately missing values.

Meters and time series

When analysing the data, it was found that some buildings had the same physical meter appearing under two or more different installation- and meter numbers, and causing duplicated measure-

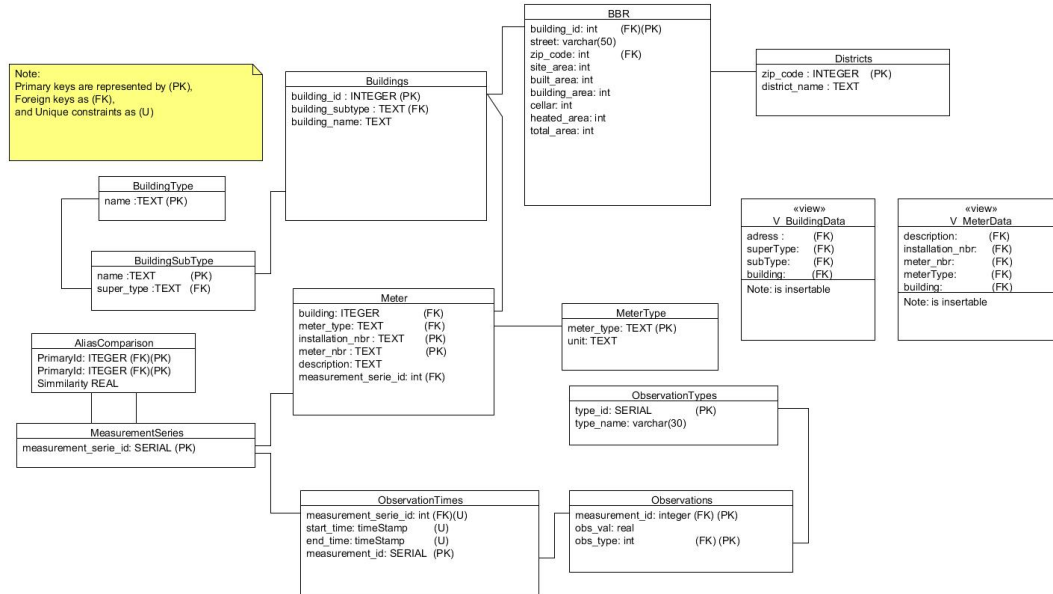


Figure 5.4: Database diagram

ments. The solution was to introduce the *MeasurementSeries* table which represents data from one real physical meter and then map multiple meters from the *Meter* table to the same measurement serie if they in fact measure the same consumption. Additionally the *AliasComparison* table was created which through a number of triggers and procedures can calculate the similarity between time series of different meters. This enables the database to identify identical meters and dynamically remap meters and series eliminating duplicate time series on a building.

Observations

The table *Observations*, contains all the measured values from all meters. Each record in *Observations* points to a record in *ObservationType* indicating its type, and to a record in *ObservationTimes* indicating the timeslot in which the consumption was measured. It is possible for many observations thus to point to the same time slot. For example the following observation types can be recorded for a certain time slot when data from a heat meter arrives: in- and outgoing temperature, energy, flow and more.

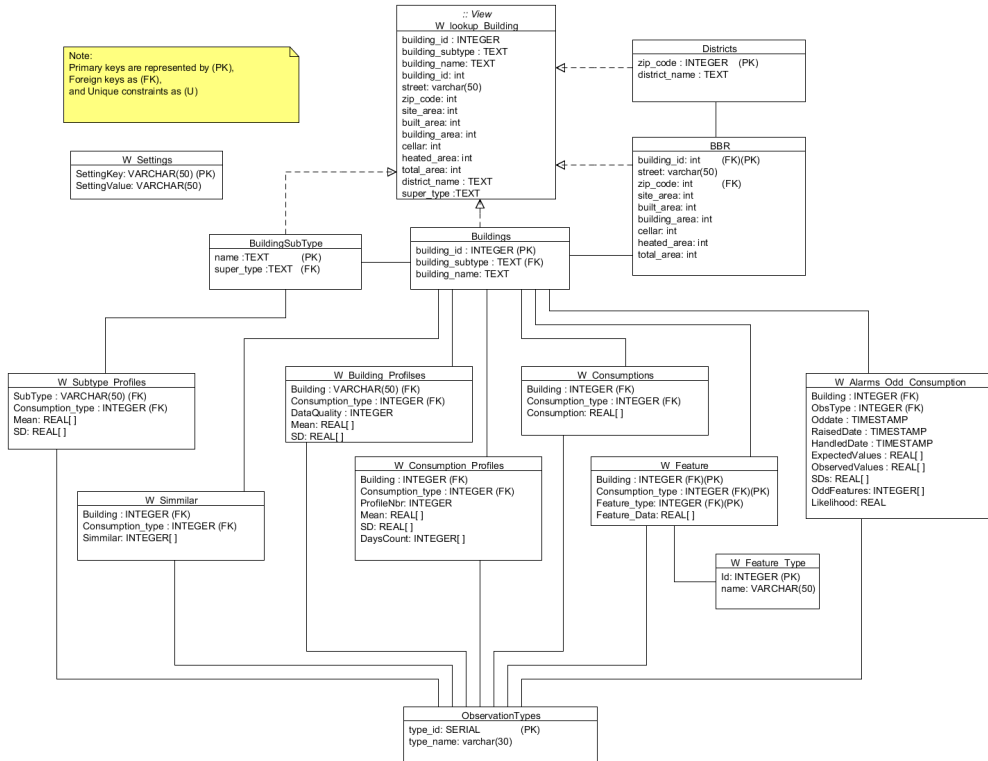


Figure 5.5: Database diagram, -Website

Processed data

Figure 5.5 is a diagram showing the part of the database which stores data that is the result of analyzing and processing the original raw data. The main goal of this design was to make the data as quick to access as possible in the way that it was going to be queried. As an example, all time series data in these tables are stored as arrays with no indication of time nor date, besides a single date in the settings table indicating the start time for all time series. This is because arrays are the most compact representation of time series data when transferred in the JSON format used by the website. To further optimize data transfer speeds, all numbers are rounded to a single decimal only. All tables and views whose names are denoted with *W_** are used by the website to retrieve information. The *W_Settings* table, holds configurable settings for both the website and other parts of the system.

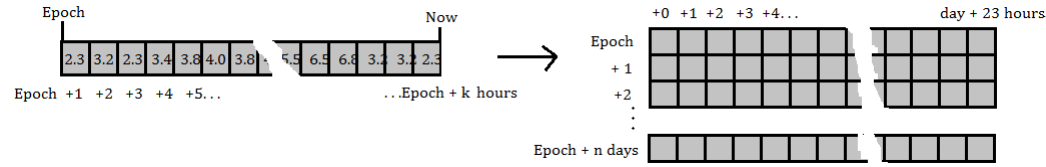


Figure 5.6: Illustration of feature extraction

5.5 Processing data

Processing and analyzing the raw meter data is done by a number of R scripts. The raw consumption data will be loaded from the database into an R session which will do various statistical calculations in order to get useful information from the raw meter data. The R session is launched once a day after new data has arrived, and runs until it has finished processing it. However, some operations are only performed on a weekly basis. The system is designed such that it remains usable while this batch job is running.

The Appendix 07 -Loading data describes how data is loaded into R and the initial data preprocessing procedures.

Extracting features

Once the raw consumption is loaded as vectors, the system will convert them to matrices with 24 columns, one for each hour of the day beginning with 00:00 at the first column and ending with 23:00 at the last as illustrated by figure 5.6.

Each row will thus correspond to a feature vector for a given day, where the 24 features are the hourly measured consumption during that day. Now for each day extra features are extracted and appended as extra columns to the original matrix. Table 5.7 shows a full list of the features that are extracted from the raw consumption figures of each day. Each feature is easily calculated in R by applying various arithmetic functions to each row of the matrix. Each feature is then appended as an extra column to that matrix, resulting in a final matrix with 39 columns. Columns 1 to 24 contain the raw consumption measures, while the extracted features occupy the other columns. Now when looking at an entire column alone, it can be considered a time serie itself with daily resolution. For instance column 25 will be a time serie of the average hourly consumption for each day since the *epoch* setting. Column 26 will be the peak consumption recorded for each day, and column 28 the average hourly consumption at night. Thus each column is considered a time serie in itself and is stored as an array in the *W_Feature* table along with ids for the building and consumption type (*water*, *electricity* or *heat*) that it describes. Thus the website can access this table to fetch time series showing the progression of various features for the buildings.

Feature	Description	Feature number
raw_consumption	The raw hourly consumption figures during the day	1-24
mean	Average hourly consumption across entire day	25
peak	Maximum consumption observed for one hour	26
min	Minimum consumption observed for one hour	27
night	Average hourly consumption between 00:00 and 03:00	28
morning	Average hourly consumption between 06:00 and 12:00	29
afternoon	Average hourly consumption between 12:00 and 18:00	30
evening	Average hourly consumption between 18:00 and 00:00	31
mean/peak	The ratio between features mean and peak	32
min/mean	The ratio between features min and mean	33
night/day	The ratio between features night and day	34
time_of_peak	At which hour does the daily peak occur	35
hours_above_mean	For how many hours is the consumption above the daily mean	36
no_of_peaks	How many peaks occur during the day	37
day_of_week	Categorical features indicating day of week	38
date	The date for when these measurements were observed	39

Figure 5.7: Table of different features index

5.6 Creating consumption profiles

To extract further information from the data a number of consumption profiles are calculated. The consumption profiles can be viewed mainly as two parallel vectors where one is the mean values of certain features, while the other contains the standard deviations for these means. Consumption profiles are generated on 3 different levels, and the procedures to generate them are only run once a week rather than daily.

Building profiles

A building profile is generated for each building and consumption type if there is enough data to do so. These profiles will contain the means and standard deviations of features 25 to 37. Thus a building profile summarizes data about a building's consumption patterns by calculating means and standard deviations for the various feature time series described in previous section. A building profile would for instance contain data about the overall averages and standard deviations for a buildings hourly consumption, night consumption, day vs. night consumption ratio etc. The building profiles are then stored in the *W_Building_Profiles* table, ready for the website to access.

$$\mu = \frac{\mu_1 \cdot n_1 + \mu_2 \cdot n_2 + \dots + \mu_k \cdot n_k}{n_1 + n_2 + \dots + n_k}$$

$$S^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \dots + (n_k-1)s_k^2}{n_1 + n_2 + \dots + n_k - k}$$

Figure 5.8: Subtype Profiles equation

Subtype profiles

The subtype profiles are created by grouping building profiles which belong to buildings of the same subtype (eg. schools, offices, nursing homes etc). Combining the means μ and standard deviations s^2 would normally be done with the formulas in figure 5.8, where n is the number of observations in each sample, which would be the number of days.

However, this approach is not used since we are not interested in the combined mean and standard deviation for each day across different buildings, but rather we will calculate new means and standard deviations using the means found in the building profiles as input. The subtype profiles are then saved in table *W_Subtype_Profiles*.

Daytype profiles

When it comes to energy and water consumption most buildings will have at least two different day types; workdays and weekends. On each type of day the consumption will follow a distinct pattern. For each building the distinct day types are identified, and a daytype profile modelling the consumption pattern is created. The day type profiles are created in four steps which include preprocessing, outlier detection, LDA and clustering. The matrix containing features for each day since epoch as explained in section *Extracting features*, is the input for these processing steps.

Preprocessing: Two adjustments are done in order to make similar days easier to detect. First the seasonality is removed by subtracting the minimum consumption for each day. Second, because all observation times are measured in GMT time, observation values measured during winter time are shifted one hour later. Otherwise the feature vectors of winter time days and summer time days would become out of sync, as people's routines would be one hour earlier during winter time.

Outlier detection: Using the PCount multivariate outlier detection algorithm, extreme outliers are removed from the dataset. The algorithm is applied to features 25, 26 and 27 (mean, peak and min), as at least one of them will be affected directly by an extremely high or low observed value at any given time during the day. It is important to remove outliers before performing LDA in the next step, because it is not a robust statistical method, and can potentially be highly affected by outliers.

LDA: Based on the knowledge that similar consumption patterns mostly occur on the same day of the week, a linear discriminant analysis (LDA) was performed to find the linear combinations of features 25 to 37 that give the best separation between the week days. The maximum number of useful discriminant functions is the minimum of $G-1$ and p where G is the number of groups and p the number of features. Thus the LDA resulted in 6 different linear combinations of the features that could best separate the days of the week, and by that also different types of days as those will most often be on the same day of the week. Using the linear discriminant functions from the LDA, each day's features 25 to 37 were reduced into 6 features called canonical variates (CVs).

Clustering Based on the CVs of each day a k-medoids clustering was performed. K-medoids was chosen over k-means because it performed well, and the implementation pamk in the R package fpc automatically detects the optimal number of clusters using the optimum average silhouette width which is a measure of how well the data is clustered [8].

Once similar days have been clustered, a day type profile is created for each cluster. This is done by finding means and standard deviations for features 1 to 37 as they looked like before any processing. The means and standard deviations are put in two parallel arrays and, along with the building id and consumption type, stored in the W_Consumption_Profiles table. Additionally an array of length 7 is stored, which indicates the number of times each weekday occur in this profile. Additionally a weight for each profile is computed as $(\frac{total_days}{days_in_cluster_n})$. For instance if 200 days were initially used for the clustering and clusters A and B were created with 50 and 150 days respectively, then two day type profiles would be created with weights 0.25 and 0.75.

5.7 Comparing buildings

To identify buildings with similar consumption patterns the daytype profiles are used rather than the building profiles. The system uses an approach similar to the K Nearest Neighbor method to find similar daytype profiles. Figure 5.9 illustrates with a fictitious example how using the daytype profiles is different from using the building profiles to compare buildings. In figure 5.9 a) 5 daytype profiles belonging to 3 different buildings are shown in a two dimensional space (each profile is assumed to be of equal weight here). Building 1 and 3 have two day type profiles each which are located pairwise near each other, while Building 2 only has 1 day type profile. Now using proximity as a measure of similarity we can see in b) that if we compare based on the daytype profiles Building 1 and 3 would be similar to each other. If we average the day type profiles belonging to the same building we would have essentially the building profiles of each building. Figure 5.9 c) shows that if those building profiles, rather than the daytype profiles, are used for comparing that now Building 1 and 2 would appear similar, rather than 1 and 3. This demonstrates that the two approaches render different results, and that by using the building profiles two buildings with very different consumption patterns can appear to be similar because

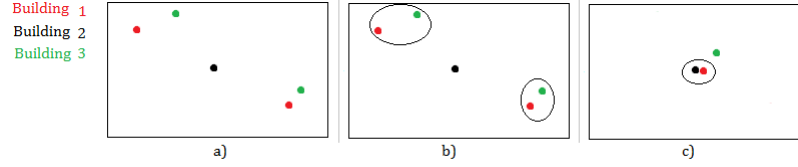


Figure 5.9: Building similarity

of the averaging of features. Therefore the comparison of buildings has been developed to be based on all the daytype profiles for each building and not the building profiles.

First a distance matrix between all day type profiles is computed using euclidean distances. The same canonical variates used for clustering are used to compute the distances. The smaller the distance the more similar the profiles will be. However as daytype profiles only partly represent the full consumption pattern of their buildings, both each row and each column of the distance matrix is divided by the weight of the profile at that row or column. This means that each measure of distance in the matrix is adjusted by the weights of both the two profiles being compared. For each building the distances across all it's profiles are summed. Then picking the profiles with lowest distance measure will be for the buildings that are most similar. The 30 profiles with lowest distance are picked out and the ids of the buildings behind those profiles are stored in an array which is saved to the *W_Similar* table from where the website can access the results.

5.8 Detecting faults

The detection of faults is based on the assumption that days of similar types have variables that are each approximately normally distributed. This assumption has been verified using a QQ-plot. The daytype profiles contain means and standard deviations of all variables for each type of day. With this information the following hypothesis test can be constructed for any daytype profile dt and a given day d :

- $H_0: d$ is of type dt
- $H_A: d$ is not of type dt

A two tailed hypothesis test is then performed using Z-scores to get the p value of this hypothesis test. First the Z score of each variable i is computed with:

$$Z_i = \frac{x_i - \mu_i}{\sigma_i}$$

where x_i is the value of variable i for day d , μ_i and σ_i are the mean and standard deviation of variable i found in the daytype profile. With the Z score the p value p_i for variable i can be calculated for the two tails as illustrated in figure 5.10: To get the total p value p_{total} across all

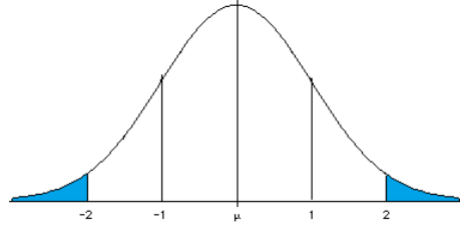


Figure 5.10: Two tailed cutoff on normal distribution

variables the following equation is used:

$$p_{total} = \prod_{i=1}^N p_i$$

where N is the total number of variables.

The significance level α indicates when to reject H_0 and say that a given day does not conform to a certain day type profile. A low α would decrease the chance of false alarms but decrease the chance that true anomalies are detected (ie. the chance of making type 2 errors increase). On the other hand a high α value increases the chance of detecting anomalies, but potentially increases the number of false alarms (ie. type 1 errors). H_0 is rejected when following equation evaluates to true:

$$p_{total} < \alpha^N$$

where N is the total number of features.

If all H_0 is rejected for all day type profiles the tested day is considered an anomaly, and the variable with lowest p value is flagged as a probable cause. The reason why a day is checked against every daytype profile individually rather than against a single building profile, is because the hypothesis test assumes that the means and standard deviations of variables come from a normal distribution. This would not be the case if a building's consumption comes from two or more distinct day types that each have normally distributed variables with different means and standard deviations. This is illustrated in figure 5.11, where two different normal distributions are plotted together, showing that the result is not normally distributed.

The features to use to detect anomalies and the α value can easily be configured by users through the website as the settings are stored in the *W_Settings* table. This will allow users to tweak the power of the test. The standard configuration is to use α 0.5 and two features, the average consumption across the entire day and the consumption at night. The first feature will detect an anomaly if the total consumption goes up or down for any reason, while the night consumption feature would be particularly sensitive if standby consumption of a building changes. When an anomaly is detected an alarm is created and stored in the *W_Alarm_Odd_Consumption*

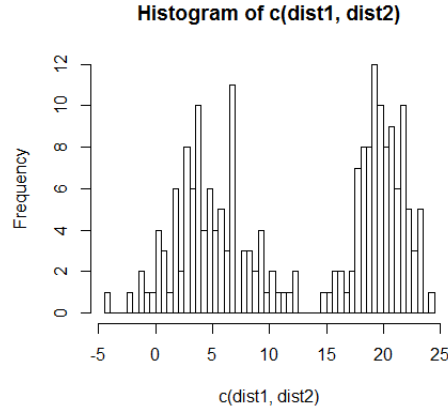


Figure 5.11: Two different normal distributions

table, where it can be accessed by the website.

5.9 Design of the website

This section will describe the design of the website. It will clarify which perspectives have been in focus, what technologies have been used. Lastly will it describe the different elements of the website.

The use case diagram from the Requirements section gives an overview of the requirements. Further does it show a grouping of the requirement into two main concepts: *Viewing a building* and *Comparing*. In relation to user interface design Steve Krug, argues that clearly defined areas and hierarchies aids the users learnability of a system and supports from getting lost [9]. Based on this, these identified concepts become important for the organization user interface, as these concepts can serve as groupings. Since the website should provide the user with a considerable amount of information, ordering and displaying it in a simple way becomes important.

In addition, each of the use cases relates to different purposes when using the system. In order to keep it simple and support the user in quickly gaining access to the information needed to perform a certain task, keeping each sub use case visually apart from each other may ease the navigation. Furthermore, by keeping them apart is it possible differ their appearance and thereby making them more memorable for the user and minimizing the risk of a user confusing/mixing up information presented.

Viewing a building

The one thing all use cases of viewing a building have in common is that they, to a certain extent only display information, and require little or no interaction from the user. Based on the findings from the previous section and with this in mind, it has been decided create and display each use case in separated bordered panels that clearly encapsulate and separates them.

Comparing buildings

Comparing buildings, in contrast to viewing a single building, requires a higher level of interaction from the user. Further, the first four use cases can be generalized into a concept of *what to compare*, and the last can be viewed as *how to compare*. This generalization, enhance the flexibility of the system, since it allows buildings, types and categories to be treated as one. Doing this, will allow comparison between buildings and types, categories and types or categories and buildings with a uniform GUI. With a uniformed GUI for the four types of content to compare, the number of steps the user have to learn is minimized, making it easier to learn, but also easier to remember. Based on this it is now possible to define three areas of the page:

- The content to compare.
- The features on which the user compare content.
- The visualization of the comparison.

In order to clarify the placement of these areas Jeffrey Venn extension to Keith Instone theory of the fundamental visibility of a system have been used [9]. Jeffrey Venn extension associates Keith Instones three question (*Where am i?*, *What is here?* and *Where can i go?*) with three areas on a page. In relation to this the content to compare have been associated with *Where can i go?*, to encourage the user to explore the data. Since the *Where am i?*, continuously throughout the website is have been allocated for navigation and information about where on the website the user is, the features to compare have been chosen to be put into main content area along with the visualization of the comparison. These two sub areas of the main content area, have been clearly divided, but still placed close together to insure the user catches the relation between them.

Flow of website

Based on the concepts and their hierarchy, discovered in previous section, the flow of the website was developed see figure 5.12. The page Building and Compare are directly causes of the groupings into hierarchies discovered, whereas the alarm pages originates from requirement *F3-02*⁹. The homepage have been created as an control and navigation page. In addition, the requirement

⁹See requirement specification.

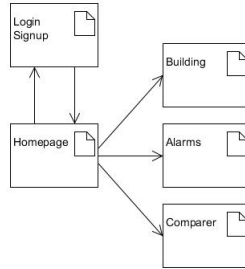


Figure 5.12: Flowchart of the website

specification expresses a need for user policies and restricted access to some features, like validation of data and handling of alarms. In order to accommodate these requirements the login and signup page have also been added.

5.10 Implementation of website

Template panels and visual components

As mentioned previously, as a design choice, it have been chosen to visibly keep different functionality clearly apart. From a system design perspective, standardizing these functionalities into components gives the software system have many advantages. Firstly, the system becomes more flexible and enhance the possibilities for code reuse. Further, by standardizing how components are created lowers complexity and easen development at a later state in the software life cycle. In addition, by standardizing reusable components the system becomes more easily changeable. Lastly by creating a “toolbox” with standard components, the view and display of information is more easy to customize to the user, if for example different users should have different views of the information. Further does it allow for generalizing functionality for collapsing, closing and reorganizing different elements on a given page.

For these reasons it have been chosen to design and implement a simple template framework (working as a factory pattern), where a controlling module contains a list of templates and how to instantiate them. Furthermore does the template framework contain a class which all templates extends. As seen in figure 5.13, an instance of the Template object contains a link or references to a HTML part. When an instance is created this HTML part is firstly cloned, and then filled out by the instance object with the provided data. The instance object, besides how to fill out the HTML, also contains all functionality regarding behavior and events. Lastly, to reduce coupling and make each component independent of other components all intercommunication between components is intended to work through events.

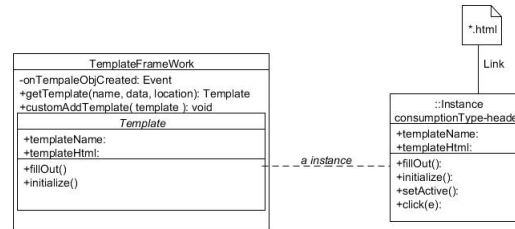


Figure 5.13: Template framework

Besides this template framework, to ensure a rapid development of the prototype a number of libraries and technologies have been used.

Technologies

Firstly, as some of the most commonly used, JQuery¹⁰ and Bootstrap¹¹ have been integrated. JQuery have been used to lessen the code needed to be written and to more easily handle events and animations. Further the Bootstrap framework for styling and in order to make the application more responsive and lean towards a mobile first development to ensure portability. In order to create visualizations of the data the D3¹² (Data Driven Documents) library have been used. Lastly for backend and client communication Dream Factory¹³ have been used. This open source platform, does with little configuration, create a RESTful service and API to database. Thereby enabling the front end application to get and set data from the database through the API. After a session have been established, through a login (either by an anonymous or by a registered user profile) data can be retrieved the client SDK.

The code snippet in figure 5.14 shows how the Comparing buildings page gets data for the charts. The *getRecords* method takes two parameters: an object containing information of what to retrieve and where to retrieve it, and a callback-function. The object-parameter has numerous configuration parameters. Mainly, in the implementation of the web-site, only three is used (*table_name*, *fields* and *filter*), which specifies the content of a basic SQL select-statement.

For further description of the implementation of the website see Appendix 06 -Implementation of the Website. This appendix discuss the implementation of the requirements on the website, the structure of the HTML and the relation between the different HTML files and JavaScripts files.

¹⁰<https://jquery.com>

¹¹<http://getbootstrap.com>

¹²<http://d3js.org/>

¹³<https://www.dreamfactory.com/>


```
4 //load feature for building
5 window.df.apis.postgres.getRecords(
6     {'table_name': 'w_feature',
7      'filter': 'building = ' + buildingId + ' AND ' +
8              'consumption_type = ' + window.activeConsumption + ' AND ' +
9              'feature_type IN(' + String(activeFeatures) + ')',
10     'fields': 'feature_type, feature_data, consumption_type'
11 }, initializeGraphOnLoad);
12
13
```

Figure 5.14: Code snippet of how to retrieve data

Chapter 6

Test and evaluation

The test and evaluation of the system have been divided into two sections. The first section will evaluate the performance of the system and its functionalities from a technical point of view, and additionally elaborate on how it can be improved. The second section, will evaluate the design of the overall system from a user's perspective. This section will focus on the initial problems as well as on the usability of the implemented requirements and use cases. In order to test and verify these, a representative from OM's Energy and Maintenance department has tested the system and given his feedback of the system in an interview.

6.1 Technical evaluation

To evaluate the various aspects of the systems performance data from 431 buildings was used. A year worth of consumption from 01. March 2014 until 01. March 2015 was loaded as trainings data, and a month of consumption from March 2015 was used as test data.

Daytype clustering

To evaluate how well the system finds and clusters similar day types, the process was run on all buildings in the year worth of training data. The k-medoids algorithm is used for clustering, and as a measure of how well the data is clustered the means and standard deviations of the average distances to a mediod will be used. For a cluster C this can be calculated as:

$$avg_dist = \frac{\sum_{i \in C} d_{ij}}{N_j}$$

where j is the medoid of Cluster C , d_{ij} is the distance from datapoint i to j , and N_j is the total number of objects in C excluding the mediod j . The clustering was tried with 5 different settings and the distances within each cluster was recorded. Figure 6.1 summarizes the results:

Setting Number	Remove seasonality	Adjust for winter	Remove outliers	Avg distance measured	Sd of average distances
1	FALSE	FALSE	FALSE	1.276	0.487
2	FALSE	FALSE	TRUE	1.219	0.484
3	FALSE	TRUE	FALSE	1.224	0.491
4	TRUE	FALSE	FALSE	1.189	0.500
5	TRUE	TRUE	TRUE	1.181	0.485

Figure 6.1: Summarized results for distances within each cluster

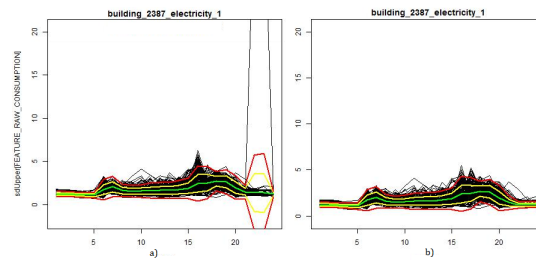


Figure 6.2: Daytype profiles with and without outlier detection

This table seems to indicate that the various preprocessing methods do lower the average distance, but the results were too subtle to show in a box-plot. This is probably due to the CVA analysis performed after the above preprocessing but before the clustering, and thus it masks the actual effects of the preprocessing. However, using raw consumption numbers rather than canonical variates for clustering, it was found that adjusting for winter time lowered the total within-cluster sum of squares almost 30% from 14654 to 10541. These results were achieved at an earlier stage in the project using k-means for clustering and no CVA.

Figure 6.2 illustrates the importance of removing outliers when creating the daytype profiles. a) shows building 2387's electricity consumption for a cluster of similar days (black lines) with an outlier present, while in b) the outlier was removed. The mean consumption is shown in green, while the yellow and red lines are 1 and 2 standard deviations away from the mean. We see in a) how the standard deviation is greatly impacted by the outlier, because the procedure was run with outlier removal disabled.

The average number day types pr. building is 2.74 with a standard deviation of 0.95. This indicates that it is almost always possible to identify at least two distinct day types, namely working days and non working days.

Often the best way to determine whatever a clustering is successful is by visually inspecting the results. Figure 6.3 shows two distinct day types and how they can be almost entirely separated using only two canonical variates, which explain 66.67% of the point variability.

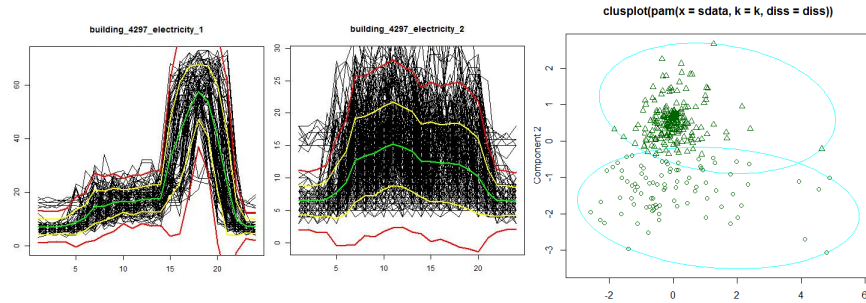


Figure 6.3: Two distinct day types and a cluster plot of the two day types

Appendix 11 -Daytype test setting contains 925 plots showing each of the day type profiles generated while running the test with setting number 5.

The clustering was tried with the DBSCAN algorithm¹. However, this did not perform very well as it would often just create one single cluster. Figure 6.3 right, gives a hint as to why this is the case. Even though the two clusters shown to the left seem clearly distinct, the cluster plot to the right shows that they have overlap. If this overlap is too dense, then due to the way DBSCAN works they will come out as one single cluster.

Detection of faults

Based on the day type profiles generated on the training data, the test data was checked for anomalies. The fault detection settings used was an alpha value of 0.05 and the features used for the fault detection were mean hourly consumption and nightly consumption. During the process buildings that did not have any consumption profiles, (due to missing) or days with missing data were discarded. The result was that 391 buildings were processed and in total 2435 days were checked for anomalies. In total 76 anomalies were detected of which 36 had to with a higher than usual consumption. Figure 6.4 shows the detected anomalies across buildings.

The 78 anomalies are spread across 20 buildings. Building 6028 has the largest number of alarms namely 12 from which 11 are due to higher consumption. Investigating this issue further shows that the reason is a higher nightly consumption which is consistently above the means of its two daytype profiles. Figure 6.5 shows the two nightly consumption means from the day types (red lines). The blue lines are the maximum one can move up or down while still being within 2 standard deviations from any of the day type profiles. The black line shows how building 6028 has had a consistently higher than usual consumption at night.

Another way that the fault detection system was tested was to simulate a higher consumption and then see how many days would be detected as having this higher consumption. This test was

¹See Related work and Methods

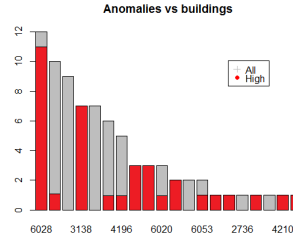


Figure 6.4: Anomalies vs. Buildings

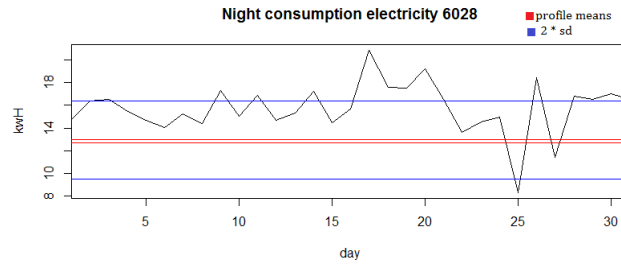


Figure 6.5: The nightly electricity consumption of building 6028

run with a simulated increase in consumption from 0 - 100%, and with α values 0.05 and 0.025. The results can be seen in figure 6.6. The figure shows that for instance increasing the consumption 20% would be detected for roughly 20% of all days. However the system will probably perform better on real data, because the increase would not be spread evenly across the entire day like the simulated increase.

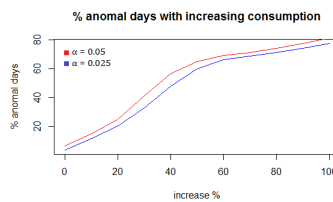


Figure 6.6: Simulated test of the fault detection

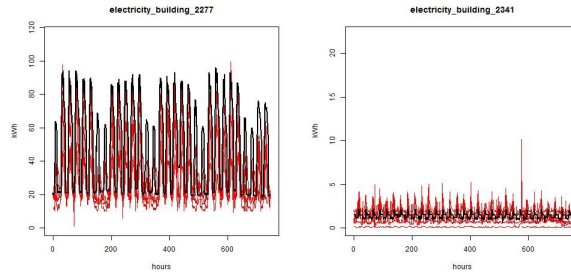


Figure 6.7: Similarity of buildings consumption

Finding similar buildings

This feature is somewhat difficult to evaluate since there is no clear metric for how similar two buildings in fact are. The approach will be to compare all buildings using the day type profiles generated in section daytype clustering, on page 35. These profiles have been generated on basis on the training data, and thus the system will find similar buildings on basis of the training data. How well the system performs in finding similar buildings will be evaluated using the test data.

Using the training data the system compared all buildings and for each building the 30 most similar were noted. Figure 6.7 shows the raw electricity consumption of two different buildings during March 2015 which was not part the training data. The black lines show the consumption of each building, and the red lines show the consumption of the 5 buildings that were rated as most similar. We see from the plots that the system did well finding similar buildings. Appendix 12 -Similar Buildings contains a plot for each building that was used.

Another way to test how well the system identifies similar buildings is to look at the properties of the similar buildings rather than just the consumption figures. Unfortunately building properties like size, number of people, age of building etc. were too inconsistent in the dataset used. However the type of building (school, office, nursing home, etc.) was available for all buildings. Figure 6.8 shows for a selected number of building types the average number of times that a building of this type appears in the list of similar buildings for a building. The red bars show this average for buildings of the same type, while the blue bars show the average for buildings of different types. The figure shows that buildings of a given type, are much more likely to have buildings of the same type appearing in their list of similar buildings than buildings of other types.

Integration of the system

BEOVulf is fully automated in importing data from OM's Energy Key system. Therefore it is possible from a data perspective to adopt the new system without any cost in effort for maintain

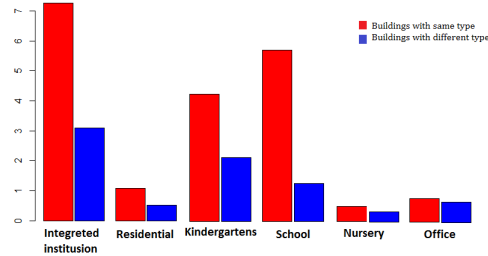


Figure 6.8: Similarity of buildings consumption

multiple systems. However updates to informations in BEOVulf do not propagate to Energy Key. The two way integration on has not been made possible from the municipality, nor has it been a goal for this project. The system is currently only operating on 505 buildings, and have yet to import the buildings from the old SonWin² system. However, this tasks should be quite easy, since the developed system now have a implemented structure in which these data could be fitted into.

Concerning the data quality in the developed system, some effective procedures have been developed which ensure that duplicated data entries can not occur. The database design along with the procedures ensure minimal usage of space and that data is consistent. All this ensures that the data quality is still the same when it is stored in the database, which again means that in order to improve data quality, the source needs to be improved, which has been out of scope for this project.

6.2 Usability evaluation

In order to test and verify the usability of the system, a representative from OM's Energy and Maintenance department was interviewed. He was a test pilot for the system, in a semi-supervised environment, where he was both using the system on his own, and also while being guided through some parts to ensure that he did not get stuck in the exploration of the system. Additionally, the guidance ensures that all features of the system were evaluated.

Learnability of system

This section will evaluate how easy the system is to learn. This is of importance, because it addresses the effort needed from the user to become competent at carrying out tasks. If the system is easy to learn the adoption will be more smooth, the user, more efficient and it will lessens the change of the user making errors.

²<http://www.sonlinc.dk/sonwin.aspx>

The test of the user interface showed that the visual design and concepts of the system were understood. The user was able to navigate around the page without help, showing that the current placement of the content is consistent with where the user expects it to be. This lessens the chance of confusion and undesirable experiences, like the frustration of getting lost or stuck.

The test showed that the content and information presented when viewing a single building is intuitive and recognizable to the user. It took the user a few seconds to look through the BBR informations and get an overview of the building information. The example shows that the quantity of information displayed is comprehensible. However, the test user did suggest that various panels could be collapsible, and that users with specific tasks should be able to have panels with irrelevant information to those tasks collapsed by default. Collapsible panels, have been a part of the original requirements and design plan, as mentioned in section Implementation of website, on page 37. Therefore the system design can accommodate this feature with little change of the existing system. The ability for individual users to customize the interface was originally specified in requirement F15-F17, but was left out of scope for this project. However, the reaction expressed by the user can also be interpreted as a slightly confusion. This could indicate that the information presented is not intuitively enough to be interpreted on its own. This confusion might not be a problem for the power-user as it will expectedly decrease as the user learned or got used to the system, but if the system is to accommodate the general public, there will be a risk of the system being too advanced. In relation to this the following countermeasures might reduce the risk and increase the acceptance rate from the general public:

- Panels could be made visually more different from each other. This will make them more memorable and reduce the risk of users mixing up graphs
- Panels could be made more self describing by adding guiding text on mouseover to different elements on the graph.
- A discrete help/information button could be added in the panel header, or alternatively as an expandable panel footer, with guiding/explaining text.
- Lastly, an alternative could be to introduce a tutorial for new users, where different elements could be explained.

Comparing buildings: On the building comparison page, the test user needed a little guidance at the beginning to understand what the right hand side bar was for. However the information needed to get him started was minimal, indicating that no additional helping functionality is needed and that the design works as intended. However, the test, showed that it is expected that clicking legend items will remove the item from the graph. This identified feature should be implemented in a further development, but is not found critical for the proof of concept prototype. The implementation of this feature in a further development will not affect the rest of the system, and will require little effort to implement.

Maintainability of the system and data

This section will evaluate the maintainability of the system. It will elaborate on what effort is needed to maintain and sustain the system and the data quality wanted by the Energy and Maintenance department.

Features for validating and changing existing data, have been included in the developed system, which means that the employees from the Energy and Maintenance department can verify and correct their data. These functionalities have also been tested both by the developers and the head of the department of Energy and Maintenance, and proven to be both easy and fast to use. As an example the head Energy and Maintenance answered the following when asked what was most important and if the system lacked any of this:

*"I'm just totally hooked on the feature with whether or not a building is validated"*³

This shows that is not always the big and advanced feature which creates the value for the user, but sometimes the features which are simple, small and easy to use. From this acceptance of the feature it have been concluded that the maintenance of the system data, will not burden the municipality with a lot of additional work. The head of the department of Energy and Maintenance stated that quite the contrary would be the consequence, because the timestamp on the panel would ensure that the personal did not do the same work twice as he expected they did quite often at the moment.

6.3 Utility and efficiency of the product

This section will evaluate the way the system supports the primary users in doing their tasks, and whether the system provides the right kind of functionality so that users can achieve their goals and sustain a high level of productivity. This section will start by looking at the page where the user can look at a specific building. From there it will move onto the page visualizing all alarms in the system, and lastly will it describe the comparison page. Lastly, this evaluation is based on the final test and interview.

Viewing a specific building After the representative from OM's Energy and Maintenance department had tested the system, an interview was conducted which focussed on the department's work and the possible applications of the system. In this discussion, it was found that the employees are mostly interested in whether or not the building is validated, rather than the informations themselves. Due to this, it was concluded that the panel took up too much space, and should be minimized when viewing a building after a validation. If this functionality is not implemented, in consequence this could lead to negative user experience, because employees will spend a great deal of time on scrolling past the panel.

³AppendixAppendix 03 -Meeting with Kim Allan 26-5-15: 7 minutes and 3 seconds in.

Requirement *F2*, and its sub requirements, concretizes the visualization of a building. The feedback on these have showed it somewhat easier to see the behavioral patterns of a building in the developed system. Furthermore, when showing the test user a building, which the group members had discovered and found odd, it took the user less than a minute to see that something was in fact wrong. In addition it took him only few minutes before he had a list of probable causes. This shows the product does provide the tools necessary to identify problems and estimate their cause. Further, that the tools are more effective than those currently used. Unfortunately, there have been no time to test the software on a large scale, and therefore it can not be confirmed as a unique case or not.

The test have found that the energy profiles is a powerful tool to obtain information of the behavioral patterns of a building. With this tool it is possible to learn a building's consumptions behavior in a matter of minutes instead of days/weeks. However, it seems that the presentation of these information can be improved, since the sliding and changing diagram can become annoying when a user is trying to read a chart/profile.

The test further showed that the seasonality functionality was an effective tool. The department of Energy and Maintenance makes these graphs, as a supporting tool for estimating the budget of a building. However, the software the department is currently using does not support them in generating these and therefore they currently have to create these graphs manually. When the Energy and Maintenance department was making the budgets for all 550 buildings, they tried creating these charts manually, but found it too time consuming and had to stop and find another procedure. Currently the procedure is to estimate with a professional guess. This makes a lot of space for making an human error. By introducing the autogenerated seasonality graph, the department of Energy and Maintenance would have the tool to create the budget based on facts, without acquiring additional effort. This would improve their current way of estimating the budgets for the buildings.

Alarms The test of the alarm feature revealed that error-code and error-messages along with visualization of observed consumption and normal consumption would be a very effective tool for the Department of Energy to diagnose the cause and analyse what counter action would be needed. Further the test concluded the department would save of time if they were using the developed alarming system.

As for the page with all the alarms the test showed that it would support their workflow and support them in doing some of the more static tasks without too many mouse clicks. Based on this it have been concluded that the feature is accepted and that it will create value for the user.

Comparing buildings Regarding the page for comparing multiple consumption the test showed that the features are accepted by the user. Further it have been proven to be easier, faster and more flexible than the closest tool the department currently have at their disposal. The tool currently used is annoyingly slow, frustrating and do not provide the information in a very understandable way.

The feature providing the possibility of plotting entire building types and compare these with a single building gives the department a new way of comparing and new understanding of the building. This feature was noted by the user, as a good way of giving a consumption context and perspective.

As for the comparison features listed on the right side of the page, a change request was made. The system should support signed in user with functionality making it possible to change/define the feature. For example should it be possible to define which interval of the day should be used as the nightly consumption. This might be needed, because if the department of Energy are to trust what they see in the diagrams, then have to know what they are looking at. In addition, by creating a system where the user can change and interact more with the graph, the system will be more flexible and thereby accommodate more different tasks. Lastly, by creating a system where the user can see and change how a number is derived, it is easier for the user to understand and customize it to the user's specific needs.

Lastly, but unexpected the feedback from the test showed that that the page/tool would help the department in localizing possible buildings for renovation. Further that the tool could provide a platform from which they could inspect/patrol many buildings faster than in their current system.

Chapter 7

Conclusion

To conclude whatever the project was successful, the developed systems capabilities will be reviewed against the project initial goals. In the introduction they are stated as being, to create a functioning prototype website capable of highlighting various trends, correlations and odd events in respect to energy consumption in buildings. Furthermore it was a goal to create a system that was tailored to the needs of the Municipality of Odense. This made usability and practical application of the system a requirement in order to make it benefit people in the real world. In relation to this, the developed systems ability to solve the municipality's five issues which were identified and described in section *Background*, will be used to conclude whether this project has been successful.

7.1 System and data maintenance

With OM's current system they do not have the resources to focus their work on becoming more energy efficient, because system and data maintenance have become overwhelming tasks. The developed system counters this problem by being fully integrated and automated with the existing system. The system uses database structures which ensure that the data is always consistent, and it actively supports and encourages users to validate that old data is still valid. This functionality ensures that redundant work in validating data is avoided. With these features, BEOVulf comes a long way in order to lessen the tasks of maintaining system and data, thereby releasing resources which can be allocated for optimizing the energy efficiency.

7.2 Visualization of buildings and consumptions

As for the second issue, the municipality lacks the tools needed in order to understand consumptions and manage the buildings, the system provides the following tools:

- Visualization of raw consumption.
- Visualization of various energy consumption patterns.
- Visualization of key consumption features including seasonality, nightly consumption, peak loads and more.

Compared to the municipality's current system, these features considerably lessen the effort involved in getting useful information about a building's energy consumption. This will make OM's Energy and Maintenance department more effective and better enabled to make informed decisions. In this respect BEOVulf has met the goals for visualization of buildings and consumptions.

7.3 Fault detection, diagnosis and resolution

Due to limited resources OM's Energy and Maintenance department can not manually inspect and monitor all buildings, but rather need to focus their effort on the buildings that need attention. BEOVulf's fault detection system has proven that it can detect anomalies in consumption patterns, and thus it is able to notify energy managers when a building needs attention. Thus OM's Energy and Maintenance department can focus their work on correcting the faults instead of on searching for the needle in the haystack.

The automatic fault detection will enable OM to respond faster to leakages and thereby minimize energy and water waste. The system is able to suggest which aspects of the consumption that triggered the fault, and even though it is not fully implemented in the prototype, the fault detection system is also aimed to be a task management system. This enables OM to effectively assign and prioritize resources towards fault resolution.

7.4 Budgeting

In relation to the fourth issue of creating better and more accurate budgets, the visualization tools provides the means for estimating more precisely the consumption of a building. With these tools the department of Energy and Maintenance will no longer have to rely on a old unvalidated data or as it was the case sometimes, a professional guess. However, BEOVulf does not provide any tools to directly support budgeting tasks. This was part of the initial aim, but was within the limited time of the project prioritized lower in favor of other requirements.

7.5 Selecting buildings for renovation

The task of selecting buildings for renovation involves comparing and screening the consumption of many buildings. This is a difficult task to do with OM's current system. To make this task easy BEOVulf implements functionality enabling the user to quickly compare many different features

for different buildings, by having them plotted in the same graph. The system also proved that it performs well in automatically identifying similar buildings, allowing the user to browse through them and visualize their consumption. These features all help users generate an initial set of candidate buildings for renovation. However, the development of the system did not come as far as having a feature that automatically identifies these candidates, which was originally the goal.

7.6 Project overall

Many improvements are still to be made if BEOVulf is to become a complete system that fully fulfils the project's initial goals. However as a fully functional prototype it comes a long way. BEOVulf is able to visualize consumption, detect anomalies, find similar buildings and a user test showed that the system is practically usable. Thus the conclusion is that BEOVulf successfully applies data mining and statistical methods to recognize and visualize energy consumption patterns of buildings.

Bibliography

- [1] Christian Beckel, Leyna Sadamori and Silvia Santini, " *Towards Automatic Classification of Private Households -Using Electricity Consumption Data*", Buildsys'12, November 6, 2012, Toronto.
- [2] Joern Ploennigs, Bei Chen, Anika Schumann and Niall Brady, " *Exploiting Generalized Additive Models for Diagnosing Abnormal Energy Use in Buildings*", Buildsys'13, November 13-14-2013, Roma, Italy.
- [3] Xiaoli Li, Chris P. Bowers, and Thorsten Schnier, " *Classification of Energy Consumption in Buildings With Outlier Detection*", IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, Vol. 57, No. 11, November, 2010.
- [4] Imran Khana, Alfonso Capozzolia, Stefano Paolo Corgnatia and Tania Cerquitellib, " *Fault Detection Analysis of Building Energy Consumption -Using Data Mining Techniques*", Elsevier Ltd, 2013.
- [5] John E. Seem, " *Pattern recognition algorithm for determining days of the week with similar energy consumption profiles*", Elsevier B.V, 2004.
- [6] Peter Filzmosera Ricardo Maronnab, MarkWernerc, " *Outlier identification in high dimensions*", Elsevier B.V., 2007
- [7] Teemu Räsänen, Juhani Ruuskanen and Mikko Kolehmainen, " *Reducing energy consumption by using self-organizing maps to create more personalized electricity use information*", Elsevier Ltd., 2008
- [8] Peter J. Rousseeuw (1987). " *Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis*". Computational and Applied Mathematics 20: 53–65
- [9] Yvonne Rogers, Helen Sharp and Jennifer Preece, " *Interaction Design, -beyond human-computer interaction*", Wiley, 3rd Edition, 2012