Gezamenlijke opleiding Universiteit Hasselt en KU Leuven

2013•2014
FACULTEIT INDUSTRIËLE INGENIEURSWETENSCHAPPEN
master in de industriële wetenschappen: energie

Masterproef

Data mining for Energy Management according to ISO 50001

Promotor :
ing. Leo RUTTEN
ir. Geert VANDENSANDE

Promotor :
Dhr. ADRIAAN BREBELS

Jeroen Fosse
Proefschrift ingediend tot het behalen van de graad van master in de industriële wetenschappen: energie

2013•2014
Faculteit Industriële ingenieurswetenschappen
master in de industriële wetenschappen: energie

Masterproef
Data mining for Energy Management according to ISO 50001

Promotor :
ing. Leo RUTTEN
ir. Geert VANDENSANDE

Promotor :
Dhr. ADRIAAN BREBELS

Jeroen Fosse
Proefschrift ingediend tot het behalen van de graad van master in de industriële wetenschappen: energie

universiteit hasselt | KU LEUVEN

Acknowledgments

After obtaining my bachelor's degree in automation, I decided to start a master of Energy-Electrical Engineering. In order to learn more about courses that were less discussed in the professional bachelor. Based upon this philosophy, getting a broad knowledge of different techniques and technologies, I decided to search for a master's thesis which was related to new technologies. Within this search I discovered Porta Capena. For many years Porta Capena is working on today's hot topic, discovering knowledge from data. Therefore, I started my master's thesis at Porta Capena on the topic of 'data mining for energy management according to ISO 50001'. However, without the help of several persons I wouldn't have been able to fulfil this task. I have done the research to find the best data mining technique and I have defined all the algorithms and solutions, whereas the programming is done by engineers in Poland. Therefore, I would like to extend my appreciation for the positive spirit of cooperation from everyone who contributed to this thesis.

First of all, I would like to thank my company tutor Adriaan Brebels for his valuable advice and for the new insights in data mining and general management. Furthermore, I am highly grateful to my school supervisors Geert Vandensande and Leo Rutten for their helpful comments and their sharp analysis of each development. I would like to extend my gratitude to all the people of BEMT and Porta Capena. The help from BEMTs people was crucial in the development of a communication link between Porta Capena and the BMS of BEMT. Furthermore, the interaction with engineers from Porta Capena Polska, especially Marzena Magda, was crucial in the programming of the algorithms. Special thanks go to Jeroen Lievens for his help and his tips in writing the fundamental text parts.

Finally, I want to express my gratitude to my girlfriend for her mental support and her help with revising my text. I also would like to thank my parents and my girlfriend's parents for their understanding and help during writing this thesis.

Table of contents

Tables list

Table of figures

Nomenclature

BMS: Building Management System
CDD: Cooling Degree Days
CHP: Combined heat and power
COM: Component Object Model
DCOM: Distributed Component Object Model
DD: Degree Days
DI: Data Integrity
E-level: Energy level
EcoSCADA: Ecology/Economy Supervisory Control And Data Acquisition
EM: Energy Management
EnMS: Energy Management System
EnPI: Energy Performance Indicator
EPBD: Energy Performance of Building Directives
EPC: Energy Performance Coefficient/Contract
GDD: Growing Degree Days
HDD: Heating Degree Days
HVAC: Heating Ventilation Air-conditioning
I/O: Input/Output
IQR: Interquartile Range
KPI: Key Performance Indicator
LED: Light Emitting Diode
OLE: Object Linking and Embedding
OPC: OLE for process control
OSI: Open Systems Interconnection
PLC: Programmable Logic Controller
ppm: parts per million
Q1: Lower quartile
Q3: Upper quartile
QQ-plot: Quantile-Quantile plot
RFID: Radio Frequency Identification
RTU: Remote Terminal Unit
SCADA: Supervisory Control And Data Acquisition
TCP/IP: Transmission Control Protocol/Internet Protocol
UTC: Coordinated Universal Time
WMS: Warehouse Management System

Summary

Porta Capena is specialised in, inter alia, data mining for energy management. Data analysis of public buildings has shown that several of these buildings consume too much energy on holidays. In order to solve this, Porta Capena wants to detect the holiday anomalies automatically. Furthermore, the company wants to provide a solution for the end-user to tackle this issue. The reference frame of this thesis is that each development must be applicable within the standard ISO 50001.

The algorithm to automatically detect holiday anomalies is developed with the use of statistical programs. Microsoft Excel imports the data and ensures the pre-processing of the data. The statistical program R selects the most appropriate data mining technique. The technique to prevent the extra energy consumption is optimized by the use of a practical case, the building of BEMT in Geel Belgium. Therefore, the installed BMS and the hardware changes determine the most appropriate technique.

The algorithm predicts an opportunity of saving € 1400 annually, electricity and gas combined, for the building of BEMT. Furthermore, it takes less than 2 minutes to generate such a year analysis. Besides, an improved- and user friendly – user-interface combined with an energy dashboard is the most appropriate implementation technique. Direct hardware communication between the systems (BMS and Porta Capena) provides a robust and easy-to-use solution.

Abstract

Porta Capena n.v. is o.a. gespecialiseerd in datamining voor energiemanagement. Uit data-analyse blijkt dat gebouwen teveel energie verbruiken op verlofdagen. Daarom wil Porta Capena de outliers op verlofdagen automatisch detecteren. Daarnaast wil de firma de eindklant middelen geven om dit meerverbruik te voorkomen. Als referentiekader moeten de ontwikkelingen voldoen aan de norm ISO 50001.

Het algoritme om automatisch verlofdagen als outliers te detecteren wordt ontwikkeld aan de hand van statistische programma's. Het rekenbladprogramma Microsoft Excel importeert de data en zorgt voor de preprocessing. Vervolgens zoekt het statistisch programma R de meest aangewezen datamining techniek. De techniek om het meerverbruik te voorkomen, wordt geoptimaliseerd aan de hand van een reële case op het gebouw van BEMT in Geel. Het daar geïnstalleerde Building Management System en aanpassingen van hardware apparaten bepalen welke techniek hierbij het meest aangewezen is.

Het algoritme voorspelt concreet een jaarlijkse besparing van €1400, elektriciteit en gas samen, voor het gebouw van BEMT. De generatie van een jaaranalyse duurt minder dan 2 minuten. Daarnaast blijkt een verbeterde – en eenvoudig bedienbare - user interface in combinatie met een energiedashboard de meest aangewezen implementatietechniek. Directe hardwarecommunicatie tussen de systemen zorgt voor een robuuste en eenvoudige oplossing.

1 Introduction

As a master student of the faculty of Engineering Technology within the course Master of Industrial Sciences Energy at the Hasselt University – KU Leuven University, I have had the opportunity to work together with Porta Capena NV in Geel. This company founded in 2001 as a group of professional programmers and engineers from Belgium and Poland. From the beginning, they have focused on general data mining with several specialisations. One of these specialisations is data mining for energy management and the development of EnMSs. The main target of this specialisation is to improve the energy efficiency [1] of public buildings. Nowadays more and more companies want to save energy. This is due to several reasons:

1. organisations want to highlight their environmental awareness;
2. strengthen their market position against competitors (costs);
3. European targets (2020) oblige countries to consume less energy.

Porta Capena uses the Internet as a common interface to tackle these issues. This results in flexible solutions which are freely accessible for customers.

1.1 Problem statement

Porta Capena starts with a building scan which analyses the historical data of the building. The scan also gathers other relevant building information, for instance opening hours and holidays. Such a building scan highlights the energy consumption problems. Based on the discovered problems, the most appropriate algorithms are selected and implemented in the EnMS. A common detected problem is that buildings consume too much energy (electricity and gas) on holidays.

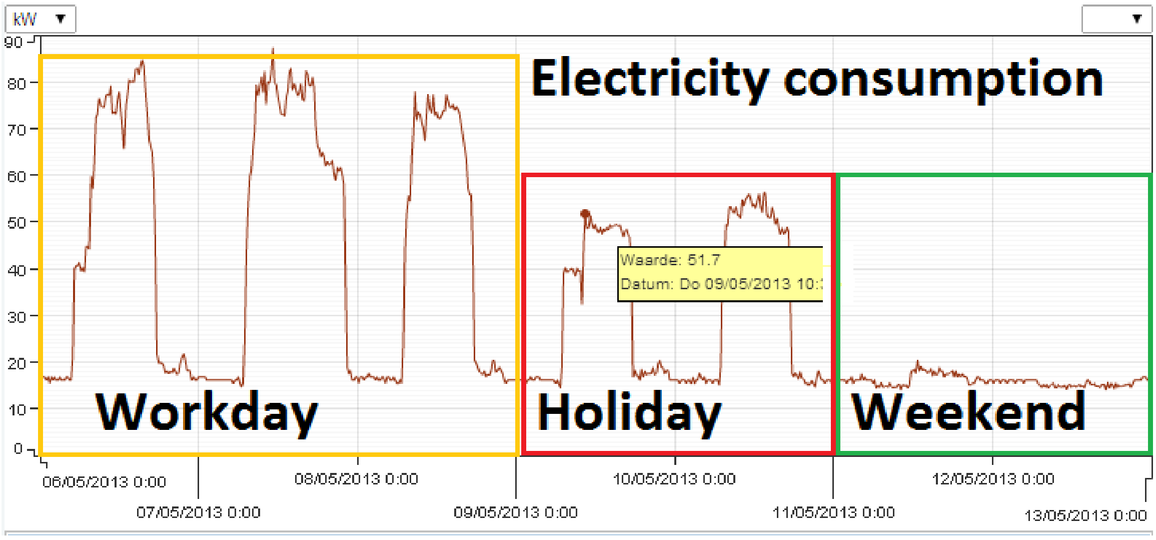


Fig. 1: holiday problem

Interviews with employees highlighted that an unfriendly user-interface is the problem of this extra consumption. This divides this problem statement into 2 parts

[1] Energy efficiency: consume less energy while the comfort either economical activity does not change.

1. Porta Capena detects this problem each time manually by using the historical data of the building. This results in a very slow data scan and an inconsistent output, since manual work is time consuming and depends on the person. This means that the holiday anomaly detection is not applicable for daily use and that the detection method is not standardised.
2. A good user interface must not only produce the right information at the right time for the user but it needs to interact with the local BMS system too. This connection is an important item from the customers perspective. Once the customer knows about the problem, he will want a solution.

1.2 Objectives

The objectives of this thesis are explained by the use of Fig. 2.

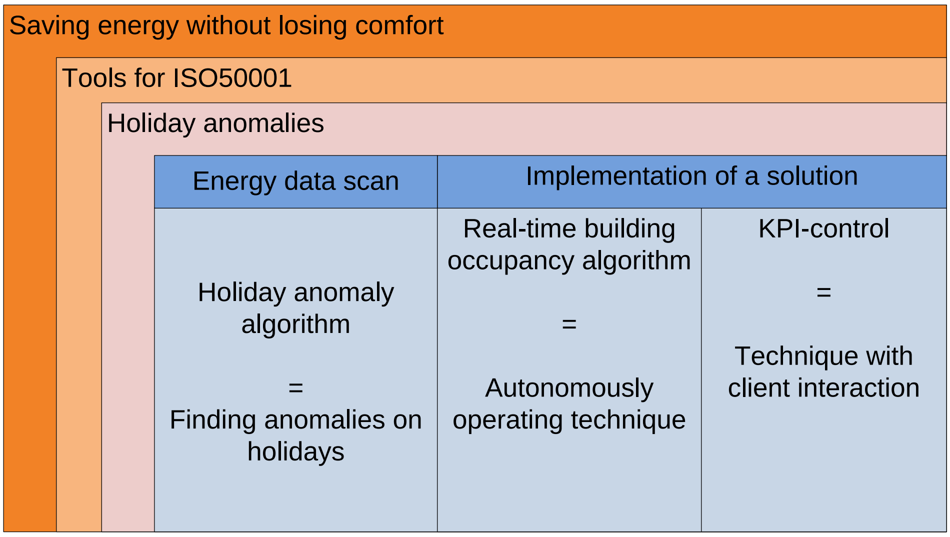


Fig. 2: thesis objectives

The reference of this thesis is visualised in the 2 largest frames. Firstly, the development of the solutions must fit within the philosophy of Porta Capena, energy efficiency. Secondly, each solution must be applicable within the standard ISO 50001. The real objective is divided into 2 parts and is situated within the holiday anomaly problem.

1. A holiday anomaly algorithm which automatically detects holiday outliers out of historical data must be developed. The output of the analysis must be visualised within an easy-to-use report.
2. The detected problem must be solved by the implementation of a solution. Therefore, 2 implementation techniques are investigated. First technique is working automatically and does not need any user input (real-time building occupancy algorithm). The second technique is based on the use of KPIs and input parameters of the user (KPI-control). The most appropriate technique

should be developed. Furthermore, a communication principle between the developed technique and the installed BMS should be established.

In addtion to these objectives, it is possible to define a list of demands. The demands for the holiday anomaly algorithm are:

- in accordance with ISO 50001,
- uses historical data,
- generate the number of outliers,
- calculates the amount of lost energy,
- estimates the potential savings,
- report can be opened directly in Microsoft Excel.

Furthermore, the demand for the selected implementation technique is:

- in accordance with ISO 50001.

Finally the demands for the interface to the BMS are:

- robust,
- failure proof,
- BMS independent,
- high quality/price ratio.

1.3 Methods and materials

This master's thesis uses software programs and hardware devices, if they are needed for tests or implementations.
The first stage will use only software programs. A data collector is used to extract data out of a database 'EcoSCADA' and to synchronise this data with data from other sources like weather stations. Since the data collector is a recent development, data should be verified with the original data from EcoSCADA. The research stage for the algorithm uses 2 software programs. Microsoft Excel is used to control and analyse the data, while the statistical program R is used to test several data mining techniques. Testing more complicated (decision tree, clustering) data mining techniques is not possible in Excel. In comparison to other statistical programs (SPSS and SAS), R is free to use and open source. Therefore it is used by many students and researchers. Depending on the selected data mining technique a relevant software package is selected, like R or Excel.

The research towards the most appropriate implementation technique and the development of the selected technique is based upon the use of software programs and features of hardware devices. A list of requirements and possible solutions will be generated. The most appropriate technique is selected based upon that list. Furthermore, the interface is exemplified by the BMS of the building of BEMT in Geel. The final selection presents a list of advantages and disadvantages and a list of prices.

1.4 The content

Chapter 2 & 3 introduce the basic concepts which are used in this master's thesis. Chapter 2 explains the state-of-the-art of this thesis, the reasons why energy efficiency and energy performance are hot topics nowadays, while chapter 3 investigates all the different aspects which are used in this thesis. The whole concept of energy management within the standard ISO 50001 is explained. Furthermore, the basic principles of data mining are described, since these principles are constantly used throughout this thesis. Moreover algorithms and KPIs are discussed in this chapter.

The first steps towards the holiday anomaly algorithm are taken within chapter 4 & 5. Chapter 4 explains which methods are used to extract data out of the EcoSCADA database and what pre-processing is about. It also analyses which of the introduced extracting methods is the most appropriate. Furthermore, data pre-processing is explained with the use of several examples. The examples are useful to learn and to understand the different principles of the pre-processing steps. Chapter 5 investigates which kind of data mining techniques might be used to develop the holiday anomaly algorithm. Therefore, the statistical program R is used to show the effects of each method on the data. The conclusion of this chapter summarizes the advantages and disadvantages of each method.

Chapter 6 discusses the development of the holiday anomaly algorithm. In this chapter, the discussed data pre-processing steps of chapter 4 are implemented in a real algorithm. But the most appropriate data mining technique for this algorithm is selected first, this selection is based on several prerequisites. Once the data mining technique is selected, the development of the algorithm starts. This development is divided into 2 parts, since the algorithm for gas uses more variables than the electricity algorithm. Therefore the electricity algorithm is explained first. The output of the algorithm is used to calculate the different parameters from the demands of the objectives. All these parameters are used within the report of the algorithm. At the end of this chapter, it is checked if the algorithm fits within the ISO 50001 requirements.

The implementation technique and communication technique which are linked with each other are discussed in chapter 7 & 8. Chapter 7 investigates the difference between an automatic control system and a semi-automatic control system. Based on requirements, advantages and disadvantages of the most appropriate technique is selected. The selected method is developed in the next sections of the chapter. This chapter also ends with a section which verifies if the implementation technique fits within ISO 50001. Chapter 8 describes how it is possible that the implementation technique of chapter 7 can communicate with the installed BMS. Several communication methods are explained and investigated. The most appropriate technique is selected at the end of the chapter.

Finally, there are 2 appendices. The first appendix explains how closed day anomalies are extracted from the dataset, which was introduced in chapter 6. This technique is used within this thesis but it is not especially developed for this thesis. The second appendix provides a more detailed overview of the tools developed by Porta Capena. All the explained tools are also applicable within the standard ISO 50001.

2 State of the art

The European Commission expects an increase of energy efficiency with 20% by 2020, but this goal is not binding to the member states. This target can only be achieved by companies and organisations who want to approach a low-carbon economy[1]. Such a low-carbon economy will result in:

- increased energy efficiency;
- energy savings;
- lower energy bills.

This approach can be supported by an energy management system. These systems help companies/organisations with the improvement of their energy efficiency and with finding opportunities for energy savings. Furthermore, companies/organisations who want to strengthen their market position or want to use the best practice method, are designated to use ISO 50001. This energy management standard is a practical guide for adopting a policy, developing an action plan, finding opportunities, etc. Besides the standard doesn't put any constraints on the different requirements[2].
For instance, within the energy review part it is necessary that opportunities for improving energy performance are identified. But the way in which it is done isn't important for the standard.

Since the release in 2011, there are different tools developed for companies interested in ISO 50001. These range from eAssessment tools (to analyse if your company is ready for ISO 50001) to special trainings to implement the standard[3]. In this thesis special attention goes to tools for:

- energy reviews,
- energy baseline,
- detecting opportunities to save energy;
- analysing improvements.

Relevant energy data combined with important building parameters (such as opening hours, floor surface) are needed to run these tools. The appropriate tool for the analysis is selected based on gathered or available data. The result will depend on the selected tool, for instance the listed tools will provide information about:

- energy reviews
    - current energy sources,
    - present energy consumption,
- energy baseline
    - the reference level which is extracted from the energy review;
- detecting opportunities
    - prediction of the opportunities which will lead to energy savings;
- analysing improvements
    - checking present consumption against the defined baseline.

Depending on the tool it will provide information about the consumption or about the detected opportunities. However, it becomes more difficult if the same analysis must be made on holidays, which are interesting for the most office buildings and schools, since the standard tool shall provide information about a specific day or period. For instance, it is possible to select one date (13th of May 2014) and it is possible to select a fixed period (1st of May 2014 until 13th of May 2014). But it is not possible to select a dataset based on different independent dates. For instance, (the 1st or 2nd of May 2014; 21st of April 2014; 1st of January 2014; etc.). Furthermore, there is a chance that the standard tool provides wrong information, since the consumption on holidays is lower than on a normal day (see Fig. 1). So the standard tool will not detect the anomaly or the opportunity. Therefore, this thesis is about the development an algorithm that detects this kind of holiday anomalies.

3 Study of the basic principles

This chapter will introduce the 4 most important items which are used throughout this master's thesis. Section 3.1 explains the concept of energy management within the standard ISO 50001. As the introduction discussed, all the developments are done within the reference frame of the standard ISO 50001. Therefore, this section is the largest section in this chapter, because the reader must have a detailed overview of this concept to understand the upcoming chapters. Energy management is followed by an introduction of data mining in section 3.2. Chapter 4 and 5 uses this section as reference, because these chapters focus on the use of data mining in real cases. Furthermore, algorithms are introduced in section 3.3. This section discusses a new lay-out for algorithms that communicate with BMSs. Therefore, each item linked to the algorithm is explained in detail. Finally, the last section 3.4 investigates the use of KPIs for energy management. The definition is followed by an overview of KPIs (categories and important knowledge). Furthermore, it is analysed how many KPIs are necessary and how a KPI is created.

3.1 Energy management according to ISO 50001

A good top management is the architecture of a well working organization. They take care of an efficient operation, lead the organization to short-term successes and they design actions to approach long-term successes. However energy management has been forgotten too many times, even though there is considerable potential to save energy and reduce costs.

People in management are most of the time skilled in economics and techniques to improve the working of an organization, but they aren't skilled in energy. This can be a reason why energy management is often neglected. Nowadays organizations are aware of the importance of energy management. The pressure of increasing energy prices and climate change legislations, forces management teams to implement energy management in their organization.

A good energy management should be applied in a manner appropriate to the philosophy and scale of the organization. The main principles are much the same in different organizations, but opportunities and the level of energy management are specific to an organization. Therefore, office buildings and industrial buildings aren't comparable [4],[5],[6].

3.1.1 Definition

Energy management enables an organization to take a systematic approach in order to achieve continual improvement of energy performance, including energy efficiency, energy use and consumption [3].

3.1.2 Why energy management is important

Reducing energy use makes perfect sense on a business level: it saves money and enhances corporate reputation. The different reasons are divided in four groups and are discussed in the following paragraphs: environment, companies, households and enhanced reputation.

3.1.2.1 *Environment*

For several years we have been struggling with the climate change on earth with as main culprit, the rising CO_2 emission. Between 1906 and 2005, the average temperature on earth increased by 0.7°C, scientists expect that the average temperature will be risen by 3°C, by 2050 [7],[8].
Besides the analysis of average temperatures, there are several indications of climate change such as: more rainfall, changing snow patterns, increasing droughts and severe storms. These changes demonstrate the risks for nature and human beings.
The advantage is that we are responsible for these problems, so it is still possible to take measures. For example the European Commission has made 5 targets for the European union in 2020, one of the 5 targets is climate change and energy [9]. The target is:

- greenhouse gas emissions 20% lower than 1990 (the most important greenhouse gas is CO_2);
- 20% of energy from renewable energy,
- 20% increase in energy efficiency.

The reduction of greenhouse gases and the increase of energy efficiency go together with energy management. This is also important for companies, which is explained in the next part.

3.1.2.2 *Companies*

Companies are the largest consumers of energy, in the Netherlands 70% of the energy is consumed by companies and in Belgium 58% of the electrical energy is consumed by them[10],[11].
The name companies is wrong in this case, because it is to general. There is a large difference between companies and so there is a difference in the savings opportunities.

For example:
A manufacturer who processes raw materials to consumer products, these companies are energy intensive. For large companies, the government can impose a maximum for CO_2 emission, the right to emit greenhouse gases is called a carbon credit [12]. Companies need to pay for these credits. With one credit you can emit one ton of carbon dioxide. By optimizing their production it is possible to decrease their emission and energy usage. In that case they need to buy fewer credits and they help the earth against its climate change.

In contrast to manufacturers we have schools, office buildings and distribution centres. These companies don't produce any consumer products so it isn't possible to save energy from the production. Energy savings are possible by implementing an energy policy, this can go from a couple of simple rules to a full equipped management system.
A distribution centre can use a WMS with on top of it a system that fine tunes the process and reports to the top management. Schools and office buildings don't use a WMS but a kind of EnMS For them it is also possible to implement the same configuration.

On top the system delivers a lot of advantages and is based on ISO 50001[3]:

- The communication between top management and the energy management team will be more flexible, because the top management gets direct information about the energy performance and the performance of the EnMS.
- Development of the documentation shall be easier when the top management has a better understanding of the whole process.
- The development of energy performance improvements and operational control in the design of renovated or new facilities will be more advanced than before. For example the sustainability of the building will be much better.
- The top management gets a better understanding of results from the internal audit of the EnMS.
- Nonconformities shall be detected faster so corrective or preventive actions can come quickly into force.

All these advantages will lead to a better output from the top management.
In general every company has the same purpose, saving money. So that they can conserve their position on the market in a world with rising prices of raw materials and energy bills.

3.1.2.3 *Final users or households*

The reason to save energy is comparable to the one for companies, except that a household doesn't need a policy, commitments or targets. But that doesn't mean that they don't need to think about energy savings. People, who are building a new house or renovating a house, must meet a number of rules. EPBD is a ministerial decree used in Belgium for new houses. The level that you need to achieve depends on the date of your urban license, when you don't achieve the E level the government will give you a penalty. The government gives you advantages (lower taxes,..) if a lower level is achieved. For existing houses, Belgium uses the EPC and it is necessary when you want to sell or rent a house. In contrast with the EPBD, you can't get advantages or penalties.

A lot of existing houses have a poor infrastructure, this reduces our comfort. Think off windows with thin glasses or uninsulated roofs which are responsible for the cold feeling at home in the winter. This means that a poor infrastructure is the culprit of many heat loses. Therefore, upgrading the infrastructure means a reduction of loses and an improved comfort, all accompanied by energy savings.
The main reason to save energy for households is saving money and increasing their comfort.

3.1.3 A vision on energy management

Implementing an energy management system takes time, because several steps need to be taken before the system is operational. Besides implementation it will continue to develop as energy performance improves and attention moves to different issues. Therefore, it is only possible to develop a roadmap Fig. 3 with the main elements which are needed for successful management [3],[4],[5],[13].

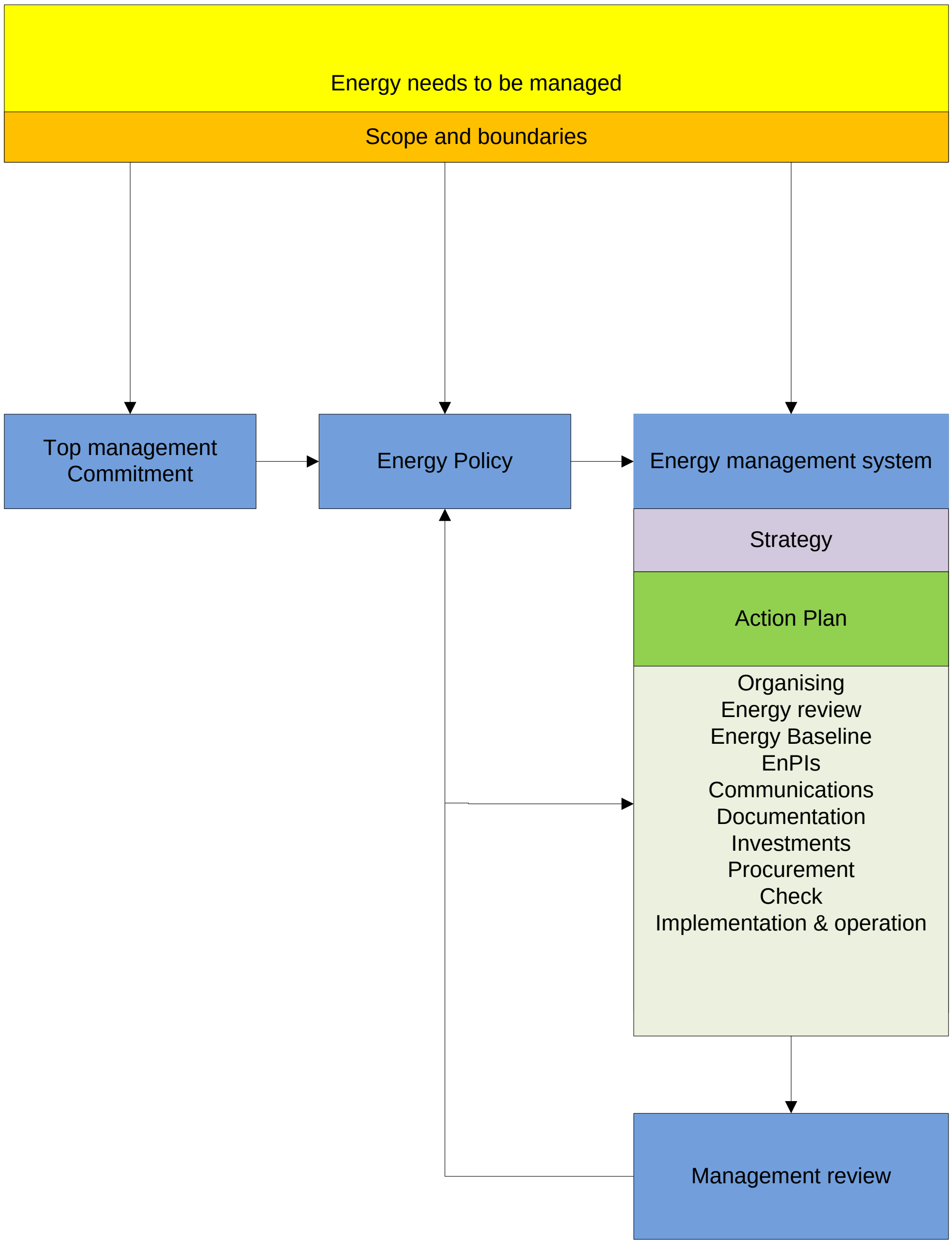


Fig. 3: Energy management roadmap

Energy management starts with a scope and boundaries which are defined by top management and linked to commitments. Top management will develop a policy, with attention for scope and boundaries. To approach the policy an energy management system will be developed. When the system is operational, top management will review the whole system at a planned interval. A continuous improvement and development of the system will be the outcome of this review. The elements of the roadmap of Fig. 3 are discussed in more detail below.

3.1.3.1 *Scope and boundaries*

The scope and boundaries of the energy management system allow the organization to focus their efforts and resources by defining the extent of the system. The scope may consist of: extent of activities, buildings, plants, facilities. The boundaries are physical or organizational limits. A couple of examples will give a clear sight on the boundaries. When a company has a data logger with a fixed sample interval of one hour, it isn't possible to sample each quarter. When it isn't allowed to install something on the roof, then this is a boundary if one want to install solar panels. It may be possible that an organization only wants to use gas for heating, so a solution with fuel would not be possible. Even the budget can be a boundary for some organizations. Top management identifies and documents the scope and boundaries of the organization's energy management system.

3.1.3.2 *Top management and commitment*

Top management hasn't a lot experience with energy management. Therefore it makes sense that they ask an energy expert to inspect their energy usage. The best way to explain issues is by giving a presentation, because it is important that they have a broad understanding of energy issues. Therefore, the first step is highlighting the discovered energy issues. These issues are followed by possible solutions. Convincing top management of the importance of energy management is the biggest target of an energy expert.

Several responsibilities for the top management:

- defining, establishing, implementing and maintaining an energy policy;
- appointing a management representative,
- providing resources (human, technology and financial);
- identifying the scope and boundaries.

They need to communicate the importance of energy management to the organization. Everyone needs to be aware of energy and needs to change their behaviour. Otherwise the project is doomed to fail.
Top management is responsible for actions to improve the system, this is discussed later in the part of management review.

3.1.3.3 *Energy policy*

A well developed policy provides the foundation of best practice in energy management. It must be relevant to the organization's nature and size. Therefore, the policy establishes and validates top management's commitment to energy performance improvement.

It is defined and documented by top management and must be understood by employees and anyone on behalf of the organization. The policy includes commitment to:

- continuous improvement,
- ensure the availability of information and resources;
- comply with applicable legal requirements,

- support the purchase of energy efficient products and services;
- the sample interval and make someone responsible for taking the manual readings.

The policy also provides the framework for setting and reviewing energy objectives and targets and is consistent with the strategic plan. The objectives and targets includes:

- evaluate energy reduction targets to determine the reinvestment time;
- improving management skills and identifying efficiency improvements;
- expressing objectives and targets in a clear and simple way.

Most of these are expressed in percentages, they need to be realistic and reflect the organization's potential. Another way to approach their targets is by defining KPIs (see section 3.4).

The policy must not be large, certainly no more than two pages. It needs to address all the items mentioned above and be relevant to the defined scope of the energy management system.

3.1.3.4 *Strategy*

This will be the first place where the policy is applied, because the policy on its own will not deliver energy savings. The size of strategy depends on the size of organization. For a small organization a robust energy plan is enough, but for a large organization a strategy is essential. Such a plan is a formal written document adopted by the management team.

Strategies should not be developed by one person or by an external organization, otherwise a lot of things will be forgotten. It should be developed by a group of employees, managers, top management and stakeholders. A group which consist of different functions will reach a higher quality, instead of one person or an external organization. A group is using different kinds of knowledge (practical, theoretical, management, logistics,..) which is gathered throughout the organization. These different kinds of knowledge are necessary for the development of a good strategy.

The strategy provides the basis for the action plan of an energy management system. There are several strategies which may be taken into account or which are interesting. For top management it is important that they know what their opportunities are. There are three possible strategies. A detailed version is discussed below.

a) Awareness of energy usage

This may be the least promoted type to save energy, is making people aware of their energy usage. People can think about their energy consumption, but that doesn't mean that they will change it, although it is very important. Energy savings up to 10% are possible with a minimum of costs [14].

For households this type can be very simple. If you take a note of your energy consumptions every week and put them in a spreadsheet, after a couple of weeks it is already possible to compare the results. If there is an increase of energy usage, you can think about its reason and plan some actions for the future. Examples of these actions are:

- turn the lights off if you aren't in the room;
- turn devices off, if possible,
- fine tune the working schedule of the heating;
- take the bike for short trips,
- prevent faucets from leaking.

In industry, awareness of energy usage is more complex, so they don't work on the same way as households. It starts with the communication of the initiative to employees and other parties such as onsite contractors and suppliers, therefore top management takes the lead [15]. After the reveal of the initiative for energy saving, the development of an energy awareness program may start. Such a program is used to inform employees, contractors and suppliers who work at the organization. It consists of: trainings, slogans, personal interviews, presentations and so on. As the energy management system continues to develop, there are additional items that employees and contractors will need to know. For example, everyone needs to know their responsibilities within the system.
The benefits of improved awareness fade with time. So it is necessary to refresh your awareness campaign. The ideal moment may be determined out of energy monitoring and targeting schemes. This is the fastest way to discover a loss of momentum.

b) Investments

This part refers to investments which upgrade our infrastructure, in such a way that an old device or tool will be replaced with a new one with the same or more functions. These kinds of investments will lead to a better infrastructure.
And this will lead to energy savings and a better comfort. You will find a couple of these investments below.

<u>Insulation of the cavity wall:</u>
Nowadays cavity walls aren't used anymore in the design of new buildings, the cavity wall is replaced by thick insulation plates. But buildings built in the past still have a cavity wall, which means that there are more heat loses trough the wall. To prevent these loses it is now possible to insulate your cavity wall with foam insulation.

<u>Installation of energy efficient lamps:</u>
The further development of technology is responsible for the new kind of lighting techniques. These techniques make energy saving possible with almost no lose of comfort. The standard fluorescent lamps in schools and offices are replaced by fluorescent lamps working with a high-frequency electronic ballast. Another option

is to use a fluorescent lamp with integrated ballast, it uses 35% less energy and loses 10% of its light intensity [16].
Compact fluorescent lamps or LED bulbs are good alternatives to incandescent lamps or halogen lights. These lamps have struggled a lot with the possibility to dim, so the dimmable lamps are only recently available on the market.

Installation of new heating techniques:
When a new building is designed, you can chose for a heat pump or a CHP. These types of heating's have some special requirements that it's easier to implement them in a new project and not in a renovation. They have a positive impact on your E level, the meaning of E level is explained in the part 'why saving energy'. A high efficiency heating or condensation heating may be selected for a renovation. For example an average household uses 2500 litres gas oil a year, by installing an condensation heating, the gas oil consumption can be reduced to 1500 litres a year.

Installation of water saving faucets:
Day in day out we are using water, so saving water can have a major impact on our water bills. The best places to install a water saving faucet is in the bathroom and in public toilets, because in these places we are using water to wash our body or our hands so the flow of water doesn't matter. These faucets save water by mixing air in the water without losing comfort so the total flow of water is decreased [17].

c) Data mining

Data mining in general is explained in section 3.2. Here it is discussed how data mining is a kind of energy saving.

A lot of companies collect data of their energy usage. Using a system which can analyze data and has an algorithm to reduce energy consumption is ideal to save energy. Beside energy saving, it is possible to identify energy wastes and to benchmark different buildings/places against each other.
Most of these applications are based on a SCADA system which is web based.

Buying such an application is also an investment. But it is surely different from investments mentioned above. Here it is possible to save energy over several energy types (water, electricity, fuel,..) and you don't need to replace devices or tools. It is just a system placed on top of the installation with fine tuning capabilities.

3.1.3.5 *Action plan*

The management identifies which actions are necessary to put energy management into effect, by the development of a strategy. The action plan describes the vital elements that need to be established. The vital elements need to be consistent with the policy and the strategy.

The format of the action plan will depend on what works for the organization, rather than reflecting a standard approach. When an organization already has a planning or project management process, it will be better to use this structure, instead of inventing a new one.

a) Organising

Organising energy management is an important element, it decides whether the energy management will work or not. Therefore, top management appoints a management representative and approves the formation of an energy management team. This team is assembled by the management representative. It is the core group which is responsible for delivering the energy policies, objectives and implementing the energy strategy across the organization. It is important that they have sufficient time, expertise and resources to perform their energy management responsibilities effectively. Time needed for good energy management depends on the nature and size of the organization. Large companies with high energy uses, often benefit from having a full-time energy management team. For smaller companies, this might be a part-time role. So they may also have other functions and duties.
The energy management team should fulfil the following functions:

- ensure that the system is established, maintained and continually improved;
- communicate and report to top management on energy performance;
- ensure that energy management actions are consistent to the policy;
- determine criteria and methods needed to ensure that the operation and control of the system are effective;
- promote awareness of the policy and objectives at all organizational levels.

b) Energy review

Most of the time, the organization does not have any idea of energy usage. So an energy review establishes a first indication of the amount of usage. Top management uses the energy review in the development of the policy. It gives an idea on the possible energy reductions. Later on it will be a basis for the energy baseline.

The following elements are important for a relevant review:

- identify energy sources,
- evaluate past and present energy use by analyzing measurements and other data;
- identify the areas of significant energy use, a special area may be the different working shifts in larger companies;
- identify and record opportunities to improve energy performance;
- estimate future energy use and consumption.

c) Energy baseline

The outcome of the energy review will be used to develop an energy baseline. It will be used by the management team to detect changes in energy performance. When they change, it is possible to establish some adjustments.
These are made:

- after major changes in the process,
- according to a predetermined method,
- when EnPIs no longer reflect organizational energy use and consumption.

The development of an energy baseline starts with defining a relevant period for the organisation and the analysis of data. This data is delivered by the energy review. To become a relevant baseline, it is necessary to apply a couple of techniques. The chapter data mining will discuss this item in more detail.

d) EnPIs

EnPIs are used for monitoring and measuring energy performance. There are several options to visualise this indicator:

- a simple parameter,
- a simple ratio,
- a complex model.

An EnPI visualises the result of the KPI in energy terms. Therefore, an EnPI might be a visualisation of the energy consumption per time, per unit of production or other relevant models. In order to check the energy performance, the selected EnPI should be recorded and regularly reviewed by the energy representative.

e) Communications

Good communications throughout the management structure is essential for ensuring that the whole organization works as a team to implement an energy management system. Fig. 4 shows a possible communication flow.

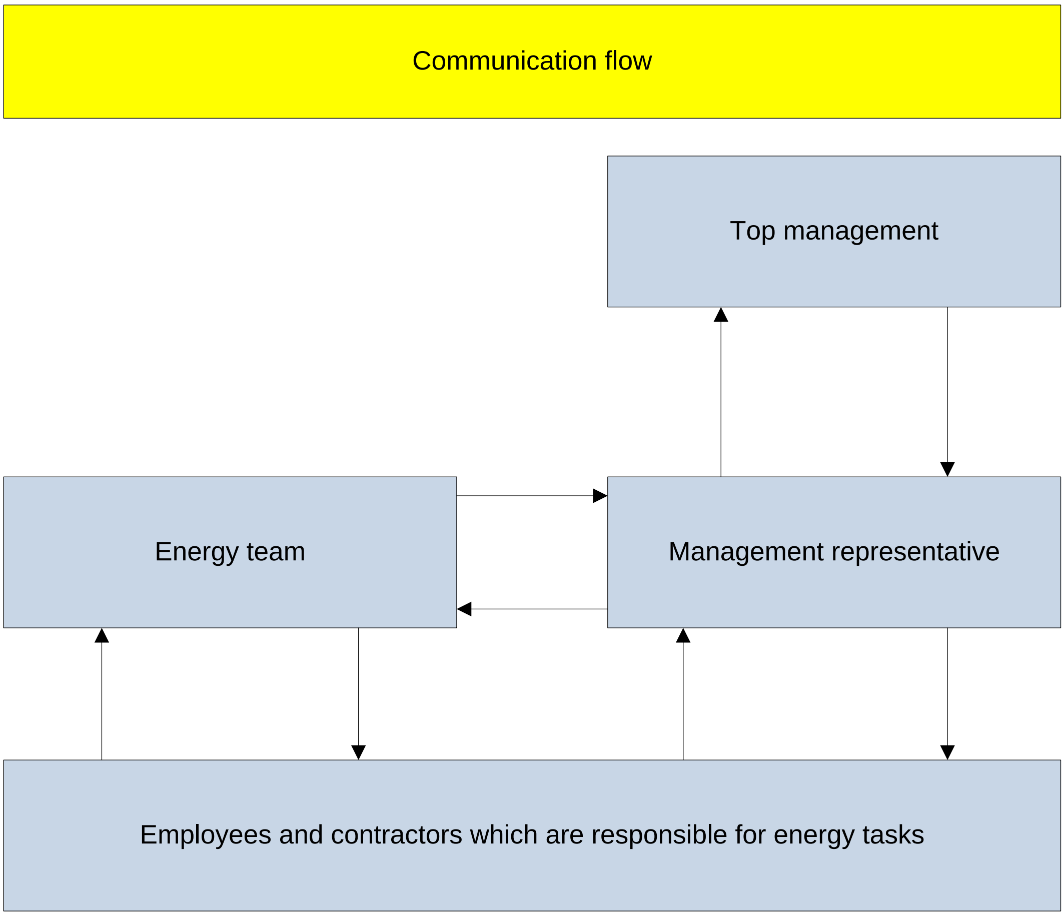


Fig. 4: communication flow

The management representative is responsible for monitoring the progress of the implementation effort and this can only be accomplished through communication with the energy team and employees. If the energy team or employees reveal an issue, this must be reported as fast as possible to the management representative. He is also the contact person of top management. The management representative must report to top management on energy performance and on performance of the energy management system. The energy team brings the policy into action and helps the employees approach their energy tasks. Therefore, a good communication between these two is convenient and important.

Improvements in communication are possible by implementing meeting schedules. A formal schedule helps ensure that these meetings are conducted. The energy team or employees shouldn't wait until a meeting to report an issue. Documenting each meeting is not necessary or practical. It depends on the nature and size of the organization, but verbal communication may also be adequate. The important thing is that the information gets communicated and issues are resolved in a way that the implementation effort is successful. Implementing a process by which employees can make comments or suggest improvements, will be useful to brainstorm and to develop new adjustments.

f) Documentation

Documents are necessary in the development, implementation and control of an energy management system. These demonstrate what has been done and provide evidence of the results that are achieved. It will be used as a guide for the management team, as well as the management representative.
Information is fundamental to any organizational process. Documentation is an example of communicating information. It may be in any format or type of medium, including electronic, magnetic, books, presentations, print, and other. In the whole process of documentation, there are two basic types which are surely necessary.

- documents
  These are focused on the present day. They are used to implement policies and commitments. The documents may be a guide for behaviour, activities and operations. Therefore, they provide answers to common questions like, when, what, who and sometimes how.
- records
  They reflect past history. Records visualize what was done or what was accomplished. They are used in the development of an energy review or energy baseline.

It isn't always clear which kind of documents are necessary to become a good energy management system. A guide in this approach may be the ISO 50001 norm and all its requirements. The ISO norm is very lean on the number of required documents. Therefore, there are no explicit requirements for documented procedures. There are only a couple of documents/records required by ISO 50001. This means that most of the decisions of a document/record that is needed, are left to the organization. Fig. 5 shows an example of these documents/records.

Documents and records needed by ISO 50001	
Documents	**Records**
• Scope and boundary of the EnMS • The audit process • Energy policy • Energy objectives • Energy targets • Energy management Strategy • Method to develop the energy review • Action plans • Decisions for communication	• Energy review • Energy baseline • Method for determining and updating the EnPIs • Records of competency • Records of needed trainings • Records of training • Results of the design activities • Results of calibrations • Results from monitoring and measurement • Results of the audit • Results of the management review

Fig. 5: required documents and records by ISO 50001

The degree of documentation will vary for different organizations. A large organization tends to use documentation more extensively than smaller organizations. Other reasons may be the complexity of the processes and their interactions or the competence of personnel. The minimum extent of documentation is given by the ISO 50001 norm. Nevertheless it is still possible to add additional information. Therefore, it isn't possible to provide the exact number of documents.

Sometimes it isn't clear which kind of additional information is necessary. It may be useful to establish rules of thumb or other criteria. These other criteria may be legal requirements, customer requirements, or prior nonconformities. Top management need to avoid to overlook the potential to leverage existing procedures and other documents, modifying them to include energy management. This makes sense and is a resource-effective approach of documenting information for the EnMS without creating new documents to be managed.

g) Procurement

There are two types of procurement which are critical in energy management. One is the procurement of energy itself, the other is procurement of energy using equipment.

1. Energy procurement
   It should represent only a small part of the energy management. The most important part of energy management is the reduction of energy. It is about getting the best price for the supply of energy. So effective procurement of energy may not save energy or carbon but it can reduce energy bills. When an organization wants the best price for energy supply, it is important that the responsible person has knowledge of the energy usage. Without knowledge of how and when the organization uses energy, the person can be approaching the market as a disadvantage. Because there are several formulas on the market to buy energy. For instance, a permanent contract where a fixed price is paid for the energy, but it is also possible to work with a click and de-click system. In such a system, an amount of energy is bought at a specific moment for the price of that moment (click). When the energy prices decrease, it is possible to sell the energy which isn't used(de-click). The process of clicking and de-clicking is a chain. After clicking, it is possible to de-click and after de-clicking it is necessary to click. Therefore, it is important that the companies consumption profile is known when the market is approached.
2. Procurement of equipment
   Rather than energy procurement this is a more important part in energy management. If a device is old or fails, it will be replaced by another one. This doesn't mean that the new device will save more energy. So the energy performance of the organization is influenced by the choices made regarding the procurement of equipment that consumes energy.
   The best way to approach good procurement is by developing procedures. These ensure that energy performance is taken duly into account when procurement decisions are being made.

h) Check

The EnMS is developed for continuous improvements. Continual improvement is also a commitment which is taken into account in the policy. Therefore, it is surely necessary that it is accomplished. For that reason the check part is necessary in the action plan. This part ensures that appropriate monitoring and measurement activities are planned and installed. Furthermore, this parts will evaluate if the system is operating in line with the policy and planned objectives. Therefore, the system is regularly checked to determine if the system meets the commitments. If commitments aren't met, plans are developed or modified. In such a way that the organization can meet these commitments and continually improve energy performance.
This part is concerned about:

- monitoring,
- measurement,
- analysis,
- evaluation of legal requirements and other,
- internal audit of the EnMS,
- nonconformities,
- correction, corrective and preventive actions;
- control of records.

3.1.3.6 *Management review*

The action plan will help to keep the energy management strategy on track, but it will come to a point that a more detailed assessment is needed. At that point the management review is used. This review is made to ensure the ongoing suitability, adequacy and effectiveness of the EnMS. The results are fact-based decisions, which are accomplished by analyzing specific information about the system and its performance. Fig. 6 shows the three most important things for an efficient review.

What you need for an efficient review

Input

- Output of the previous review
- The energy policy
- Review of energy performance and ENPIs
- Evaluation results of compliance with legal requirements
- The energy objectives and targets
- EnMS audit results
- Status of previous actions
- Projected energy performance of the following period
- Recommendations for improvement

Decisions to be made

Which strategic changes or modifications are needed?
Are there changes in external requirements that will affect the system?
Are there changes internally that will affect the system?
Do the current measures provide the correct information?
Is there a need to change, add or delete any objectives?
What changes are needed or expected to improve the system?
Is the system suitable for the organization?
Is the system effective?

Output

- Changes in the energy performance of the organization
- Changes to the energy policy
- Changes to the EnPIs
- Changes to objectives or targets
- Changes to the strategy

Fig. 6: management review

The management representative is responsible for the input. Therefore, he ensures that the appropriate information is collected, organized and presented in a way that enables management to make informed decisions. This doesn’t mean that all the information for decision making comes from one person. The management representative identifies documents/records as necessary for the review. The next step is to know who has the needed information. For the representative this should be simple, because he is responsible for defining responsibilities and authorities throughout the EnMS. Lastly the management representative must organize the collected information for the presentation. The information should show the big picture and should be detailed enough to allow management to make fact based decisions, which are defined in the output.

The best way to approach a good energy management is to schedule a review at intervals of three or six months. A full analysis is recommended every year, this keeps the strategy on course. For this analysis it is important that everyone is involved, from the top level right through the organization. This may be useful to develop new energy saving strategies or other management strategies.

3.2 Data mining

Analyzing large amounts of data is a necessity. The fact that data mining has been discussed previously, confirms this statement.
We are deluged with data: scientific data, medical data, financial data,.. This explosive growth in stored data is the reason of the need of new techniques and automated tools [18]. Therefore, several automatic tools are needed to analyze data, classify , summarize and characterize trends in the data.

3.2.1 Definition

Data mining is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases, the Web,.. [18]

3.2.2 What is data mining

Data mining can be defined in many different ways. So even the term data mining doesn't present all the major components. For instance the mining of coal from rocks and sand, is called coal mining instead of rock or sand mining. Analogously, data mining should have been more appropriately named, knowledge discovery form data or KDD. This term is in fact a process and data mining is one of the steps in the process. Fig. 7 visualises the process of data mining [18].

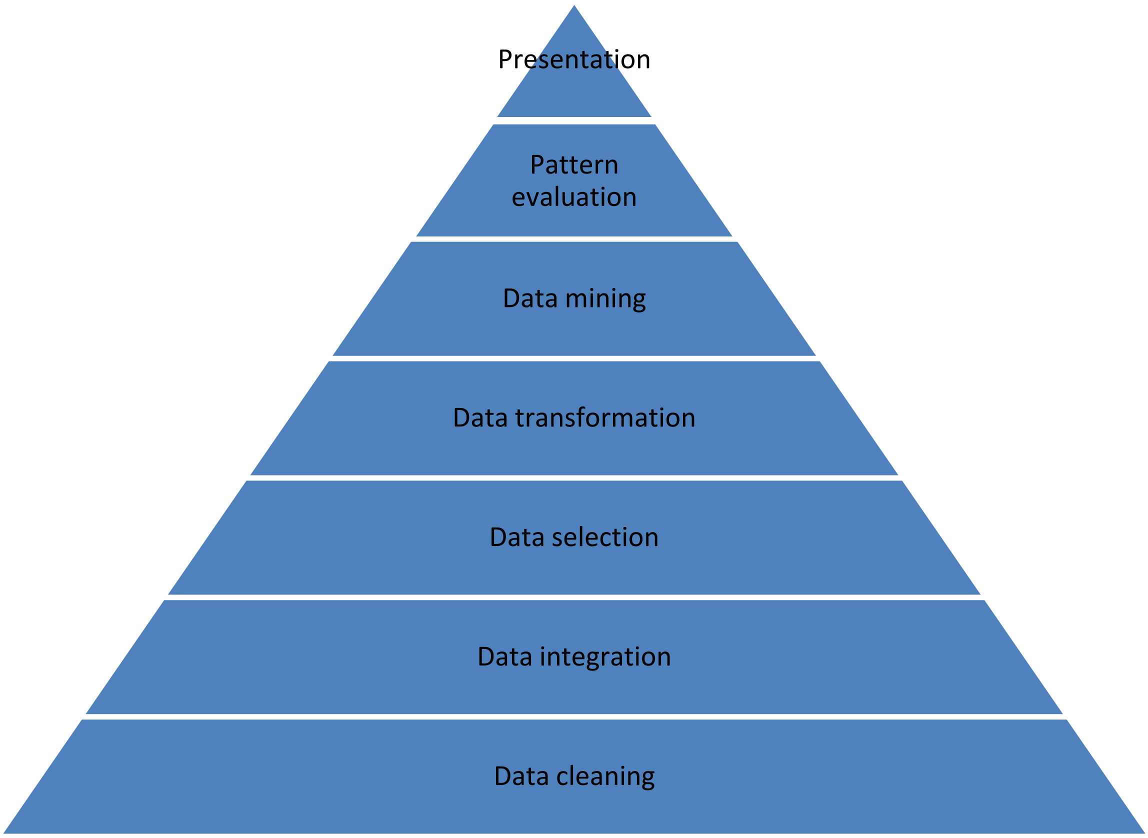


Fig. 7: data mining steps

1. Data cleaning: to remove noise and inconsistent data
2. Data integration: combine several data sources
3. Data selection: retrieve relevant data for the analysis from the database
4. Data transformation: data are transformed and consolidated into forms appropriate for mining
5. Data mining: an essential process, intelligent methods are applied to extract data patterns
6. Pattern evaluation: to identify the truly interesting patterns representing knowledge based on the use of specified models
7. Knowledge presentation: minded knowledge will be presented to users

Steps 1 through 4 are different forms of data pre-processing. In these steps the data is prepared for mining in step 5. Although data mining is a process step here, the term data mining is often used to refer to the entire knowledge discovery process.

3.2.3 Kinds of patterns that can be mined

The purpose of this thesis is to implement an EnMS. Therefore there is a necessity for data. To become an efficient EnMS, it is important to discover interesting patterns. These patterns may for example visualize a higher energy consumption than necessary. However, in order to find these patterns, knowledge of patterns which can be mined is necessary. The next part discusses all these patterns, so that a relevant pattern may be chosen later [18].

3.2.3.1 *Class/concept description: characterization and discrimination*

Data entries can be associated with classes or concepts. For example, in an HVAC store, classes of items for sale include: boilers, valves, faucets. Concepts of customers include: people who buy spare parts and people who buy a new device. So it can be useful to describe individual classes and concepts in summarized, concise and yet precise terms.

a) Data characterization

This is a summary of the general characteristics or features of a target class of data. For example, a manager of a store orders the following data mining task: Summarize the characteristics of customers who spend more than €2000. The result is a general profile of these buyers. This result can be presented in different forms.

b) Data discrimination

This is a comparison of the general features of the target class data objects, against the general features of objects from one or multiple contrasting classes. For example, an engineer may want to compare the general features of lamps. He wants to compare lamps with a reduction of energy usage by 10% against lamps with a reduction of energy usage by 20%, during the same period. The way to present the results is similar to that one used for data characterization. However discrimination descriptions should include comparative measures that help to distinguish between the target and contrasting classes.

3.2.3.2 *Mining frequent patterns*

These are patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent item sets, frequent subsequence's.

a) Frequent item sets

Refers to a set of items that often appear together in a transactional dataset [2](is the information recorded from transactions in case that several product are sold together). For example, a data logger and sensors, which are frequently bought together in electricity stores by many customers.

b) Frequent subsequence's

Is also known as sequential patterns or the fact that each action is followed by another action. A frequently occurring subsequence is the pattern that customers, tend to purchase a laptop first, followed by a tablet and then a memory card.

Mining this kind of patterns leads to the discovery of interesting associations and correlations within data.

---

[2] Is the information gathered from transactions, when several products are sold together.[19]

3.2.3.3 *Classification and regression for predictive analysis*

Classification is the process of finding a model or function that describes and characterizes data classes or concepts. The model is developed based on the analysis of a set of training data (class-labelled). Such a model is used to predict the class label of objects for which it is unknown. The output of this model can be applied in different forms, such as: if-then rules, decision trees, mathematical formulae, or neural networks. Therefore, classification is used to predict categorical labels (discrete, unordered), while regression models continuous-valued functions. This means that regression is used to predict missing or unavailable data values, rather than (discrete) class labels.
Classification and regression may need to be preceded by relevance analysis, which attempts to identify attributes that are significantly relevant. Such attributes will be selected for the classification and regression process. Irrelevant attributes can be excluded from consideration.

3.2.3.4 *Cluster analysis*

Clustering analyses data without the necessity of class labels. Therefore clustering is very useful, because most of the time data is not class labelled in the beginning. So clustering can be used to generate class labels. Fig. 8 shows an example of several clusters.

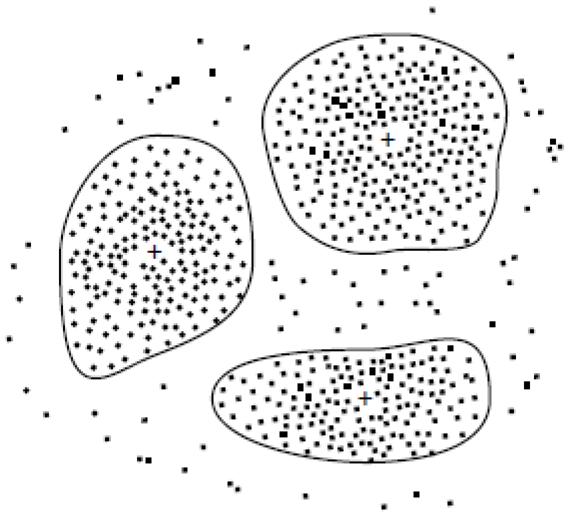


Fig. 8: clustering

Source: Data mining concepts and techniques [18]

The objects are clustered or grouped based on the principal of maximizing the intraclass similarity and minimizing the interclass similarity. Therefore, clusters will contain objects with a high similarity. The elements in such a cluster are dissimilar to elements in other clusters. So each cluster can be viewed as a class of objects, from which rules can be derived.

3.2.3.5 *Outlier analysis*

A data set may contain objects which don't fit in the file, these data point are outliers or anomalies. Normally, this kind of data points are removed from the data set, because they aren't reliable for analysis(impossible values, problems during data collection, etc.). However, in some applications the rare events can be more interesting than the more regularly occurring ones. For example, in the research for energy waste. Most of the time energy is wasted at periods at which more energy is used than necessary. So the energy usage at these moments will be measured as outliers.

3.2.3.6 *Are all patterns interesting?*

A data mining system has the potential to generate thousands or even millions of patterns. However, not all patterns may be interesting. Only a small fraction of the patterns would be of interest to a given user. But what makes a pattern interesting? There are several requirements for a pattern to be interesting.

1. It must be easily understood by humans
2. Valid on new or test data with some degree of certainty
3. Potentially useful
4. It must be new

In addition, a pattern is also interesting if it validates a hypothesis that the user sought to confirm. In general an interesting pattern represents knowledge.

3.2.4 Technologies that are used

The introduction already mentioned, that there is a need for new technologies. Therefore knowledge of other domains such as statistics, machine learning, database systems and information retrieval has been used to develop new techniques and methods. This section gives some examples of disciplines that are used in data mining [18].

3.2.4.1 *Statistics*

A statistical model is a set of mathematical functions that describe the behaviour of the objects in a target class in terms of random variables and their associated probability distributions. As a result this model can be the outcome of a knowledge discovery task. Alternatively, data mining tasks can be built on top of statistical models. This means that statistics can be applied in several forms. For example, we can use statistics to find outliers or gaps in data. However statistics can also be used to verify data mining results, by applying a statistical hypothesis test.

3.2.4.2 *Machine learning*

This item investigates how computer programs can learn based on data. The purpose is to teach a computer program how to detect complex patterns in data. A typical machine learning problem is programming a computer in such a way that it can automatically recognize handwritten postal codes on letters, after learning from a set of examples. There are several types of machine learning:

- supervised learning,
- unsupervised learning,
- semi-supervised learning,
- active learning.

Unsupervised learning is a technique which can be used for outlier detection. Therefore, this type is discussed in the chapter of data mining techniques.

3.2.5 Major issues in data mining

Data mining is a dynamic and fast expanding field with great strengths. Different issues have risen, which have been addressed in recent data mining research and development. The issues continue to stimulate further investigation and improvement in data mining. The major issues in data mining are listed below [18].

3.2.5.1 *Mining methodology*

Because data mining is fast expanding, researchers have been developing new data mining methodologies. This involves the investigation of new kinds of knowledge. Let's have a look at these various aspects of mining methodology.

- mining various and new kinds of knowledge;
- mining knowledge in multidimensional space;
- data mining, an interdisciplinary effort;
- boosting the power of discovery in a networked environment;
- handling uncertainty, noise, or incompleteness of data;
- pattern evaluation and pattern, or constraint guided mining.

3.2.5.2 *User interaction*

The user plays an important role in the data mining process. A couple of important items will be introduced here.

a) Interactive mining

Doing data mining interactively will help to discover new patterns. In this case is interactive associated with flexible. A flexible user interface makes it possible to dynamically change the focus of a search.

b) Incorporation of background knowledge

Background knowledge, rules and other information regarding the domain under study should be incorporated into the knowledge discovery process. This knowledge would be useful for the evaluation or it will be a guide for the search towards interesting patterns.

c) Presentation and visualization of results
d) Results of data mining must be presented in such a way that it is easily understood and directly usable. So it requires the system to adopt expressive knowledge representations, user-friendly interfaces, and visualization techniques.

3.2.5.3 *Efficiency and scalability*

Data mining algorithms must be efficient and scalable, in order to effectively extract information from huge amounts of data in many repositories. In other words, the running time of data mining algorithm must be predictable, short and acceptable for applications. So efficiency, scalability, performance, optimization, and the ability to execute in real time are key criteria that drive the development of many new data mining algorithms.

3.2.5.4 *Diversity of database types*

Like already mentioned, we are deluged by data. This means that there is a wide diversity of data types. It is logical that there are not data mining tools developed for all these types, because the data type will differ for different data mining applications. Therefore, systems will be developed domain specific and in such a way that in-depth mining isn’t a problem. However the construction of tools which can discover knowledge from different sources poses great challenges to data mining.

3.2.5.5 *Data mining and society*

It is for sure that data mining is a necessity. Although discovering knowledge out of data may have an impact on our society. Therefore, it is useful to mention the next cases.

a) Social impacts of data mining

With data mining penetrating our everyday lives, it is important to study the impact of data mining on society. How can we use data mining technology to benefit society? How can we guard against its misuse? These questions must be explored in representing studies.

b) Invisible data mining

Expecting that everyone (each person) will know data mining techniques isn’t possible. Therefore, it is necessary that systems will have built in data mining functions. Thanks to this kind of integrations, people can perform data mining or use data mining results simply by mouse clicking, without any knowledge of data mining algorithms. For example, search engines on the internet are using invisible data mining techniques.

3.3 Algorithm

Top management must take decisions to make a building energy efficient. These decisions are used for the development of a strategy. This is discussed in more detail in the part about energy management. But one of the items was data mining, data mining explained that algorithms may be used to approach efficient energy management. Nowadays there are algorithms for almost each thinkable problem. Therefore, it is important to document and discuss the details of an algorithm. There exist a lot of algorithms each dedicated to an application. Therefore, this section will discuss the general concept of an energy algorithm. The concept is visualized in a figure, following the definition of an algorithm [20].

3.3.1 Definition

"*An algorithm is any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output. An algorithm is thus a sequence of computational steps that transform the input into the output.*" Introduction to algorithms; Cormen, Leiserson, Rivest, Stein: second edition [20]

3.3.2 Energy management algorithm

The introduced principle will focus on energy management for buildings. Therefore, it cannot be used for production processes or other purposes. Fig. 9 shows the roadmap of an energy management algorithm for buildings. Each separate block in the figure is explained in more detail. Furthermore, examples within these sections refer to HVAC systems.

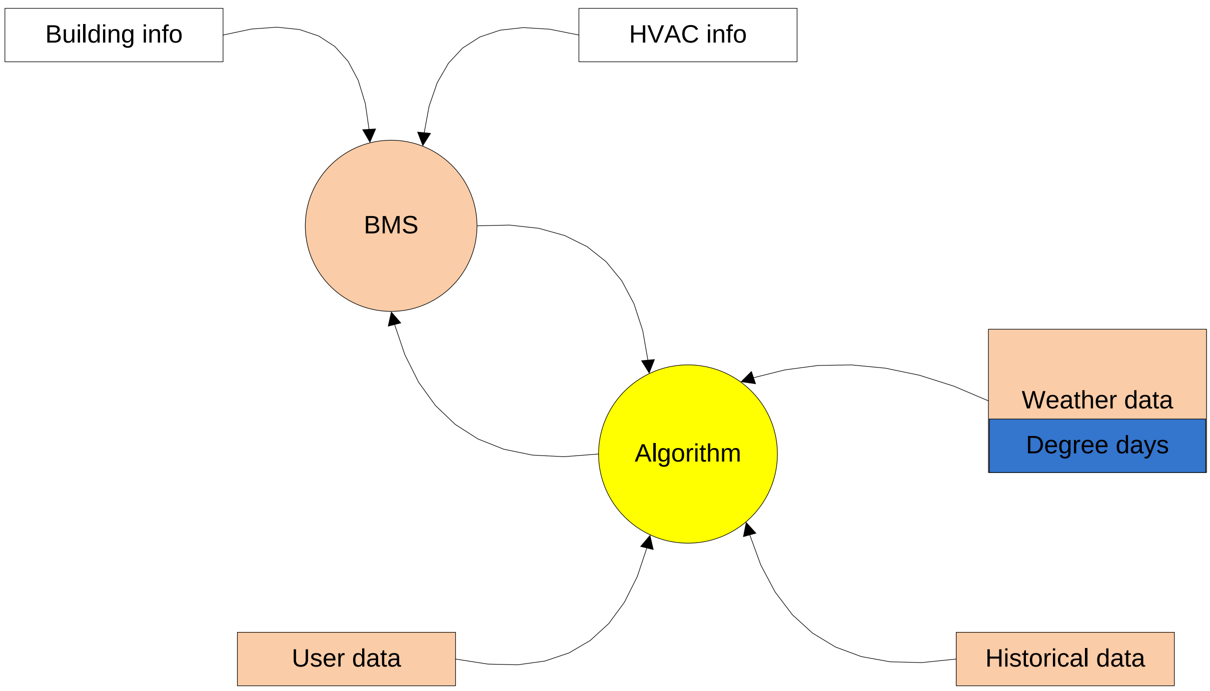


Fig. 9: Energy management algorithm roadmap

This concept of Fig. 9 is based on four parameters. These parameters are providing the input of the algorithm. When the algorithm receives input information, it is ready to produces some value as output. The output of the algorithm depends on the purpose of the algorithm. However, within this kind of algorithms, the output is always directed to the BMS. Furthermore, it is more a rule than an exception, that several algorithms are used in an EnMS.

For example, an algorithm is implemented to determine the ideal start-up time of the heating. Another algorithm can be implemented to prevent that the installation will work out of control. So it is for sure that one algorithm isn't enough to approach efficient energy management.

In general, the input parameters are pretty much the same for different outputs. Therefore, each parameter will be discussed in detail. It is important that correct information is provided since wrong inputs provide wrong outputs, this kind of faults must be avoided.

3.3.2.1 *User data (left bottom of Fig. 9)*

The algorithms are developed for buildings, most of the time for offices or schools. The purpose is to approach efficient energy management. Efficient may stand for not losing comfort, because if there is a loss of comfort, people in the building will complain. This means that people have an important role in the development of such an algorithm. Therefore user data is a parameter of the concept. User data is related to the occupancy of a building, because comfort must be achieved at moments when people are in the building. In this case comfort is associated with temperature. Therefore, the user provides two kinds of data.

1. Time schedules
   The algorithm needs to know the opening hours of the building. Besides this data, it must also know when the building is closed for holidays.
2. Desired temperature
   Since comfort is associated with temperature, someone must define a desired temperature as setpoint (or a band of temperatures). This temperature will be used as a baseline.

User data may have an effect on the whole HVAC cycle, because heating and air-conditioning will affect the temperature. Besides the moment when the desired temperature must be achieved, time schedules are needed for the correct functioning of ventilation systems. When people occupy a building the CO_2 will raise. The ventilation is responsible for the insertion of clean air. This system works in such a way that the CO_2 stays under control. In general CO_2 is expressed in parts per million, the baseline value may be 800 ppm. Since studies have proved that the number of faults that are made during a test is linked with the CO_2 level [21]. Therefore, this value is ideal for maximal knowledge transfer.

Besides how and on what user data has affect, it is important to understand that a building consists of several rooms. Therefore, it is possible that a couple of rooms are controlled separately. For each room user data is necessary. Because the algorithm needs to know when the comfort of these rooms needs to be achieved and if the comfort level is already achieved.

3.3.2.2 *Historical data (right bottom of Fig. 9)*

This kind of data is always presented as datasets. These datasets are pre-processed, so that only relevant data enters the algorithm. The algorithm will use this data for knowledge discovery. There are several kinds of knowledge, these are discussed in the data mining part. For instance the information retrieved from these datasets, can be used to make predictions on what is going to happen. Historical data can also be used for the development of baselines or as an indicator of good/bad control. An example will show how historical data may be used in energy management algorithms.

For instance an algorithm to prevent the out of control of the heating: assume a not programmed holiday in a building where gas and water consumption is logged. The algorithm used in this example will use the data of the gas and water consumption as calculation parameters. Gas data provides information of the working times of a heating system. While water data will be used to detect occupancy of the building.

Based on this explanation, the heating start-up like every morning. In general the algorithm is programmed that the output is generated after 10 a.m. Until then it is using historical data for calculations.
The output is the result of a comparison between a reference value and the actual value. When the water usage based on the historical data for that day, is nearby or below the defined baseline value. The algorithm will perform a switch off action,

because it has detected that the heating started up while the building wasn't occupied.

3.3.2.3 *Weather data (right top of Fig. 9)*

Weather data is used for predictions and for calculations relating to a building energy consumption. There are two types of weather data which are important for an EM algorithm. The first type will be weather forecasting and the second will be DD.

An efficient algorithm uses data from weather forecasts. Each night around 12 pm the weather prediction is loaded from a station close to the building. Once the data is loaded, reliable information for the next day is available for the algorithm. As a result it uses the information to predict what's going to happen the next couple of hours. These predictions may affect the decisions which the algorithm must take.

For example: an algorithm is developed for calculating the optimal starting point of a heating system. Heating will be necessary if the room temperature is below 20°C. Furthermore, the weather forecast provides a sunny day with temperatures above 25°C.
If only the room temperature is analysed, heating will be necessary from the moment on that the temperature is 19°C. If the algorithm also analyses the weather forecasting, it will predict that the temperature will rise fast. This means that the building temperature automatically will rise to a higher value, probably higher than 20°C. Based on this prediction, the algorithm decides that heating isn't necessary. This is just one example of how weather forecasting can be used in an EM algorithm.

3.3.2.4 *Degree days (right top of Fig. 9)*

Degree days are a special type of weather data. But they are also related to historical data, because the data which degree days provide is analysed each week or month. DD are calculated from readings of outside air temperature. These are often used for calculations which are related to building energy consumption. The three main types are: heating degree days, cooling degree days and growing degree days, all these are similar. If you understand one, all of them will be clear. HDD will be explained in detail, CDD and GDD in general and a word will be spend on the problems and opportunities [22].

HDD principle and calculation method:
HDD visualize how much (in degrees), and how long (in days), the outside temperature was below a certain level. This level will differ by the classification of the building. A building which is used as a working place can have a base temperature of 16°C. But a hospital needs for instance 21°C. Therefore baseline values are developed. A baseline in the UK is 15.5°C, in Belgium 16.5°C and in the Netherlands 18°C. Isn't it strange that heating isn't necessary if the building temperature is above 16.5°C? No, because there is also an average internal heat

gain and the baseline can differ from building to building. Internal heat gain refers to a few degrees of free heating. These degrees are derived by: people (sensible and latent heat), lights (sensible heat) and equipment (sensible and latent heat). 3,5°C is a basic value for the internal heat gain of buildings. 16.5°C + 3,5°C is 20°C of inside air temperature. This value is correct, if the baseline was followed. Therefore, it is important to understand, that you don't need to follow the presented baseline value. Because a lot of people are unaware of the fact that DD can come in any base temperature [22].

For the calculation of DD, one need to turn temperature readings into DD. Actually this action is very straightforward. A simple example shows the philosophy behind the calculation. Assume a building with a base temperature of 16.5°C and consider a single day, August 15th. The outside temperature was 15.5°C throughout the whole day. Temperature varies throughout a day, but a constant temperature makes the calculation easier. The calculation is:
1 degree (below base temperature) * 1 day = 1 HDD on August 15th

There won't be a HDD, if the outside temperature is above the base temperature. Like already mentioned, temperature will vary throughout a day. So the calculation will be a summation of different intervals. Where each interval has had a constant temperature. Nowadays, the shortest intervals which are measurable are these of an hour or half an hour. The result of such an interval summation is a DD.

The introduction mentioned that DD are used every week or month. A new DD will be presented each day. If the DDs are analysed each week, one must take the sum of each DD for that week. For month analysis the same method is usable. The result of the summation represents the number of DD for that week/month.

<u>The use of HDD:</u>
HDD are used as an indicator for energy efficiency. For example, the heating system used 252,10 kWh gas, in 2011. At the end of 2011, the roof was insulated to reduce the gas consumption. In 2012 the gas consumption was 237,90 kWh. So the difference between these two values isn't large. Does this mean that insulating the roof wasn't a good idea? No, it is possible that 2011 was colder than 2012. To proof this statement HDD are used. The average value for 2011 was 3,3 and for 2012 4,1. Now it is possible to compare the gas consumption with the HDD's. This calculation provides:

- kWh per HDD in 2011 = 252,10/3,3 =76,39
- kWh per HDD in 2012 = 237,9/4,1 =58.02

The conclusion of these 2 methods is that the heating efficiency in 2012 was about 30% better than it was in 2011. So insulating the roof was a good idea [22].

<u>CDD and GDD:</u>
CDD are the opposites of HDD, since they are a measure of how much and how long, the outside temperature was above a base value. GDD are calculated in the

same way as CDD. But the base temperatures used are based upon the temperatures above which certain plant or insect growth occurs. GDD are beyond the scope of this thesis.

<u>DD issues:</u>
When DD are applied to real-world buildings, common DD techniques suffer with a number of problems, which can easily lead to inaccurate results. These problems are important for DD users, therefore several issues are explained [23].

a) Base temperature problem

Like already mentioned in the text above, is the base temperature now 16.5 °C or 21°C? It is important to pick an appropriate base temperature for DD calculations. As a building's base temperature typically varies throughout the year, even the most appropriate base temperature is usually only an approximation.

b) The base load energy problem

DD are used for analyzing the heating (gas, fuel) consumption of buildings. But in many applications gas is also used for cooking or domestic hot water. The amount of gas that isn't used for heating, isn't related to the external temperature. That is the base load value which is on the y-axis. The problem with this value is that this value isn't measured, most of the time. The base load value then depends on the used base temperature. This could be a problem when the wrong temperature is chosen. Fig. 10 shows this problem.

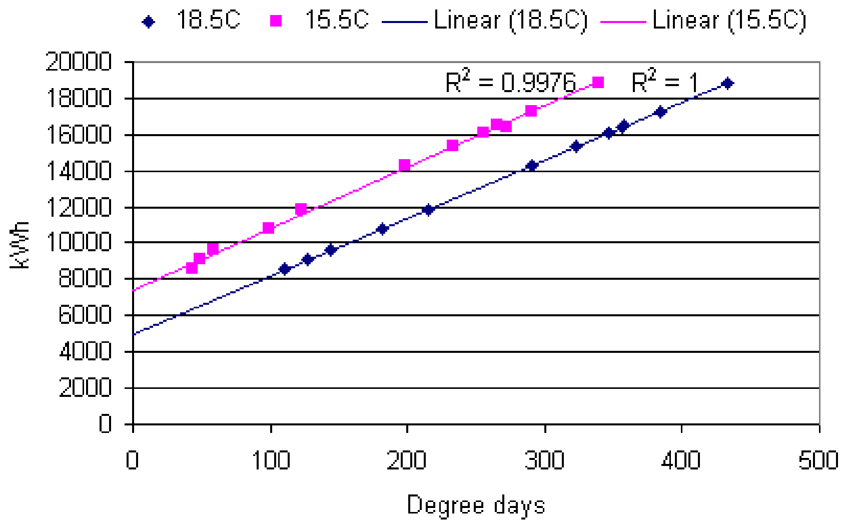


Fig. 10: base load problem

Source: energylens [23]

Two different base temperatures are used. If the determining method for the base load is used as mentioned, 2 different base loads are calculated. This shows how important a correct base temperature is.

c) The “ideal” temperature problem

This part will explain an inaccurate zone of DD. When the outside temperature is close to the base temperature of the building, the building will often require little or no heating. However, DD calculations are inaccurate under such circumstances. For instance: the gas consumption must be normalised for a specific day. The heating system used 122 kWh. The average outside air temperature was 16°C and the defined baseline is 16.5°C, so this results in 0.5 HDD. We need to divide the amount of gas by the HDD, for the normalization. The normalized gas consumption is now 244 kWh/HDD.

Hints to tackle the DD issues:
Suggestions to tackle the most common DD issues might be:

a) Use the most appropriate DD data you can

The DD data that is readily available, is unlikely to be entirely appropriate for calculations. Ideally, you would be able to get DD data for an appropriate base temperature and covering just the hours at which the building is heated. However, such data is unlikely to be freely available.

b) Ignore periods with an ideal outside temperature

One of the problems mentioned that DD are inaccurate, if the base temperature is close to the outside temperature. It is often best to simply leave these periods out of any analysis.

c) Get interval metering

This kind of metering will provide high resolution data, because it is possible to takes samples at half-hourly intervals. Analysis of this data can, in seconds, reveal patterns of energy wastage that could never be revealed by weekly or monthly regression analysis. On the other hand, interval metering solves the problem of manual meter readings, because this kind of metering works automatically. Therefore, data is always saved at the same time. This shall bring accuracy to a higher level.

d) Get interval sub metering

Separate sub metering can significantly reduce or even eliminate the base load energy problem. If the heating energy consumption is metered separately, any base load will be minimal. Use mean degree hours for DD calculations. The most mathematically precise method of calculating DD, is to sum hourly temperature differences and divide by 24. It is important that only positive differences are summed. The formula for this process for HDD is [23]:

$$HDD = \frac{\sum_{j=1}^{24}(\delta b - \delta o, t)_{((\delta b - \delta o, t) > 0)}}{24}$$

Where HDD is the daily DD for one day. δb is the base temperature and $\delta o,t$ is the outside temperature in hours j. The subscript denotes that only positive values are taken. Daily DD are then summed over the appropriate period.

3.3.2.5 *BMS (left top of Fig. 9)*

The BMS is the most important part in the development of energy algorithms. It will provide information as input for the algorithm, but it will also perform the control actions. This kind of algorithms are difficult to use when there isn't a BMS, because of the fact that the BMS establishes the link between the software part (algorithm) and the hardware part (installation). Not all of the installations are controlled and visualized at a central place without a BMS. This makes it complex to gather all the necessary data. Therefore, it is highly recommend to work with a BMS [24],[25].

Definition:
A BMS is used to control and to visualise all the installations of a building on a central place. These installations can be heating systems, doors, lighting, safety systems,… The main advantage of a BMS is that all the data is located at a central place [26].

Kind of data:
BMSs have a large range of possible applications. The smallest BMS are responsible for heating and lighting. When the largest BMSs are analysed, applications as door control , safety and sun blinds are integrated. Therefore, the extent of data that is available will depend on the BMS configuration. There are two things that a BMS always can provide. They are already visualized in Fig. 9, building information and HVAC information. These are providing general information which is used by almost every energy algorithm.

- Building information:
  - location of the building,
  - number of rooms, lights,
  - information of the building envelope.
- HVAC information:
  - time schedules of heating start-up,
  - measured temperatures (water, air, room),
  - set temperatures,
  - capacity of the HVAC devices,
  - kind of HVAC system.

Complex algorithms will need more information, and will use a BMS with a larger bandwidth. The larger bandwidth is required to gather the necessary data which means that the cost price of the installation will rise.

3.4 KPI

A KPI is a part of the verification element which is discussed in the action plan of EnMSs. Therefore KPIs are very important to the top management [3], [27]. Because they are useful in decision-making and can be used throughout all levels of a company. The purpose of a KPI is to measure success in achieving outcomes of a planned action. Furthermore, this part discusses different KPI categories, what is important when KPIs are used, how many of them are necessary and how a KPI is created. This part will start with a definition.

3.4.1 Definition

A KPI is a kind of measurement that enacts how good a process is performing in comparison to a reference. KPIs are used for important aspects of organisational performance. These kind of aspects are related to improvements or processes that must be kept between specified limits [3].

3.4.2 KPI categories

It is possible to divide the KPIs in a number of categories [28]. Some of them are:

- quantitative KPIs: related to the amount of an object;
- qualitative KPIs: says something about product or process quality;
- cost efficiency KPIs: the unit cost of achieving a specified amount of service;
- cost effectiveness KPIs: Cost of approaching a fixed amount of service to a designated level of quality;
- timeliness or Responsiveness KPIs: Related to the time taken to carry out a service;
- productivity KPIs: The amount of outcome of a machine or process.

3.4.3 Important knowledge about KPIs

There are several interesting things to know when using KPIs. Things that are forgotten very often but that are useful. They are all highlighted separately [28].

a) Strategies and actions will change over time.

It is important that related KPIs are changed together with the strategy or objective. Management should also explain why the KPI is changed or why it is still relevant.

b) Relationship between strategy and KPIs

There must be a relationship between the developed strategy and the defined KPIs. Otherwise it would be impossible to provide enough background knowledge to other interested people (managers, team leaders,..) of enough background knowledge.

Without background, they can't asses the strategies adopted by the company and the potential to succeed.

c) Focus on an open communication policy

When a company approaches an open communication policy, it is necessary that they explain the components of a measurement and how it is calculated.

d) Explain why a KPI is relevant for a specific target or objective.

e) Provide data sources used for calculations

Provide the data sources used for calculations, if people ask for it. Explain also the assumptions, if there were assumptions. Such an action will give managers and team leaders confidence in the top management. Besides they can make their own assessments of the reliability of the workflow.

f) KPIs as benchmark

Top management may use KPIs for benchmarks against a relevant external group. In such a case they need to explain why these groups were chosen. This explanation gives an idea about the company's competitors in the opinion of the top management.

3.4.4 Number of KPIs used

KPIs are used in several sectors (financial, IT, industry,..), so it is difficult to define an exact number of KPIs which are necessary. Everything depends on the company and their strategy. However, based on experience four to ten KPIs are acceptable for most types of companies [28].

3.4.5 Creating a KPI

There are several steps that must be followed when creating a KPI [28]. The sector for which a KPI is developed may have an effect on the number of steps. For instance, some KPIs are based on actions. Thinking off an emergency door which is used when there isn't an alarm. A KPI for such a problem may be: the door can be opened less than 3 times a year, when there wasn't an alarm. Otherwise the KPI isn't met. It will be more complex when we need to develop a KPI for an EnMS. To provide a general overview, several steps are shown below.

- KPIs definition
  - define the KPIs;
  - analyse the company's strategy ,
  - brainstorm on the KPIs,
  - think of a plan to gather data.
- Monitor & analyse data

    - monitor KPI compliance,
    - gather and analyse the data.
- Improve the processes or components
- Refine KPIs

Firstly, Top management must decide what kind of KPIs they want to implement. Therefore, it might be necessary that the strategy of the organisation is analysed in order to become a relevant KPI. In the future a KPI must be analysed to verify if there are been improvements or not. Therefore the organisation should think about a plan to gather data. Once this is done, it is possible to step-up all the measurements. Based on the compliance of the KPI and the gathered data, it will be possible to analyse if improvements are necessary/possible. If improvements are possible and the system is updated, the old KPI must be reviewed, because it is possible that the old KPI is not relevant anymore. Finally, these steps only provide a general overview. Therefore, it is possible that an organisation does not use all these steps.

4 Data collecting and pre-processing

The introduction appointed the target for this thesis, developing algorithms based on data. Before the development can start, it is necessary to have data which has been prepared for data mining. This chapter describes the process towards prepared data. Firstly, it is explained where the data used in this thesis come from. Secondly, it is clarified which items are essential in data pre-processing. At the end there is a conclusion which summarizes the most important items. Fig. 11 shows which steps within the data mining pyramid refer to the pre-processing of data [18].

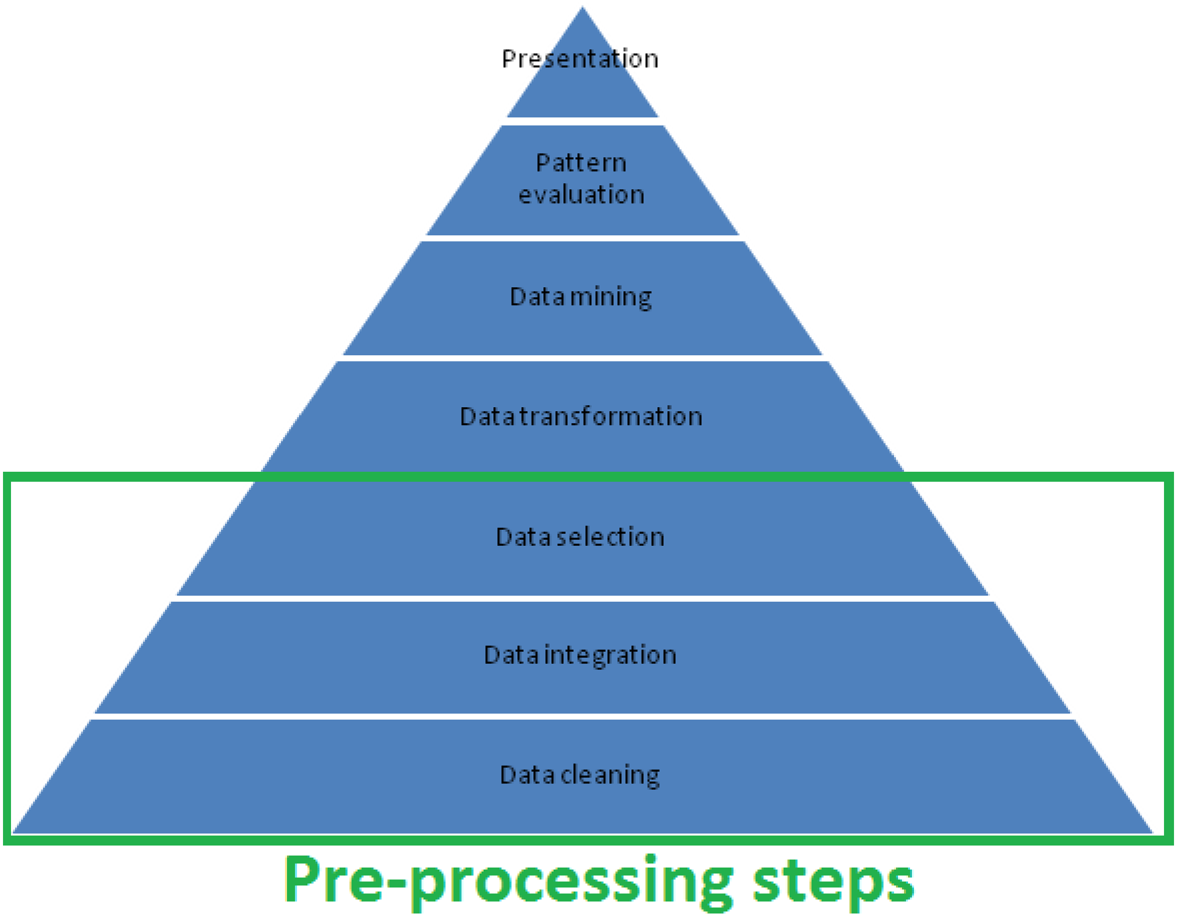


Fig. 11: pre-processing steps

4.1 Collecting methods

In general there are a lot of different methods to collect data. The most important difference is the one between manually and automatically data storage. When data is used for development it is definitely preferred to use automatic data storage. Because manually collected data isn't saved at a constant interval.
For instance: One is responsible to collect data each day at 2 p.m. The first day the meter reading is done at 1.59 p.m. and the second day it is done at 2.10 p.m. This example makes clear that the accuracy isn't good, while the accuracy is important for the later analysis.
Data provided by Porta Capena is normally based on automatic meter readings and pulses. How this data is extracted is explained in part 4.1.1 and 4.1.2. The last part

4.1.3 explains how data which isn't measured by Porta Capena can be used for analysis.

4.1.1 EcoSCADA

EcoSCADA is the web-based solution of Porta Capena for everything what is related to energy (analysis tools, reports, dashboards, etc.)[29]. The webpage of EcoSCADA provides all the data which is measured. Each customer has a login and password which gives him access to the buildings of his responsibility. The advantage of this approach is that there is no need for your internal IT network, no extra servers or databases, no extra support. Only a standard web browser is necessary.

The data of interest for the customers can be exported from an energy consumption profile. Fig. 12: consumption profile shows the electricity data of the reference year 2013 for BEMT.

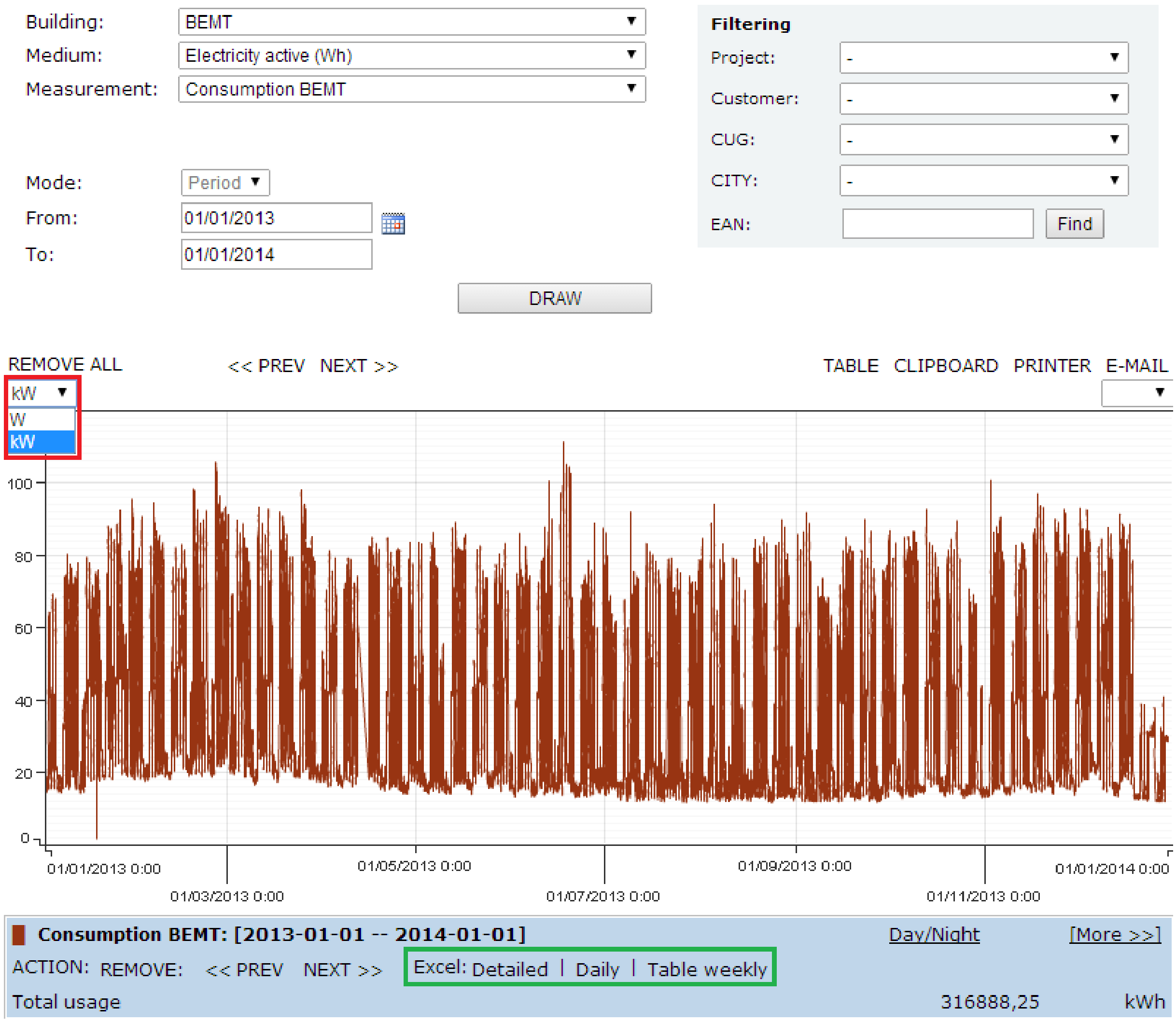


Fig. 12: consumption profile

Before the data is extracted one should decide what the unit of interest is. This will depend on what kind of device or installation is measured. For the development of a holiday algorithm the unit is kW, since the decision about the holidays which are out of control for a building is based on the total consumption. Once the unit is chosen it is possible to extract the data. The green rectangle shows the possible methods. It is recommended to use detailed (15 minute values) information to develop an algorithm. Since in the beginning it isn't known which data mining method is the best, and because it is possible to calculate all the other values (daily, weekly) from the 15 minute values.

4.1.2 Data collector

The data collector is a recent development to extract data from the database and to synchronise them with other data.
There are also 2 options to run this collector:

1. in a webpage,
2. as an application.

Fig. 13 shows the lay-out of the data collector when the application is used.

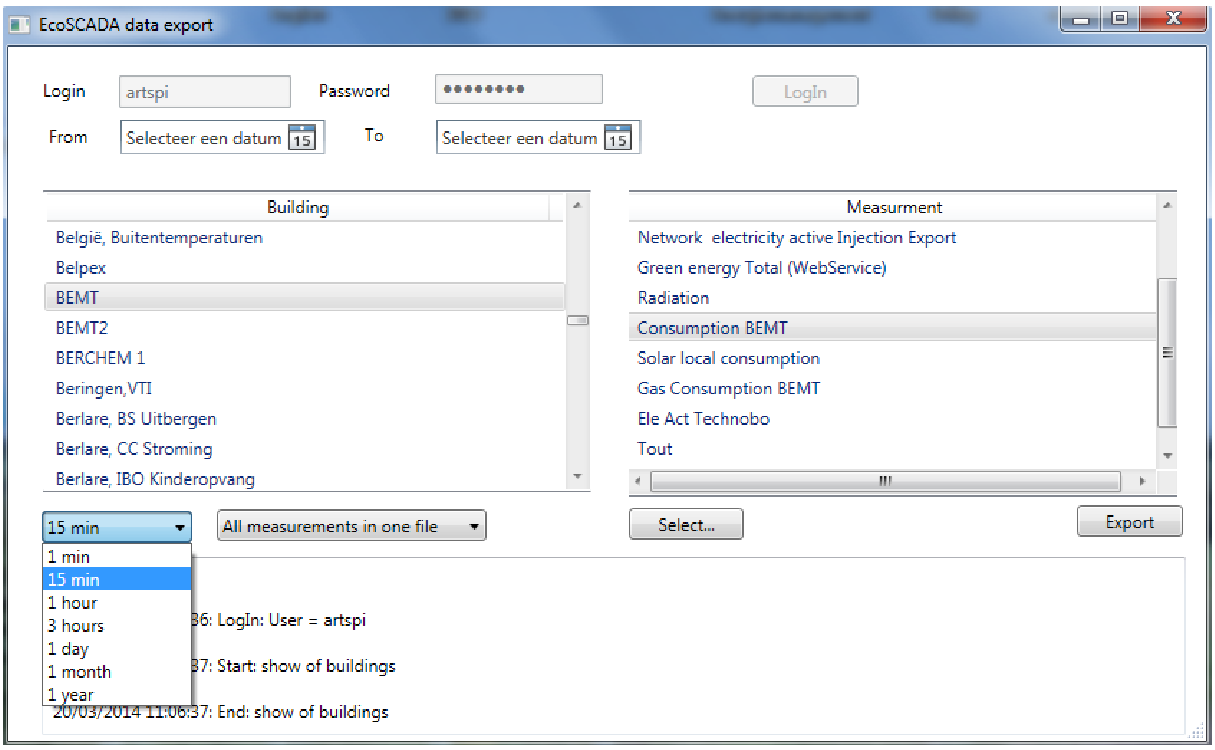


Fig. 13: lay-out data collector

The main difference between the collector and the consumption profiles is the presentation of the data. As Fig. 13 shows, it is possible to extract data from 1 minute values until yearly values. Also other options can be selected. Furthermore, one must select the path where the data is saved, this wasn't possible in EcoSCADA. The other part of handlings which are necessary to extract data are similar to the previous method. So the collector has more options than EcoSCADA. Besides the extra options the collector is built to provide data which is prepared for analysis. Therefore, the data cleaning issues explained in 4.2 are normally tackled by the algorithms of the data collector.

4.1.3 Collecting third-party data

The database doesn't have data of new customers. Therefore, if a new customer is interested in an energy data scan, he should provide a relevant dataset. Once the building is configured in the software and the dataset is imported into the database, one may run an energy data scan. This approach is only possible if the customer already stores his energy data. However there are a lot of buildings that aren't equipped with such devices. In such a case it is possible to contact the distribution network operator, for Flanders (Belgium) this is Eandis or Infrax. They can provide data of historical consumption for the building. This is possible for households and professionals, in this thesis only professional data is relevant. How the data is presented depends on the type of measurement[30]. For professional users 3 types are possible, whereby the selected type depends on the consumption of the building. These 3 types are:

1. manual meter reading on yearly base;
2. manual meter reading on monthly base;
3. automatic meter reading.

The introduction of this part already discussed the difference between manual and automatic meter reading, manual meter reading isn't appropriate to develop an algorithm. Although monthly readings may provide enough information for a basic analysis. For a detailed analysis and to develop algorithms more detailed readings are recommended. Therefore the automatic meter reading of the distribution network operator is preferred, if the data isn't measured and the customer can't provide a dataset.

4.2 Data pre-processing

The study of the basic principles showed that data mining starts with data pre-processing. Normally they consist of 4 different parameters. But they aren't all explained in this part, because data transformation (the 4th) is linked to the used data mining technique. Since the development of the algorithms isn't started, there is also no data mining technique chosen. So this item will be discussed in the next chapter.

4.2.1 Data cleaning

4.2.1.1 *Time zones*

Each country has a time zone which has a uniform standard time. Nowadays there are 40 time zones in use which means that each date is used for 50 hours somewhere on earth. The time in a specific zone is shown as the difference against the UTC time [31]. For instance, Belgium is situated in the time zone UTC+ 1. If a database is working with timestamps in UTC, it would mean that timestamps from meter readings are recalculated to UTC time. This means that data exported from that database isn't correct, because it uses UTC time and UTC +1 is expected. Such a problem may be detected by comparing a timestamp against the same timestamp in a consumption profile (Fig. 12). This is possible, because the building must be configured to visualise a consumption profile. Since a building is situated in a country and a country in a time zone, the time in the consumption profile is adjusted to the countries time zone. Fig. 14: time zone issue shows an example of this issue.

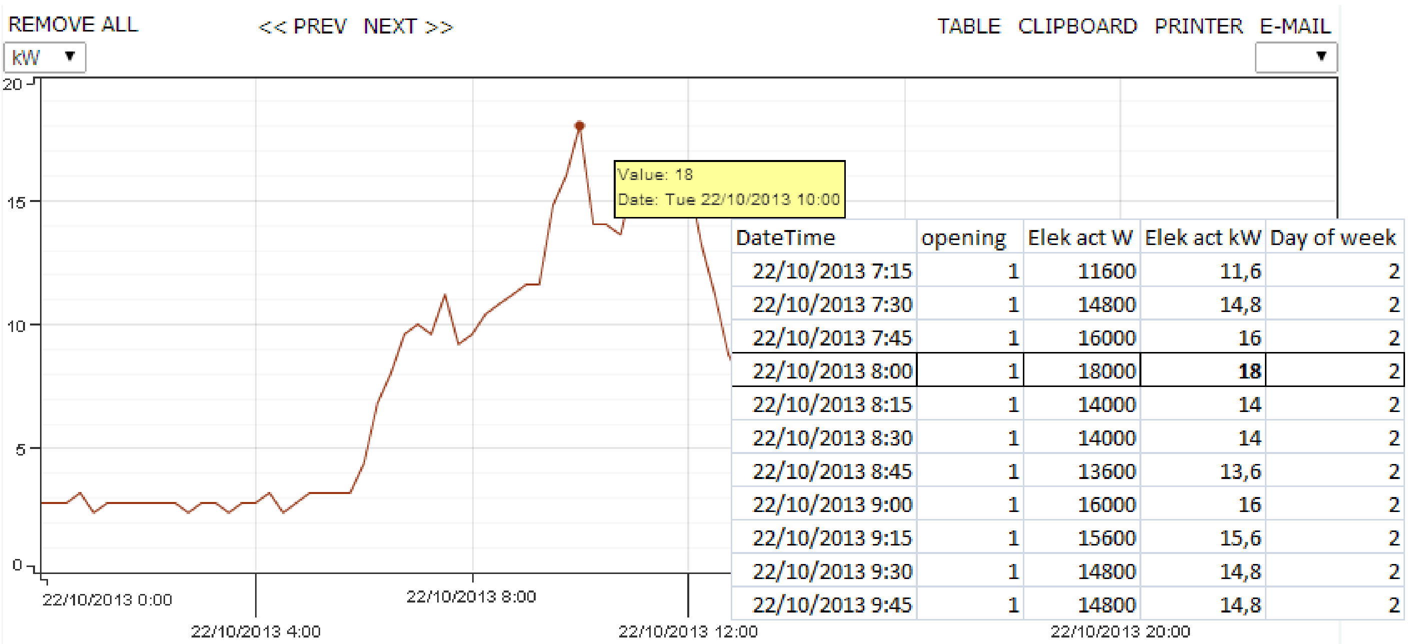


Fig. 14: time zone issue

4.2.1.2 *Data integrity*

Energy data is measured by a data logger which is installed in the field. So there are a lot of factors which may affect the working of the logger:

- problem with the power supply,
- broken cable between measurement device and data logger;
- problem with the data logger,
- problem with the measurement device.

If there is a problem gaps may occur in the measured data which result in a decreased data integrity. When there are gaps in a dataset and the set isn't prepared mistakes will be inevitable. Table 1 shows a dataset which isn't prepared.

Table 1: unprepared gaps

Timestamp	Value in kW
12/04/2013 15:15	67,9
12/04/2013 15:30	64
12/04/2013 15:45	61,4
16/04/2013 9:15	18,1
16/04/2013 9:30	82
16/04/2013 9:45	80,2
16/04/2013 10:00	79,6
16/04/2013 10:15	82,5

The yellow records visualise where the gap is situated. Data have been logged until a quarter for four on 12/04/2013. The next timestamp isn't measured so a problem occurred between four and a quarter for four that day, whereby the problem was solved on 16/04/2013. This data isn't prepared because nothing happened with the gaps. Furthermore, it is very difficult to detect these kind of gaps because they aren't visualised in the dataset.

Therefore gaps should at least be visualised in the dataset and it should be better if they are filled in with values. In data mining there are different methods to handle this kind of missing values. Firstly, in some datasets it is possible to use the most probable value to fill in the missing value. For instance: a dataset with temperature measurements has a gap of two hours. For such a small gap it is a good idea to use linear regression to predict the missing values. This method is only possible for linear processes, as temperature is. Furthermore, the gap must be limited because linear regression for 2 days isn't reliable.
Another method would be decision tree induction. This may be used for electricity and gas predictions if there is enough data, because a decision tree takes decisions on the provided data.
Secondly, gaps may be filled by using a measure of central tendency(e.g., mean or median). This method is less appropriate for gaps in temperature measurements, because it isn't accepted that a gap is filled with the median value of 16.5°C while the temperature of the latest measurement was 12°C. Such an optimisation would mean that the temperature changes with more than 4°C in 15 minutes, which is not relevant. On the other hand, it can be used for electricity or gas, since calculating median/mean values is less complicated than using a decision tree.
Finally, it is possible to fill gaps with a global constant (unknown, ∞, blank item). This method visualises the timestamp but doesn't decide what the missing value should be. This is useful when data is delivered to an end-user. The user will detect immediately a gap or multiple gaps and he can fill these gaps with the method of his choice.

4.2.1.3 *Time shifts: summer- and wintertime*

Europe uses a guideline which is provided by the European Union. This guideline explains the policy which is used by the member states, for summer- and wintertime. This policy says that: summertime starts on the last Sunday of March. On this Sunday 1 o'clock UTC time becomes 2 o' clock, for Belgium (UTC +1) 2 o'clock then becomes 3 o' clock. Summertime ends on the last Sunday of October. On that Sunday time goes back to normal UTC time[32].

For people this is not a problem the time changes in the night so in the end everyone has slept a hour extra or less. But it is more complicated for databases. They are working with UTC time and the data is shifted in time if another time zone is needed. This results in 4*24-4=92 values for a summertime change and 100, 4*24 +4=100 values for a wintertime change. These issues are solved with prepared data. Summertime will receive 4 blank items which results in 96 values. Wintertime has too much values so the data must be reduced to 96 values. A reasonable method is to delete the values which are measured in the 'old time'. But there are probably more methods which are useful in this case, but that is behind the scope of this thesis.

4.2.2 Data selection

In data mining literature this is the third pre-processing step (see Fig. 11) [18]. This is only correct if the cleaning and integration is done on the database and not on a smaller dataset which is used for further analysis. In this case cleaning is done in the database but the integration isn't. So this is a different situation which means that the data selection isn't the third item, but should be the second one, because data selection is the step where relevant data is retrieved from the database.

4.2.3 Data integration

Data which is extracted from a database doesn't always have all the required information. The extra information which would be relevant in the data mining step can be added to the original set. This process is called data integration, within the literature of databases this is known as an inner join. This data may be:

1. from another dataset,
2. derived from the original step.

If data from another set is used one needs to be sure that the data of the new set refers to the same attribute of the original step. This problem is called the entity identification problem. Within the energy business it isn't a problem to tackle this issue, because each measurement or value is linked to a timestamp. Therefore the attributes are linked to each other by the timestamps. In case of derived information there isn't a problem at all, since one always knows from which measurement either value the information is extracted. Table 2 shows a dataset without data integration while Table 3 shows one with data integration.

Table 2: energy data without integration

Day	Elect [kWh]	Data Integrity [%]	Elect Norm by DI
01/Jan/13	562,80	100	562,80
02/Jan/13	933,20	100	933,20
03/Jan/13	996,23	100	996,22
04/Jan/13	963,48	100	963,47
05/Jan/13	399,78	100	399,78

Table 3: energy data with integration

Day	Elect [kWh]	Data Integrity [%]	Elect Norm by DI	Event	Holidays	Weekday/Weekend
01/Jan/13	562,80	100	562,80	0	1	0
02/Jan/13	933,20	100	933,20	0	1	0
03/Jan/13	996,23	100	996,22	0	0	0

04/Jan/13	963,48	100	963,47	0	0	0
05/Jan/13	399,78	100	399,78	0	0	1

These tables are an example of data integration. It doesn't mean that only these extra parameters are relevant. In some cases opening hours, temperature or others are also interesting, this example shows 3 new columns. Firstly, the event column is displayed. This information is extracted from a dataset of the building owner. These values indicate if there is a building event in the weekend either on a holiday or not. Secondly, the information about holidays may be extracted from 2 datasets. One set which provides information about the legal holidays and another one which gives information of the building holidays. For instance: Thursday 9th of May 2013 was a legal holiday in Belgium while Friday 10th of May 2013 was a building holiday. So everyone had a vacation of 2 days. It is also possible that a building owner brings this data together in one dataset. Lastly, the weekends are displayed in the last column. In contrast to the 2 other columns the data in this column is derived from the data in the left column. So these 2 tables give a good overview of the possibilities with data integration.

4.3 Conclusion

Before data is retrieved from the database it should be cleaned. Therefore the data collector is developed, because this collector has the algorithms to clean and prepare the data for further steps. Furthermore, the part of data cleaning is an advice for everyone who needs to work with data. Besides the data collector there is EcoSCADA which is more recommended for a quick overview.
After the data is collected (data selection) it is possible to add data which is based on a new dataset or which is derived from available information. Any kind of data may be added as long as this data is relevant for analysis. Based on this information it is possible to draw a new data mining pyramid, Fig. 15 shows this new pyramid.

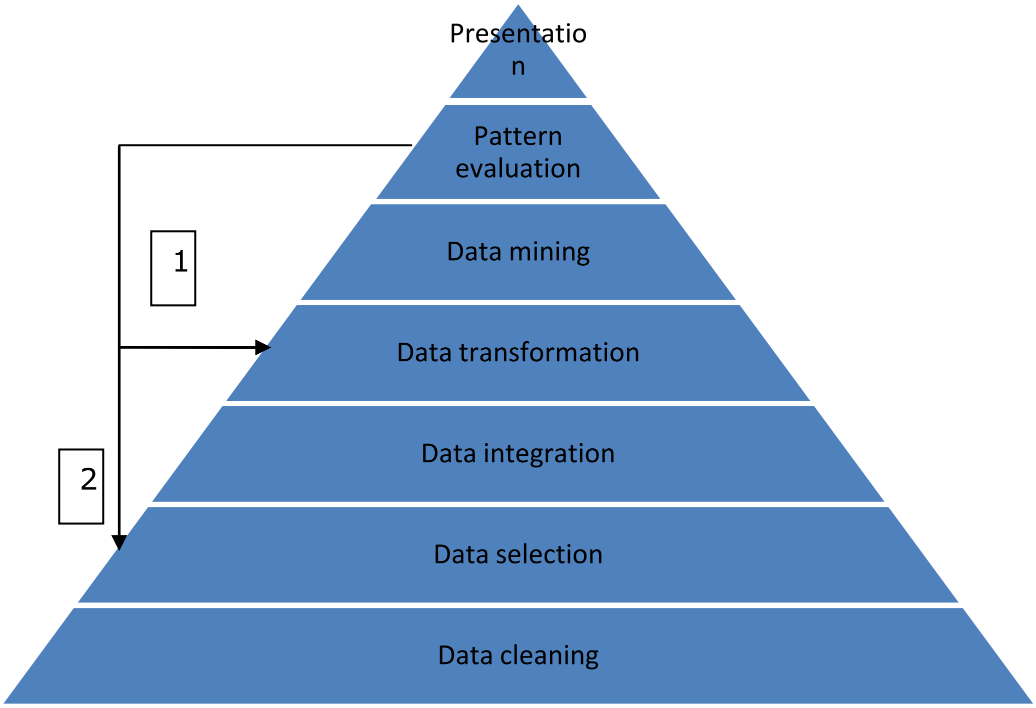


Fig. 15: data mining according to thesis

Firstly, the data selection step is interchanged with the data integration step, like discussed in 4.2.2. Furthermore, if the evaluated pattern doesn't provide the expected result it is recommended to start from the data transformation step (number 1). When another data mining technique doesn't provide better results, one should restart from the data selection step (number 2).

5 Data mining techniques

This chapter starts with an analysis whether data transformation is applicable for the holiday anomaly algorithm either the real-time building occupancy algorithm. The data transformation part is followed by an overview of several data mining techniques which can be used for the defined algorithms. Finally, the conclusion explains which technique is the best for anomaly algorithms.

5.1 Data transformation

This data mining step will transform or consolidate the data from the extracted dataset. In this thesis 15 min values are extracted, these values are used when a transformation is preformed. Data transformation is used to simplify the data mining results which ends up in an easier pattern understanding. Since easier results are appreciated in decision making, it is reasonable to check if data transformation is applicable on the used dataset.

The applicability of data transformation depends on the expected result. So the demands of the algorithms must be evaluated. The main demand of the holiday anomaly algorithm is detecting the holidays which are out of control. Once these days are known it is a matter of calculation to define an EnPI and KPI. Furthermore, the main demands of the real-time building occupancy algorithm are:

- detect occupancy,
- decide to switch off unnecessary systems or not.

Data transformation is applicable in case of the holiday anomaly algorithm, because anomalies are detected on a daily base. Therefore, it is more designated to use daily consumption values instead of 15 minute values. Such a transformation results in a huge reduction of the data. Each day has 96 values when the sample rate is 15 minutes. This results in 35040 values for one year. The data transformation from 15 minute values towards daily values results in a reduction of 34675 values.

It is more complicated for the real-time building occupancy algorithm, because the algorithm is bipartite. Firstly, it must be detected if the building is occupied or not. This analysis isn’t based on gas either electricity consumption, since these values are the parameters which the algorithm will control. Once it is known that the building isn’t occupied, the algorithm must check the gas either electricity consumption. If the consumption is higher than on a normal holiday, a shutdown of the responsible system is required. Furthermore, the time intervals of the analysis are limited because it must be in real-time. Therefore, it is recommended to use 15 minute values for this algorithm.

5.2 Data mining techniques

This part gives a general overview of different data mining techniques which are appropriate for outlier detection in energy data. Such an overview gives information about the fundamentals and in which case it may be used. The study of the basic principles already introduced the concept of 5.2.1 statistical analysis and 5.2.2 analysis based on machine learning.

5.2.1 Statistics

Statistics is represented by mathematical functions which have the force to describe the behaviour of the analysed group/class [18]. These kind of functions are widely used so people recognise these functions. The most interesting methods for this thesis are methods which analyse the dispersion of the data. Because these are useful for identifying outliers. The two most common methods are described, the standard deviation & variance analysis and box plot analysis.

5.2.1.1 *Standard deviation & variance*

Standard deviation and variance indicate how widespread a data distribution is. The standard deviation is calculated as the square root of the variance. A low standard deviation indicates that most of the measurements are close to the mean. A high standard deviation means that the measurements are spread out over a large range of values [18].

There are two methods used to calculate the standard deviation. The selected method depends on what kind of observations are used. If a sample of data from a whole population is used, the method is called sample standard deviation. In such a case the variance is calculated by:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

With:

- s^2 = the variance
- N = number of observations
- x_i = one specific observation
- $\bar{x}$ = the mean of all observations

The other method is called the population standard deviation. This method is used when the whole population is used to calculate the standard deviation. The variance is calculated by:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

With:

- σ^2= the variance
- N = number of observations

- x_i = one specific observation
- μ = the mean of all observations

The used data is normally a sample of the whole population, since top management selects a data period which is relevant for the organisation. Therefore, it is recommended to use the first method.

These formulas may be used on each kind of data, although it isn't always the best solution. The best way to use these formulas is with a dataset which has a standard distribution, then it is possible to calculate outliers correctly. It must be tested if the dataset is a standard distribution or not [33][34]. This test could be done with a chi-square test or with a quantile-quantile plot. Although, a first test could be done by analysing the figure. Fig. 16 shows a standard distribution which is expected for an outlier analysis.

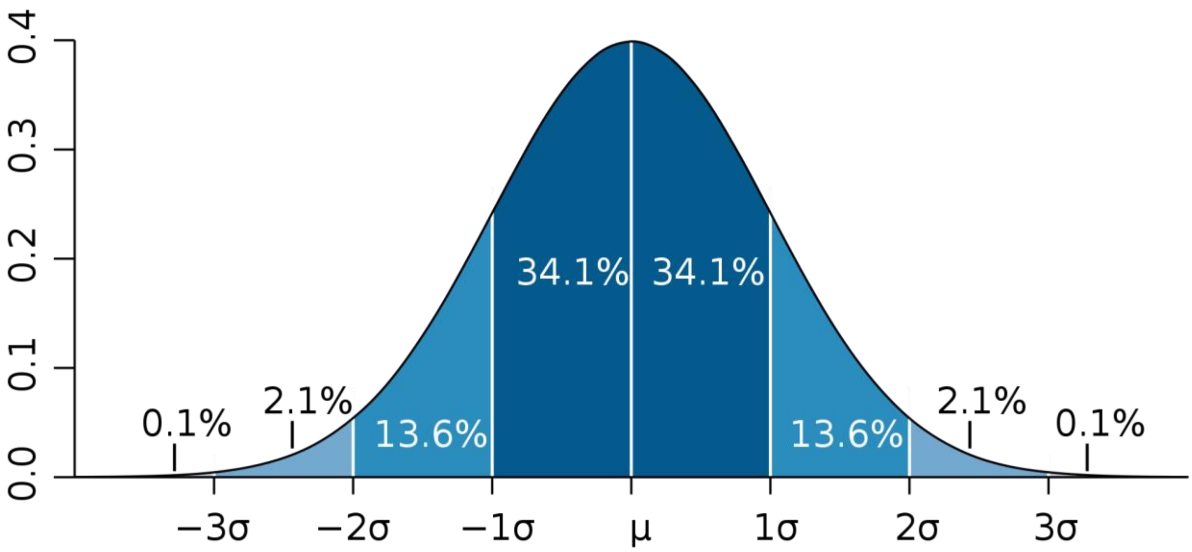


Fig. 16: standard distribution
Source: http://en.wikipedia.org/wiki/Normal_distribution [35]

A dataset which is distributed like this has 99.7% of its measurements between the -3σ & 3σ limits. Furthermore these limits are used for the outlier calculation. A value is identified as an outlier if it is further than 3σ away from the mean μ [4]. For instance: $\mu = 10$ & $\sigma = 2$, a measurement is identified as an outlier if its value is lower than $10 - 3 * 2 = 4$ or its value is higher than $10 + 3 * 2 = 16$.

5.2.1.2 *Box plot*

A box plot is a popular method for visualising a distribution. This method is not as detailed as a standard distribution or a histogram but a box plot is especially useful for the analysis of outliers [36]. Furthermore, a box plot is not restricted to any kind of distribution which truly is an advantage. Fig. 17 shows an example of a box plot.

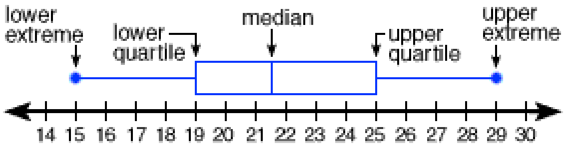


Fig. 17: box plot example

Source: http://www.statcan.gc.ca/edu/power-pouvoir/ch12/5214889-eng.htm [36]

The five titles of the figure are also known as the five-number summary. This summary is explained as follows:
The boundaries of the box are the quartiles respectively Q1 (lower quartile) and Q3 (upper quartile). The median is always marked with a line inside the box. The length of the box is calculated as the difference between Q1 and Q3 and is called the interquartile range IQR. The whiskers (two lines) connect the box with the lower extreme and the upper extreme either the minimum and maximum observations.

Although a box plot is commonly used as a visualisation tool, it is possible to calculate all the values of the five-number summary. Most software packages like excel are using algorithms to do this. The principle of these algorithms are all the same but the output values may differ. The principle of calculating a box plot is as follows: the maximum and minimum value doesn't need extra information. The median is calculated by taking the middle value of the dataset. If the number of observations is even, the median is calculated as the average of the two middle values. Q1 is the middle value between the smallest number and the median while Q3 is the middle value between the median and the highest number. Finally, outliers are defined with respect to the quartiles. An outlier is defined as an observation which is 1.5 x IQR beyond the quartile (Q3 or Q1). This results in two outlier boundaries:

1. minimum boundary = Q1 - 1.5 x IQR
2. maximum boundary = Q3 + 1.5 x IQR

Fig. 18 shows an example of a box plot with an overview of the five-number summary.

Q1	359,05
min	317,825
med	380,2875
max	834,775
Q3	436,2938
IQR	77,24375
outlier	552,1594

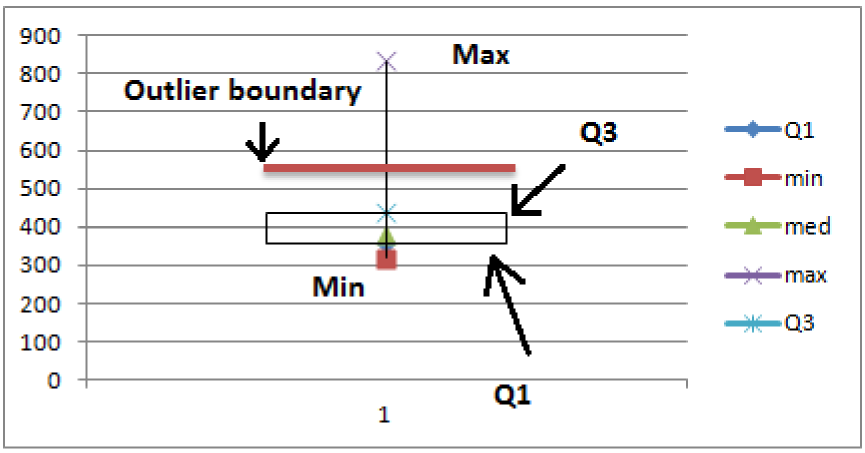


Fig. 18: box plot with five-number summary

The box plot contains all the parameters which have been discussed before (minimum, maximum, Q1, Q3 and median). The redline in between of the maximum value and Q3 is the outlier boundary. Each value which is higher than the red line or 552.1594 kWh is defined as an outlier.

5.2.2 Machine learning

The principle of machine learning, making intelligent decisions based on data, is already explained in the study of the basic principles. The first data mining technique which will be discussed is a special type of machine learning, unsupervised learning. This is unsupervised since the input variables are not class labelled, which means that it is not possible to put the data into standard groups. Therefore, this method is a synonym for clustering [18]. For instance, the input may be a dataset which contains the yearly electricity consumption. Suppose that 4 clusters are found where each cluster stands for:

- consumption during weekends;
- during the week,
- during events,
- during holidays.

The model provides the semantic reason of the cluster, since the data is not class labelled. Therefore unsupervised means that the model cannot check the decisions against training data. Therefore, feedback from another input parameter is always needed.
Secondly, Rule-Based classification is explained. Within this principle of machine learning the learned method is represented by a set of rules.

5.2.2.1 *k-means clustering*

k-means clustering was one of the first used clustering techniques and is still one of the most popular techniques. The reason therefore is that the technique uses a simple and straightforward principle to classify a dataset into a certain number of clusters (k clusters). Since k-means clustering is a centroid-based technique, the algorithm starts with selecting k-initial centroids. The number of k-centroids is defined by the user of the algorithm [37], [38]. Therefore, it is important that the user knows his data. Fig. 19 shows a plot of a dataset which will be used for clustering.

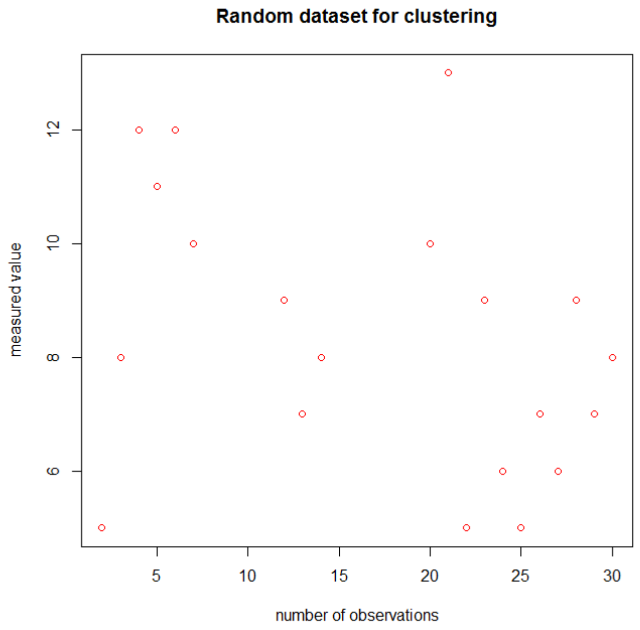


Fig. 19: dataset for clustering

Now, the user must define the number of expected clusters. As normal, each different number results in a different output. Fig. 20, Fig. 21 and Fig. 22 show a cluster with each time a different number of k-centroids.

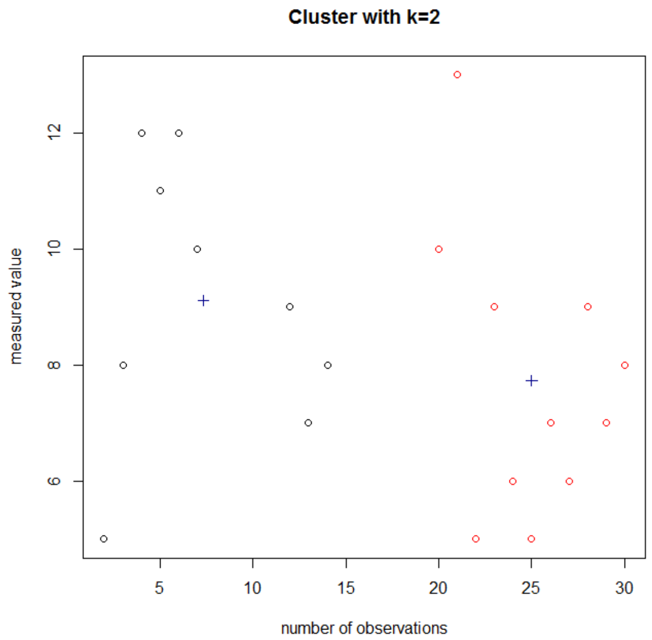


Fig. 20: cluster with k= 2

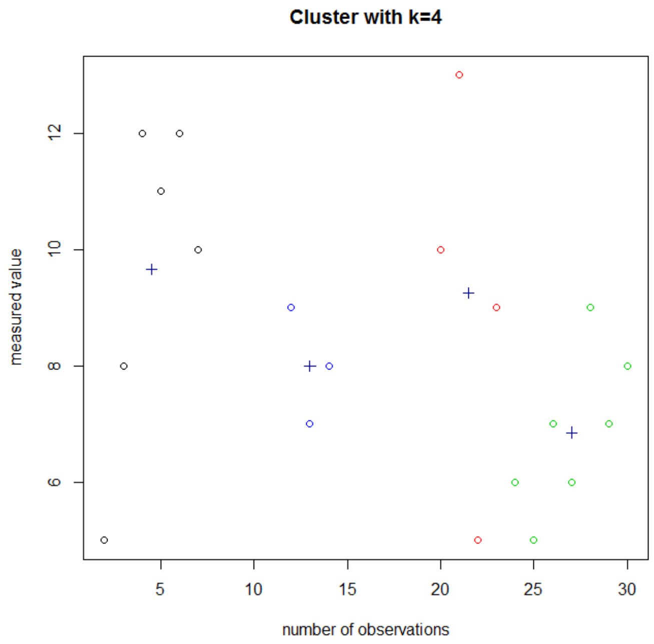


Fig. 21: cluster with k= 4

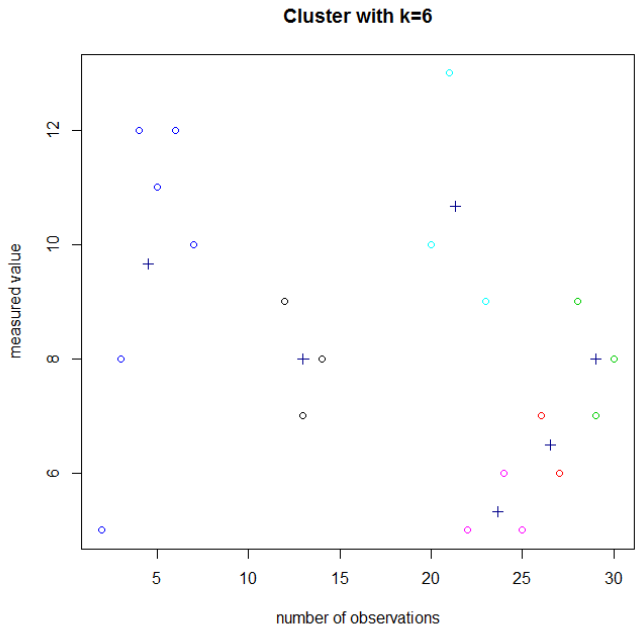


Fig. 22: cluster with k= 6

In these figures each different cluster has another colour, while the centroids represent the centre of a specific cluster. These are visualised with crosses.

Once the user has selected a number of clusters the algorithm can start its analysis. First, it selects k of the observations in the dataset D. If for instance the selected number of clusters is 2 then k has the value 2. The selection of the observations to start with is done randomly. Each of the selected observations represent a cluster mean, since it is impossible to select all the observations of D as centroid, there are still other observations that aren't used. These observations are assigned to the clusters which are most similar. This process is based on the

Euclidean[3] distance between the observation and the centre of the cluster. Furthermore, the quality control of a whole cluster is based on the sum of squared error between the observation and the centroid. This calculation is given by the formula:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, C_i)^2$$

With:
E = the sum of squared error
k = the number of clusters
p = a specific observation
C_i= the centroid of a cluster

In other words, for each observation in each cluster, the distance from the observation to its cluster is squared and finally all the distances are summed [18].Therefore, this formula tries to make the resulting clusters as compact and separate as possible. Fig. 23 shows an initial dataset for cluster analysis.

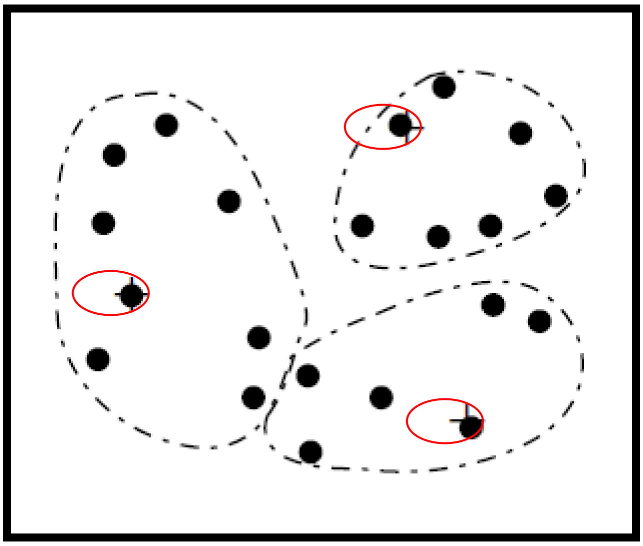


Fig. 23: initial dataset [18]

The defined number of clusters is 3. Based on the selected observations as centres (observations with a cross) the clusters are calculated. The k-means algorithm then iteratively improves the quality of each cluster. Based on the observations assigned to the cluster in the previous iteration the new means are calculated. All the observations are then reassigned using the new means as the cluster centres. Fig. 24 provides a schedule which makes this iteration process more clear.

[3] The Euclidean distance is the normal distance between two points that someone would measure with a ruler. [39]

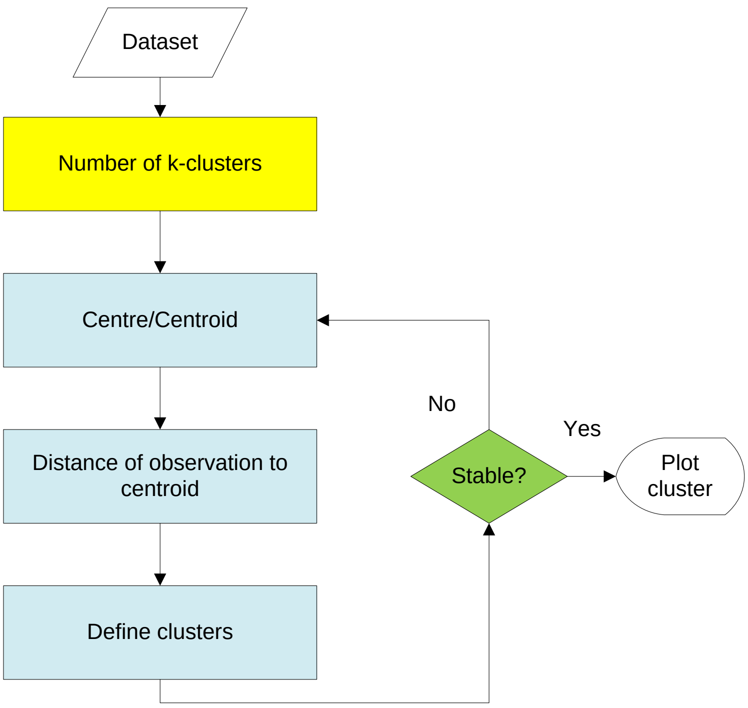


Fig. 24: iteration process

Furthermore, Fig. 25 gives an example of the reassignment of the centres and the observations within the clusters.

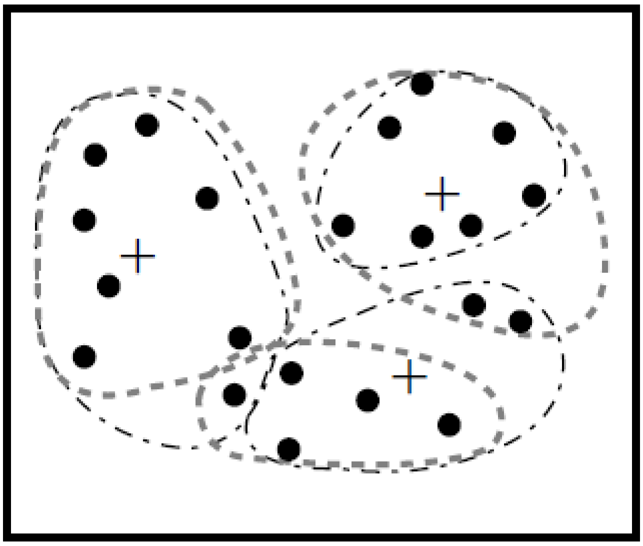


Fig. 25: cluster iteration [18]

The dotted curves represent the initial clusters from Fig. 23. During the iteration process new clusters are formed, these are visualised by the dashed curves. Once the iteration becomes stable the output is performed like Fig. 24 shows. Fig. 26 shows this final cluster.

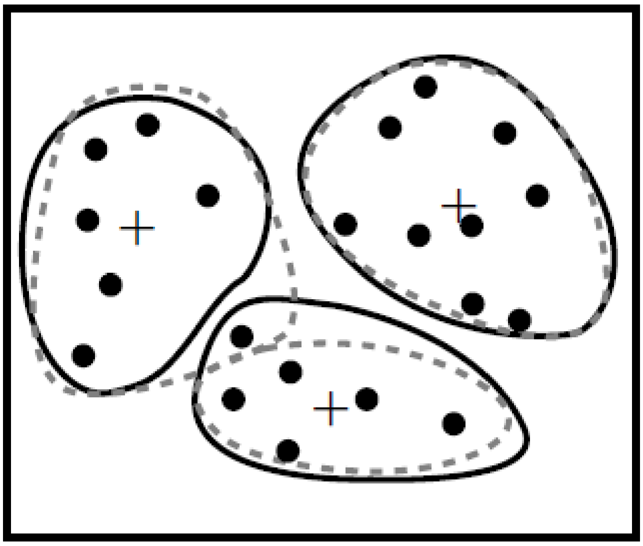


Fig. 26: final cluster [18]

The dashed curves in this figure are the last curves from the iteration process. The final cluster is represented by the bold curves.

5.2.2.2 *Rule-Based classification*

Rules are an efficient way of representing information about the requirements and the impact of decisions. Such a rule-based classifier uses IF-THEN rules for its decisions. This rule is used as follows: IF *condition* THEN *conclusion,* where the IF and THEN part have their own meaning. The 'IF' part is known as the precondition and the 'THEN' part is known as the rule consequent. The rule consequent takes place when the precondition of the rule is fulfilled.

A problem may occur when several rules are triggered at the same moment. In such a case it is required to have a conflict strategy. Two of these strategies are commonly used:

- size ordering,
- rule ordering.

Both methods classify their rules based on priority. Size ordering sorts the rules by its precondition size. The longer a precondition is, the tougher the requirement is. Therefore, the highest priority goes to the toughest precondition. Rule ordering sorts the rules beforehand. In this case there are 2 options. Firstly, the rules may be sorted based on decreasing importance. For instance: all the rules of the most frequently used class come first, the rules of the next prevalent class come next, and so on. Secondly, it is likely that rules are sorted into a priority list. This decision list is based on another input parameter which is seen as relevant. For instance: accuracy, size, etc.

It is important to build a rule-based classifier in a good way. Only then the rules are acting like expected. There are 2 ways to build rules:

- extraction from a decision tree;

- sequential covering algorithm.

A decision tree is commonly used for classification. Once the decision tree has generated its results (the conditions of each split) it is possible to translate this tree into rules. One rule is created for each path from the root to a leaf node. Each split stands for a precondition while the leaf node forms the rule consequent. Fig. 27 shows a decision tree which predicts the electricity consumption in case of a weekday, a weekend without event or a weekend with an event.

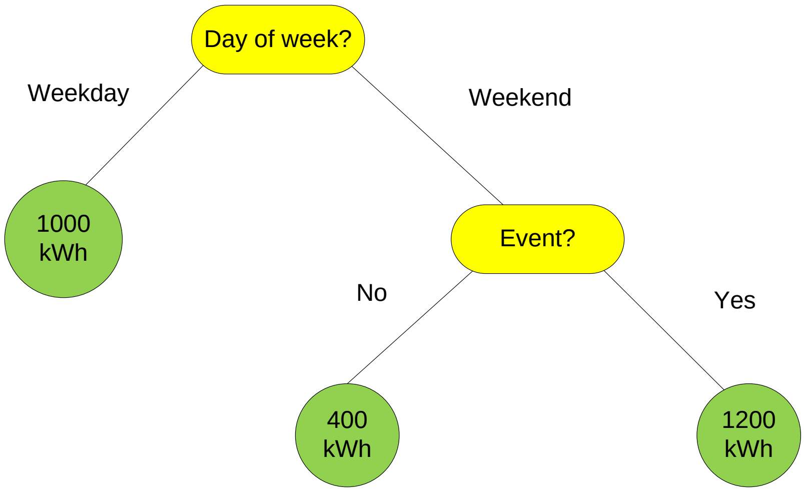


Fig. 27: decision tree

The conversion of this tree towards rules goes as follows:

R1: IF Day of week = Weekday THEN electricity consumption = 1000 kWh
R2: IF Day of week = Weekend AND Event = No THEN electricity consumption = 400 kWh
R3: IF day of week = Weekend AND Event = Yes THEN electricity consumption = 1200 kWh.
So for this tree it was very clear what the results were, but for a large tree this sometimes becomes difficult. Then it will be useful to convert the decision tree into rule-based classifiers.

A sequential covering algorithm extract rules directly from the data. Within this method rules are learned one at a time. So they are grown in a general-to-specific manner. Which means that it is possible to start with an empty rule and then progressively keep adding things to the next rule.

5.3 Conclusion

The standard deviation and variance are extremely useful when used with a standard/normal distribution. This advantage forms also a disadvantage, because it must be tested if the dataset is a standard distribution. Furthermore, this method is not a good measure if the data is highly skewed. Therefore, it provides fast and reliable results if the dataset is a standard distribution. The box plot doesn’t have this kind of issues, since it doesn’t put constraints to the type of distribution. Therefore, the box plot is easier in use which results in faster results, while the statistical methods don’t always detect outliers, the k-means clustering does. Since the algorithm starts with the expected number of clusters, it always generates k-number clusters. This may be seen as a disadvantage of this technique. Finally, the rule-based classification technique may be used in combination with one of the previous methods or in case of easy decisions. Like IF value > previous value THEN lights on. Rule-based classification in combination with one of the previous methods is used in the same way as with a decision tree.

6 Holiday anomaly algorithm according to ISO 50001

This chapter explains the development of the holiday anomaly algorithm. The first section 6.1 describes which technique(s) is/are the best for the algorithm, this section is based on the previous chapter. Once the best technique is chosen, it is possible to start with the development. Section 6.2 describes the internal structure of the algorithm. The electricity part describes the general concept. While the gas part illustrates the normalization of the data and explains how this extra step is implemented into the general concept. Section 6.3 explains how KPIs and EnPIs are extracted from the algorithm results and how the savings are calculated. Furthermore, section 6.4 gives a general overview of the algorithm, this overview is used to explain how the algorithm is used in practice. At the end of this section the final algorithm report is explained. Finally, section 6.5 describes the link between ISO 50001 and the holiday anomaly algorithm.

6.1 Data mining technique for holiday anomalies

This algorithm uses historical data to find holiday outliers. This means that a whole dataset (a relevant period for the organisation) must be available to run the algorithm. Therefore, the algorithm can use one of the following (previous introduced) techniques:

- standard deviation and variance technique;
- box plot,
- k-means clustering.

Like the conclusion of the previous chapter discussed, the rule-based classification technique is not appropriate for detecting outliers. It may be used to supervise the controller which communicates with the BMS.

The best data mining technique will be chosen based on several prerequisites and a comparison of the advantages/disadvantages of each technique. The next prerequisites are important in this selection:

- the users must have access to the outlier value;
- a building may have no outliers.

The first requirement gives the user extra information about the anomalies. This information can be used in benchmarks. Therefore, it is relevant that the outlier value is provided. The second requirement puts a constraint on the k-means clustering technique, since this techniques starts with the expected k-number of clusters. The number of expected clusters will be 2. The first cluster stands for all days which are in control and the second cluster stands for all days out of control or vice versa. But this technique will always find 2 clusters. Which means that the cluster will find outliers, even if there are no real outliers. Therefore this method is not used in the algorithm.

These prerequisites are not a problem for the statistical techniques. The standard deviation and variance technique have the advantage that the calculation method is clear and straightforward. But it needs to be checked if the dataset has a standard distribution. Fig. 28 shows a density plot of the daily electricity consumption. The dataset used for this test is a set from BEMT, the building which is used as a practical case.

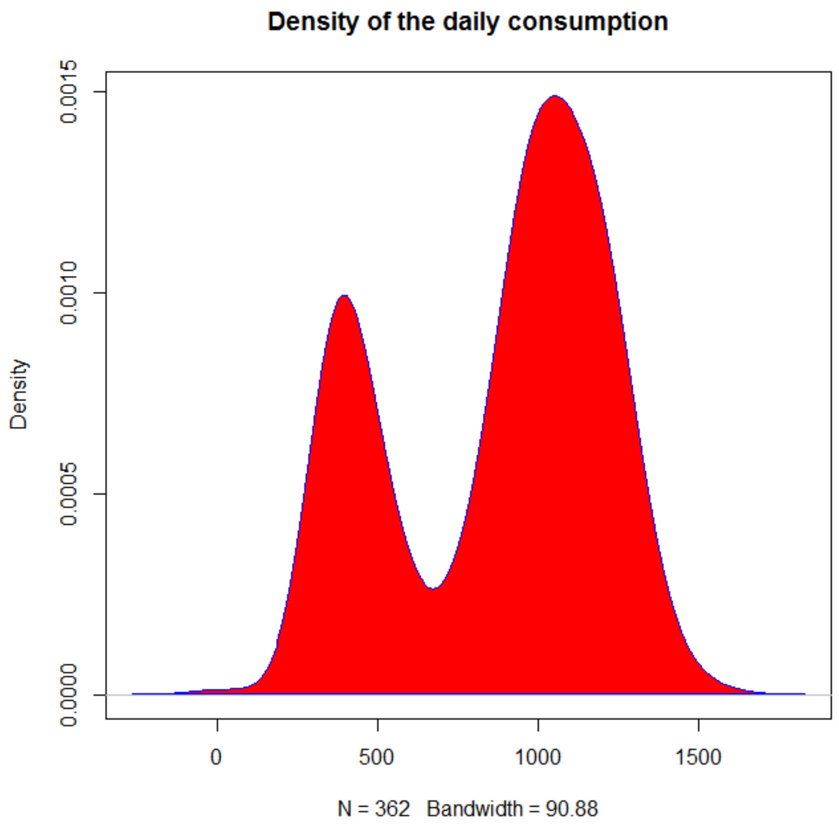


Fig. 28: Density of the daily electricity consumption

This figure shows 2 consumption peaks. The first one is situated around 350 kWh/day(x-axis), this peak belongs to the weekends (closed days). The second peak is situated around 1200 kWh/day and belongs to the weekly consumption. This figure makes it clear that using the whole dataset is not relevant, since it is not a standard distribution. The data which will be useful in the algorithm are the closed day values. As an example the weekend data is used for the previous described tests. Firstly, Fig. 29 shows the density plot of the weekend consumption with on the x-axis the consumption/day and on the y-axis the density.

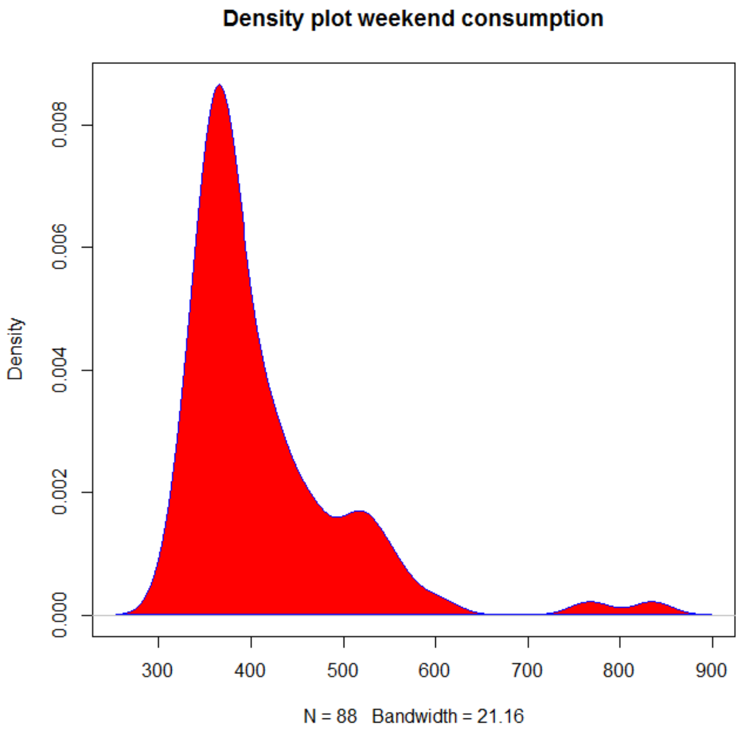


Fig. 29: density plot weekend consumption

If this data can be used by the standard deviation and variance technique depends on the results of the chi-square test and the QQ-plot. Both of these tests are done with the statistical program R. Fig. 30 shows the QQ-plot of the closed day consumption.

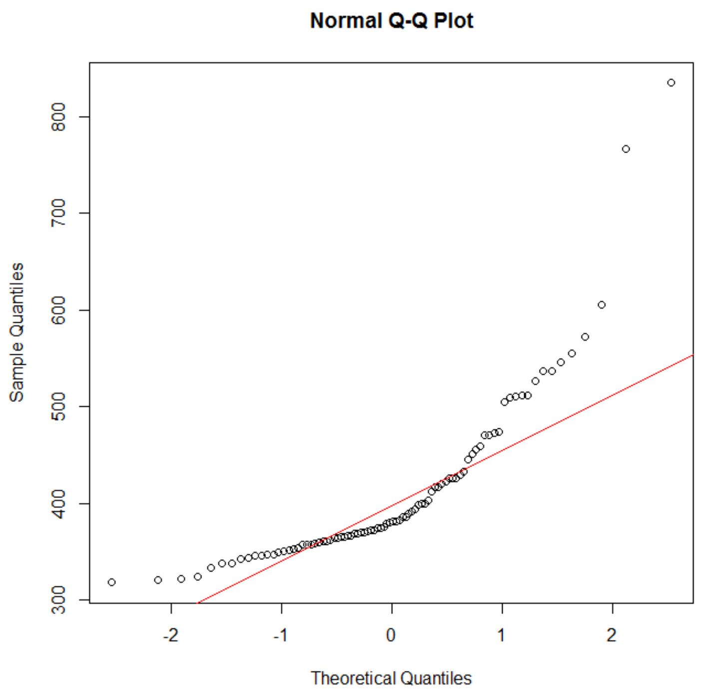


Fig. 30: QQ-plot

Normally, the observations are situated around the red line, which is the trend line. This figure already shows that the dataset of closed days is not a standard distribution. This is also confirmed by the chi-square test. Chi-squared test for given probabilities calculated in R:

X-squared = 1638.767, df = 87, p-value < 2.2e-16

With:
X-squared = the chi-square sum
df= the degrees of freedom
p-value = the probability of a statistical test.

The most important value of this list is the p-value. If the p-value is higher than 0.1, the dataset is a standard distribution. The value 0.1 or 10% is the most common value that is used in chi-squared tests. Like the QQ-plot already showed, this dataset is not a standard distribution. This technique is not the best one because the example showed that a dataset doesn't always have a standard distribution. If a test discovers that the dataset doesn't have a standard distribution, another technique must be chosen. Therefore, it is better to start directly with another technique, the box plot because it is not bound to a specific distribution and the way of calculating the outlier value is straightforward and easy to use.

6.2 Development of the holiday anomaly algorithm

A general description of algorithms is given within the study of the basic principles. That description explained the importance of the BMS. In this algorithm the BMS is needed to supervise the installations. Finally, the devices of the implementation will communicate with the BMS to get control over the installations. Although, the BMS at its own is not the most important part in this algorithm, because the BMS only receives control actions from the developed algorithm. The most important in these algorithms is the data. However, it is possible that this data comes from the BMS.

6.2.1 Electricity algorithm

As the introduction of this chapter explained, the electricity algorithm consists of the basic concept. Therefore, this section will be larger than the gas section. Furthermore, this section is divided in several parts. Since there is a need for data to develop algorithms and a specific data mining technique is needed, section 6.2.1.1 describes the steps which are taken. Section 6.2.1.2 provides a detailed overview of the defined problem. This analysis gives insights into the data which is necessary for the data selection and integration, which is discussed in section 6.2.1.4. This section also describes how the box plot is used in the algorithm. Section 6.2.1.3 explains the use of data integrity in practice. While section 6.2.1.5 describes where the reference data which is used in 6.2.1.4 comes from. The detection of the holiday anomalies is explained in section 6.2.1.6. Finally, section 6.2.1.7 provides a roadmap of the basic concept of the algorithm. The report is not included in that roadmap because it will be explained in section 6.3.

6.2.1.1 *What is done*

The chapter of data selection and pre-processing already explained how the data is extracted from the database. Like the introduction highlighted, data is used from the building of BEMT. But data of the city hall of Vorselaar is also available for use. Furthermore, it discussed a couple of pre-processing steps. Several of these steps will be reviewed later in this part. Moreover, section 6.1selected the best data mining technique. So the box plot is the technique used for this algorithm.

6.2.1.2 *Review of the problem*

The introduction explained the general problem of holiday anomalies. But it did not provided a detailed overview. In this section it is recommended to analyse the problem in more detail, because it results in a better insight of the expected algorithm. Fig. 31 shows a week profile of the electricity consumption for BEMT which visualises the problem.

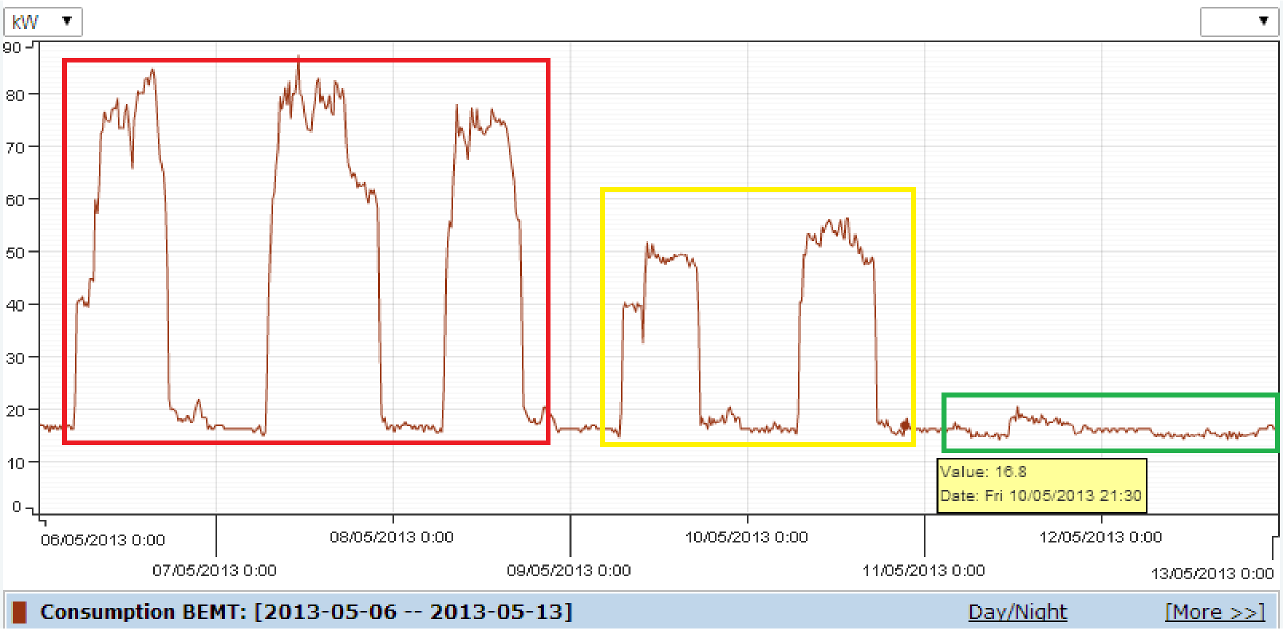


Fig. 31: electricity consumption with anomaly

The problem is situated in the yellow rectangle. The first day is the 9th of May 2013 and is a legal holiday, Ascension day. The second day is also a holiday which is defined as a building holiday. In comparison with the weekend consumption or closed day consumption (green rectangle) the consumption of the holidays is much higher. In comparison with the normal workday consumption (red rectangle) the holiday consumption is lower. Therefore, these holidays are definitely anomalies. Furthermore, based on this overview it is clear that several datasets could be extracted from the original dataset (daily electricity consumption over a relevant period) for the box plot analysis.

6.2.1.3 *Data integrity*

The algorithm gets an original dataset from the database. This set consists of the electricity consumption in 15 minute values, over a relevant period for the organisation. Furthermore it is recommended to use data in kW. To find the holiday anomalies, the algorithm will extract several times data from the original dataset. But it is important that the data (daily consumption values) has a good quality, since the daily consumption is calculated as the average of the 15 minute values. Therefore, the 15 minute values are used to calculate the integrity. Other calculations are all based on daily consumption values.

The algorithm uses several rules to deal with the integrity. A day with less than 75% of its values or in other words a day with more than 24 missing values is excluded from the dataset. Because 24 missing values stands for 8 hours without data which is already a broad tolerance. Other days with a data integrity between 100% and 75% are scaled to their number of measurements. For instance, a normal day has 96 measurements. Therefore, the daily consumption of a day with an integrity of 100% is calculated by the sum of each value divided by 4 (to change the units kW ➔ kWh). While the daily consumption of a day with a data integrity of 80% or with 77 measurements is calculated by the sum of all values divided by 4 and multiplied by 100 and divided by 80 (data integrity). The data integrity is also shown in the final report. Extra information is available in section 4.2.1.2.

6.2.1.4 *Data selection and integration for box plot analysis*

The box plot is a statistical technique to find outliers/anomalies. Therefore, the five-number summary can be calculated on any kind of extracted data. But it does not mean that the generated output is relevant for the user. Based on the detailed review of 6.2.1.2 there are several datasets available in the original set,

1. workday consumption as reference + holiday consumption;
2. workday consumption as reference,
3. closed day consumption as reference + holiday consumption;
4. closed day consumption as reference,
5. holiday consumption.

The best dataset is selected on the previous introduced constraints for the algorithm. The general objective is repeated to make it clear: *the algorithm uses historical data to detect all the holiday anomalies. Once these days are located, the results must be visualised (number of anomalies, **amount of electricity lost** and amount of money lost).* Furthermore, section 6.1 introduced an extra prerequisite, the user of the algorithm must have **access to the outlier value**.

The bold parameters are the constraints for the dataset, since a predicted consumption is needed for the calculation of the lost electricity, it is a constraint of the data which is used. Based on this constraint the first and the second proposal are not relevant, most of the time the workday consumption is higher than the holiday consumption. Therefore defining a predicted value with values that are higher is not relevant. The second bold parameter is related to the constraints of a box plot. A box plot always uses the whole dataset to calculate the five-number summary. It also uses the outliers/anomalies in its calculation. These outliers/anomalies have an impact on the defined outlier value. This is shown in the next example: Table 4 shows a dataset without outliers. The values are generated in R.

Table 4: normal dataset

normal data							
277,15	297,31	320,08	300,52	283,44	287,29	281,63	309,51
314,55	331,39	292,33	295,51	303,19	300,88	311,62	308,34
297,13	300,19	309,77	313,41				

For the box plot analysis, the five-number summary is calculated. Table 5 gives an overview of this summary.

Table 5: five number summary normal data

normal data	
Q1	294,72
min	277,15
med	300,7
max	331,39
Q3	310,23
IQR	15,518
outlier	333,51

The outlier value is around 334 while the maximum value is 331. For sure, this dataset doesn't have outliers or anomalies. To analyse the difference, the second dataset uses the same 20 values together with some outliers. Table 6 shows the new table with outliers

Table 6: outlier dataset

data with outliers							
277,15	297,31	320,08	300,52	283,44	287,29	281,63	309,51
314,55	331,39	292,33	295,51	303,19	300,88	311,62	308,34
297,13	300,19	309,77	313,41	420	333	401	387
365	412	343					

The presence of the outliers in the dataset results in a different five-number summary. Table 7 shows the new summary.

Table 7: five number summary outlier data

outlier data	
Q1	297,22
min	277,15
med	309,51
max	420
Q3	332,2

IQR	34,975
outlier	384,66

An analysis of Table 5 and Table 7 results in an IQR which is more than the double of the IQR of the dataset without outliers. Furthermore, the difference between the outlier values is 50. This has a serious impact on the output of the algorithm. So, this example provides 2 conclusions. Firstly, data of holidays is not allowed in the tested dataset, since they have a strong influence on the outlier value. This means that proposal three and five are not relevant for analysis. This means that proposal four is the best dataset. Secondly, it is strongly recommended to prevent outliers in the dataset which is used as a reference. There are 2 possibilities to have anomalies on closed days. Firstly, there was an event during the day. Secondly, one of the installations did not change its work modus. Which means that the energy consumption has been higher than normal. Event anomalies are deleted by data integration, which is done within the algorithm. Normally, the information is taken from the configured building in EcoSCADA, but events are also an input parameter of the algorithm. So, each closed day with an event is excluded from the set. Furthermore, an additional algorithm is developed to exclude anomalies on closed days, this is explained in the appendix. The final dataset which will be extracted from the original set consists of the consumption values and timestamps on closed days without events and without anomalies.

6.2.1.5 *Practical case closed days*

Section 6.2.1.2 discussed for the first time the concept of closed days. These days are used as a reference for the box plot analysis 6.2.1.4. Closed days are integrated in the original dataset by the algorithm (data integration). This information comes from the three layer calendar in EcoSCADA and is linked to the building. Fig. 32 shows this information for BEMT.

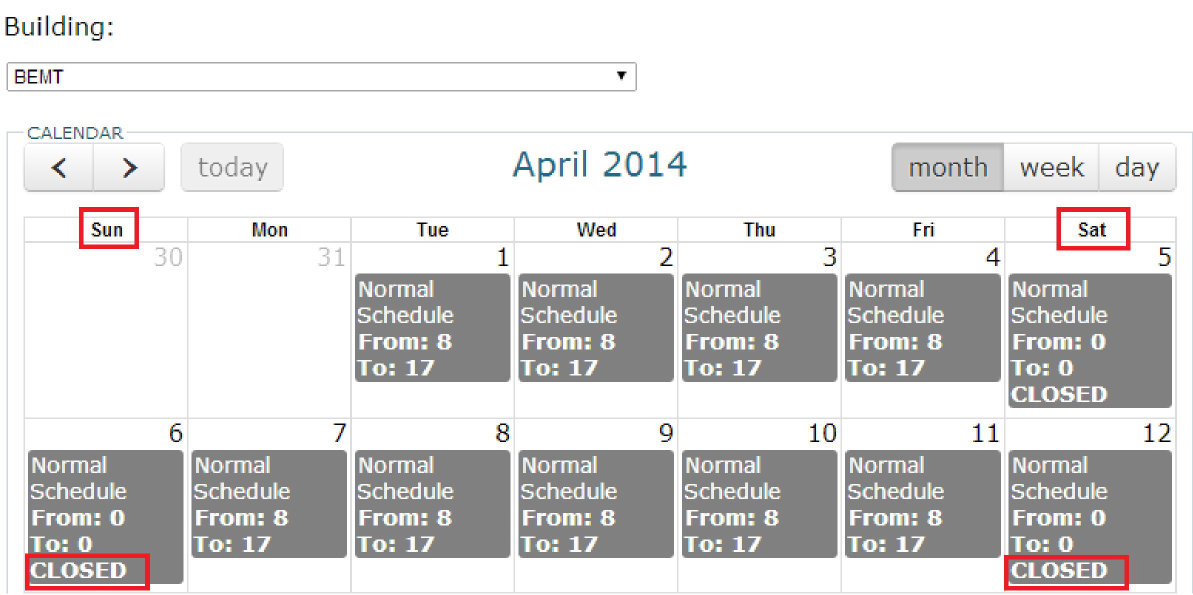


Fig. 32: closed days for BEMT

The closed days for BEMT are situated in the weekends. The use of this data makes the algorithm more flexible. Since there are several buildings like supermarkets which are open on weekends but which are closed a day or multiple days in the week. For instance, a supermarket is open on Saturday and is closed on Sunday and Monday. If the algorithm used only weekends as closed days, it would have had information about Sunday but not about Monday. Since each Saturday would have been an event for the algorithm. Therefore, it is recommended to use the extra input of Fig. 32, because then the algorithm would have used Sunday and Monday as closed days.

6.2.1.6 *Holiday anomaly calculation*

Data of closed days without events and anomalies is used as input for the box plot. Based on this information the box plot calculates the five-number summary. Out of this summary, the median and outlier value are the most important. While the median is discussed in section 6.3, the outlier value will be used now. The outlier value is implemented in a rule, as described in the previous chapter. Generally, the consumption of each holiday is compared with the outlier value. Each holiday with a consumption higher than the outlier value is described as an outlier. The algorithm knows which days that are holidays by data integration of the algorithm. This holiday information comes from EcoSCADA, but it is also an input parameter of the algorithm.
In syntax code this rule may be:

```
FOR i=first_holiday TO n-holiday
        IF outlier_value < consumption_holiday THEN
                Holiday is outlier
        ELSE
                Holiday is in control
        END_IF
NEXT
```

6.2.1.7 *Roadmap summary*

Fig. 33 shows a roadmap summary of the holiday anomaly algorithm. The algorithm summary is based on this figure.

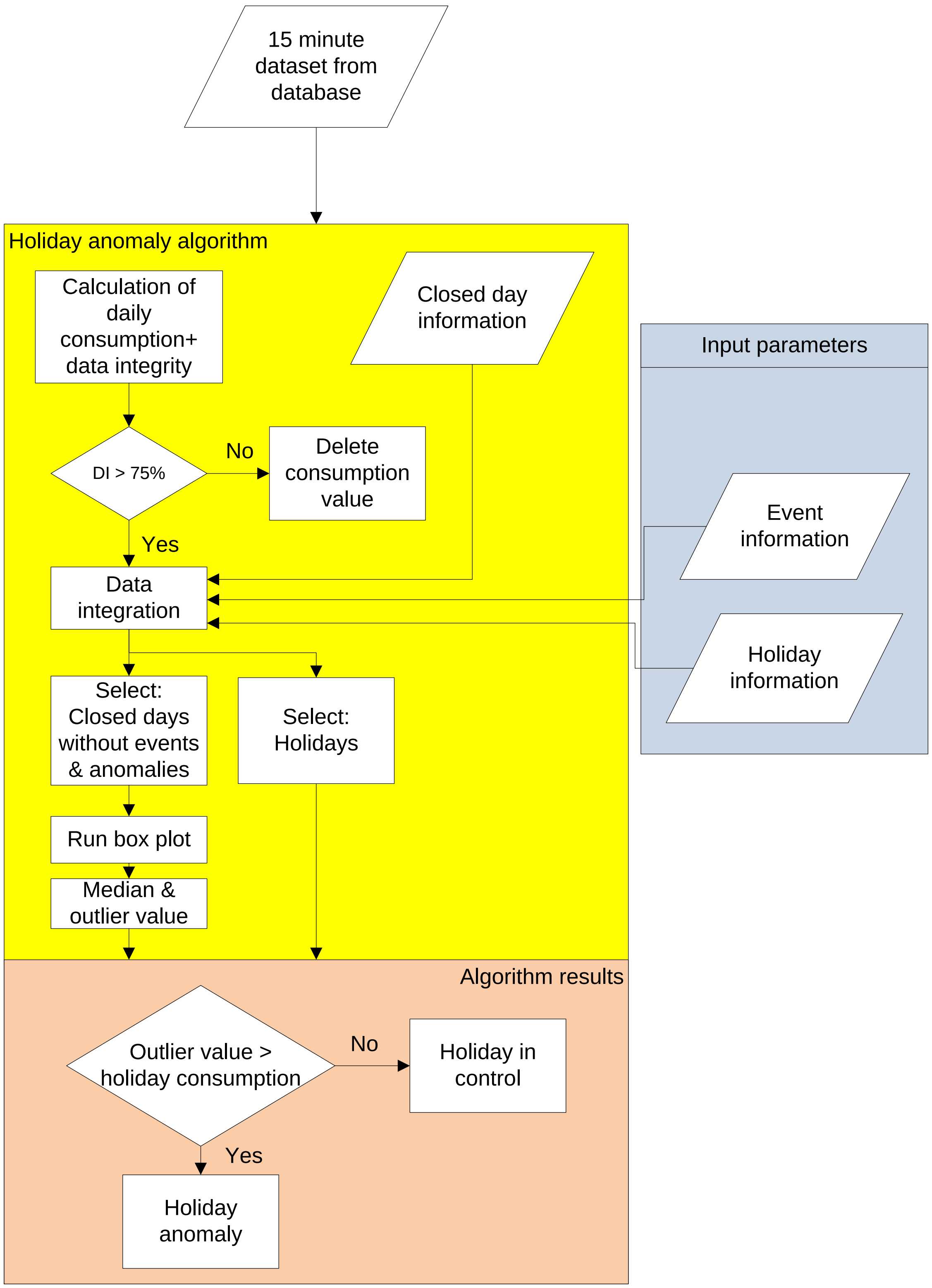


Fig. 33: holiday anomaly algorithm roadmap

Before the algorithm can run, the user has to configure minimum 3 things. Firstly, he has to provide a 15 minute dataset from the database in kW. Secondly, he has to select a building, since all the data of closed days, events and holidays is linked to a specific building. Although, like the roadmap shows, it is possible for the user to configure the holidays and events before the algorithm is started. As a third item the user must provide the cost price of the electricity that is used for the savings. Once these items are configured, the algorithm can be used for analysis of the dataset. The output of the algorithm consists of the timestamps and consumption values of the holidays with an anomaly. This data is used in section 6.3 to calculate the KPI, EnPI and the savings. Finally, all the results are presented in a report.

6.2.2 Gas algorithm

The principle of finding holiday anomalies in the gas consumption is the same as with the electricity algorithm. Although, the gas data needs more pre-processing before the algorithm can use the data. Furthermore, one extra input parameter is needed in comparison to the electricity algorithm. Because of these reasons, the gas algorithm adds an extra step and parameter to the basic concept. Therefore, section 6.2.2.1discusses this extra step and parameter, while section 6.2.2.2 visualises the final roadmap of the algorithm.

6.2.2.1 *Pre-processing of gas data*

The algorithm uses the electricity data as absolute values. There is no general accepted way to normalize electricity data. This is different for gas data. The study of the basic principles highlighted the use of degree days to normalise the gas consumption. Furthermore, it also explained the calculation of DD and discussed where the weather data comes from. Therefore, the pre-processing of gas data means the normalisation of the data with degree days.

To start the calculation of DD a base temperature is required. This temperature is selected in function of the building. But nowadays, there are more and more buildings that don’t have one set temperature any more. For instance in BEMT, a room which is not used has a set temperature of 18°C. But when somebody enters the room the heating starts-up and the room is heated towards its set temperature. Each room has the capability to control its set temperature by a local temperature controller. It is difficult to select a relevant base temperature for such a building, since the number of degree days depends on the selected base temperature. Therefore, it is possible to select the base temperature by linear regression. Fig. 34 shows the regression plot between the consumption and the number of DD.

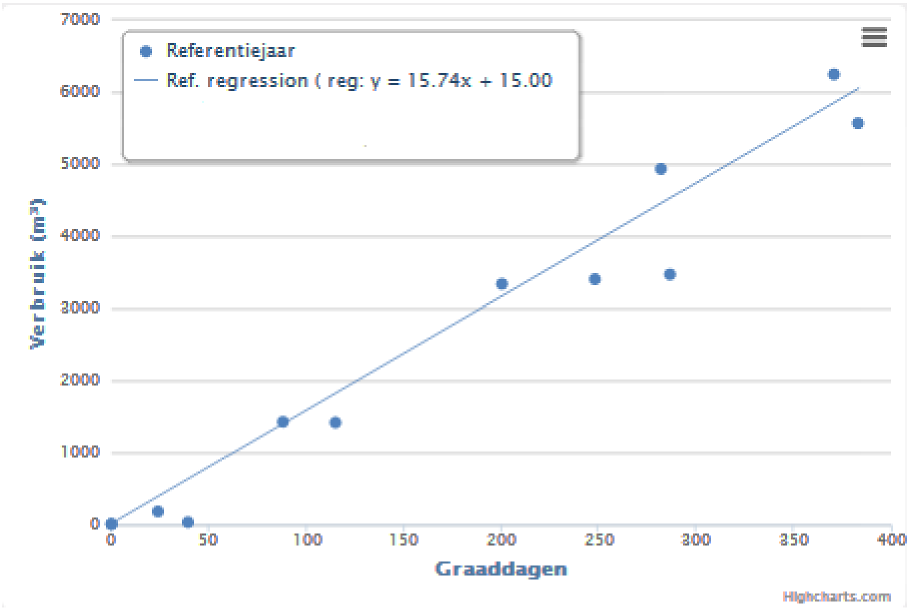


Fig. 34: example gas normalisation

The base temperature for this figure is 18.5 °C. If another temperature was selected, the slope of the regression line would have changed. Selecting another base temperature doesn't mean that the correlation with the DD would be better. Therefore, the best DD temperature is selected by using statistics. In statistics the quality of the correlation between the number of DD and the gas consumption is given by the coefficient of determination, R^2 (Proportion of variance explained). This coefficient is a number between 0 and 1. An R^2 of 1 indicates that the regression line perfectly fits the analysed data. Furthermore, an R^2 of 0 indicates that there is no relationship between the variables. This means that the highest R^2 is seen as the best correlation. The algorithm uses this kind of analysis to find the best base temperature. Table 8 provides an overview of the R^2 for the BEMT building.

Table 8: R^2 overview

	DD T13	DD T15	DD T16,5	DD T18	DD T21	DD T28
R^2	0,771848	0,776591	0,768644	0,757704	0,734921	0,722257
R^2 max	0,776591					
Best DD	15.00					

There is not so much difference between the different DD calculations. But the max R^2 belongs to the DD calculation with 15°C as temperature. As a result this temperature is used for the calculation of the DD for BEMT. As the study of the basic principles discussed the weather information is loaded from a weather station. Therefore the extra parameter which is used by the gas algorithm is the selection of the weather station. It is recommended to use a station close to the building.

6.2.2.2 *Pre-processing roadmap*

Since the main principle of the algorithm is not changed, the pre-processing of the gas data is done in the beginning of the algorithm. Therefore, Fig. 35 provides an overview of this pre-processing step together with the extra input parameter.

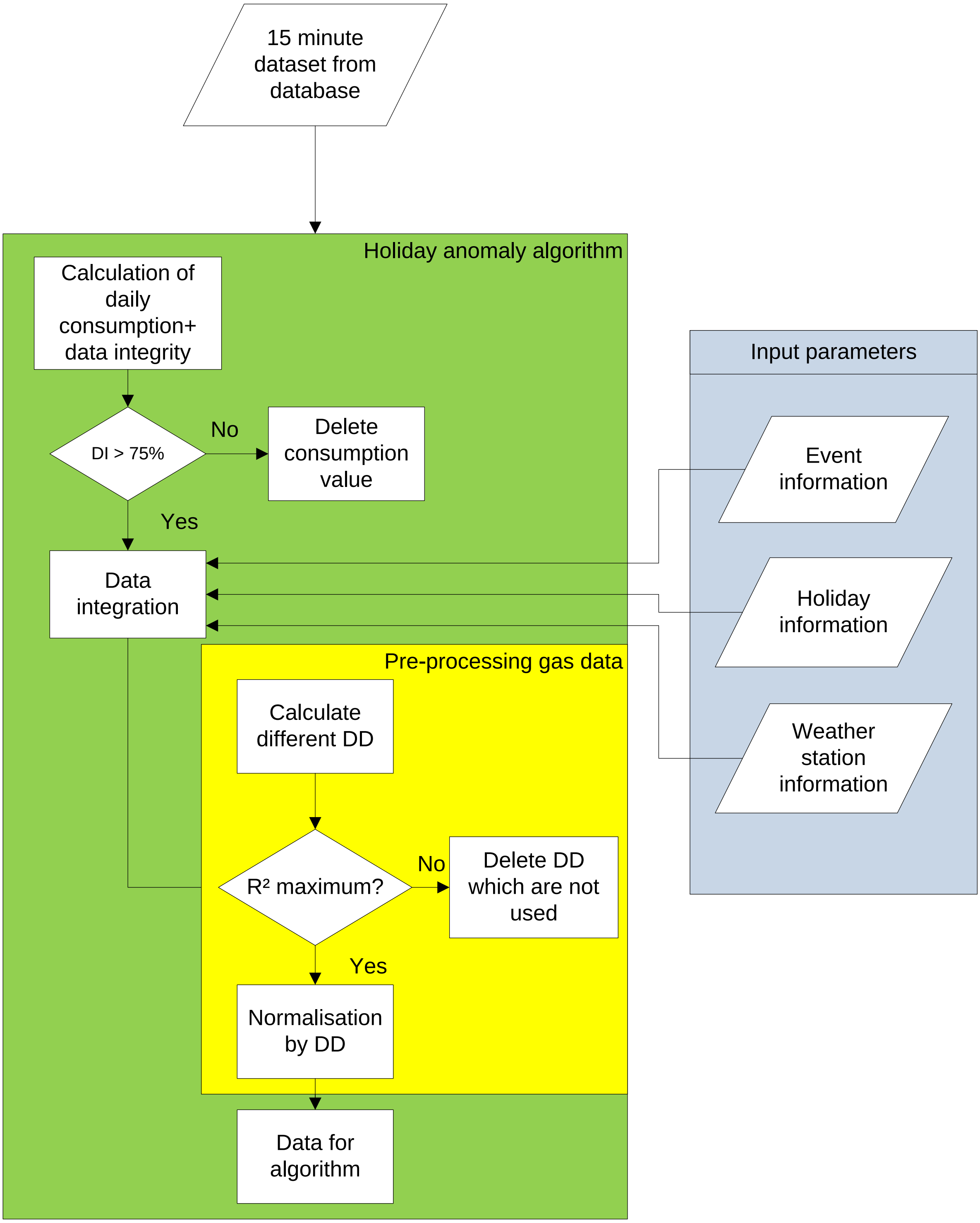


Fig. 35: gas pre-processing roadmap

The provided dataset must contain gas data with a 15 minute timestamp and the unit must be m^3. Furthermore, the blue figure shows the extra input parameter 'weather station information'. Moreover, the data integrity test is still the first step in the algorithm, because the output quality strongly depends on the quality of the data. Therefore, the pre-processing is situated between the data selection (explained in the electricity algorithm) and the data integration step. Finally, the algorithm will use gas data which is normalised by DD.

6.3 Holiday anomaly report

This chapter consists of several separate sections. Each section focuses on one item of the report. The number of sections is extracted from the definition of the algorithm: *the algorithm uses historical data to detect all the holiday anomalies.* ***Once these days are located****, the results must be visualised (****number of anomalies, amount of electricity lost and amount of money lost****).*Based on the definition one can conclude that there are 4 sections. Firstly, the detailed results of the algorithm are visualised. These results are used for the calculation of the other parameters as explained in the definition. Since the algorithm must fit within ISO 50001, the other 3 requirements are renamed.

- number of anomalies = KPI;
- amount of electricity lost = EnPI;
- amount of money lost = savings.

6.3.1 General algorithm output

The results of the data scan must be visualised once the anomalies are detected. Therefore, the final report starts with the output results of the algorithm. Table 9 shows how the output of the algorithm is shown in the report.

Table 9: algorithm output

Day	Electricity [kWh]	Data Integrity [%]	Electricity Norm by DI	Holidays	Event	Closed day	Outlier Holidays
02/01/13 0:00	933,20	100	933,20	1	0	0	Outlier Holidays
01/04/13 0:00	933,35	100	933,35	1	0	0	Outlier Holidays
01/05/13 0:00	874,72	100	874,72	1	0	0	Outlier Holidays

This table visualises the different data integration and data selection steps of the previous chapter. The first two columns represent the data selection. The third, fourth and last column indicate the data transformation step while the other columns represent the data integration step.

There is a difference between the output of the gas algorithm and the output of the electricity algorithm. The values of the gas algorithm are normalised by DD. The values of the electricity algorithm are not normalised. Therefore, both algorithms use a different calculation method to find the EnPI and savings, which are explained in sections 6.3.3 and 6.3.4.

6.3.2 KPI

Since a KPI is useful in decision-making process and may be used throughout all levels of a company, it must be easy to understand for everyone. Therefore, the visualisation of the KPI is the same for the gas- as for the electricity algorithm.

The easiest way to tell top management that there is a problem with the energy consumption on holidays, is by giving them the number of days with a holiday anomaly. For instance: assume that a building has 21 holiday anomalies on 26 holidays and that the sum of lost electricity is 10000 kWh. The energy representative must explain these results to top management. He may use the number of days to explain the problem or he may use the 10000 kWh. He tells top management that their building has consumed 10000 kWh too much on holidays. Top management will have difficulties to know if it is a significant problem or not. But this is not a problem when the number of days is used. Because everyone knows that 21 anomalies on 26 days is a significant problem. Therefore, the KPI for the holiday anomaly algorithm is represented by the number of anomalies.

6.3.3 EnPI

The energy management part of the study of the basic principles highlighted the use of EnPIs, measuring how good something functions. This EnPI is used to measure how well the BMS acts on holidays. Within ISO 50001 there are several methods to calculate an EnPI. Such an EnPI may be a simple parameter or ratio but it may also be a complex model. Furthermore, there are a couple of possibilities to visualise EnPIs. It is possible to visualise EnPIs per time, per unit of production and variable combinations are possible [3]. In case of the holiday anomaly algorithm the EnPI will be displayed per unit of consumption. This is for gas in m^3 and for electricity in kWh.

But which EnPI value indicates a good energy performance on holidays and which value does not? This step is important to understand the calculation method of the EnPI, since an EnPI is derived from a KPI it is relevant to use the result of the KPI for this analysis. The KPI provides information about the number of holiday anomalies. If the KPI = 0, there is no problem. On such a day there was an amount of energy consumption but this consumption was not significantly higher than on closed days. If the KPI >0 then there are holiday anomalies. On these days the energy consumption is higher than normal. If there are no anomalies the EnPI is also 0 otherwise the EnPI is > 0. Finally, the calculation method is divided in 2. One method in case of the electricity algorithm and one in case of the gas algorithm.

6.3.3.1 *EnPI for electricity algorithm*

This EnPI is visualised in kWh and shows the amount of lost electricity in comparison with a normal closed day. This is calculated by the use of a simple mathematical function.

$$EnPI = \sum_{k=1}^{k=n} \left(Real_{consumption} - Predicted_{consumption} \right)$$

With:

k = the index of the holiday anomalies
n= the maximum number of holiday anomalies (KPI)
$Real_{consumption}$= the electricity consumption of the analysed anomaly
$Predicted_{consumption}$ = the electricity consumption in case of a normal holiday

The electricity consumption is not normalised by another dataset. So it is difficult to find other similar days with a normal consumption profile. Therefore, the median value of the closed day dataset is used as $predicted_{consumption}$ value. This is also the reason why the median value is extracted from the box plot.

6.3.3.2 *EnPI for gas algorithm*

This EnPI is visualised in m³ and shows the amount of lost gas in comparison with a normal closed day. The calculation of the EnPI is also based on a mathematical function. Although, this function is more complex than the function of the electricity algorithm.

$$EnPI = \sum_{k=1}^{k=n} \left(Real_{consumptionbyDD} * (number\ of\ DD) - Predicted\_consumption \right)$$

With:

k = the index of the holiday anomalies
n= the maximum number of holiday anomalies (KPI)
$Real_{consumptionbyDD}$= the gas consumption normalised by DD of the analysed anomaly
Number of DD= the number of DD of the analysed anomaly
$Predicted_{consumption}$ = the gas consumption in case of a normal holiday

As the mathematical function shows the gas data is normalised by DD. The integration of the temperature in the dataset makes it possible to find similar days. Such a similar day predicts the consumption of a closed day with almost the same temperature. The technique used to predict this consumption values based on the temperature is a decision tree. Fig. 36 shows how this technique is used.

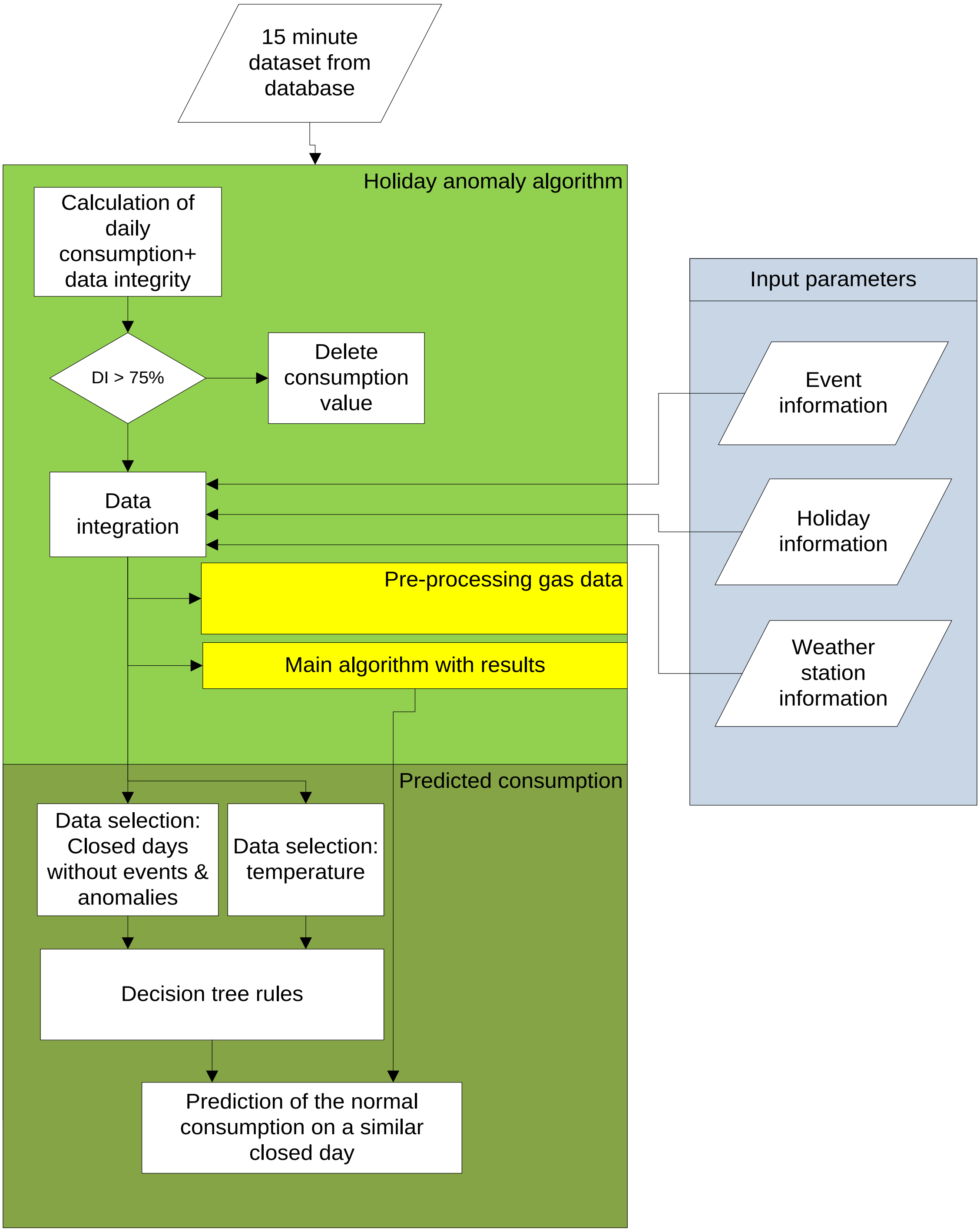


Fig. 36: prediction technique

The previously explained techniques and algorithm is not visualised anymore. These visualisations are replaced by the yellow rectangles. The focus of this figure is situated in the predicted consumption. The same data as for the box plot is selected to run the decision tree. Besides the consumption data on closed days the decision tree also uses the external temperature to calculate its rules. Once the rules are generated the results of the algorithm are used to predict the consumption of a similar closed day. This is visualised in the last rectangle. The statistical program R is used to calculate the decision tree rules. The decision tree method used in this example is the regression tree, this method is based on the smallest sum of squares. Therefore, the decision tree algorithm creates a new leaf if this leaf reduces the sum of squares [40]. Fig. 37 shows the decision tree which is based on the data of BEMT.

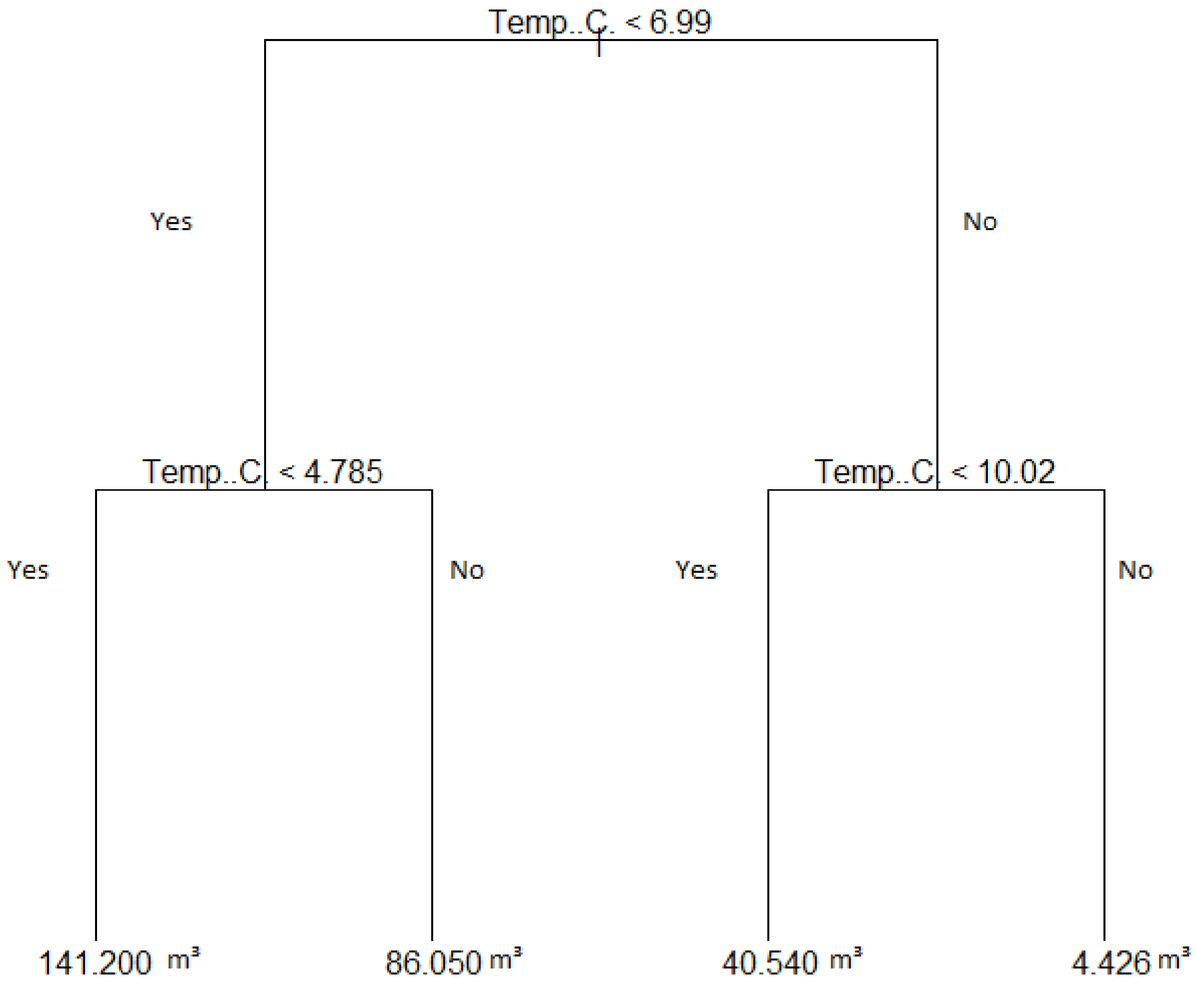


Fig. 37: decision tree

If an anomaly holiday has a temperature below 4.785 °C the predicted consumption for that day is 141.2 m³. The same principle is used for the other temperatures.

6.3.4 Savings

The savings are calculated based on the EnPI value. Because the EnPI visualises the amount of lost energy. Therefore, it is a small step towards the calculation of potential savings. Once the energy price is known the potential savings can be calculated. So the energy price is the last input parameter necessary for the algorithm. Fig. 38 shows how the potential savings are calculated.

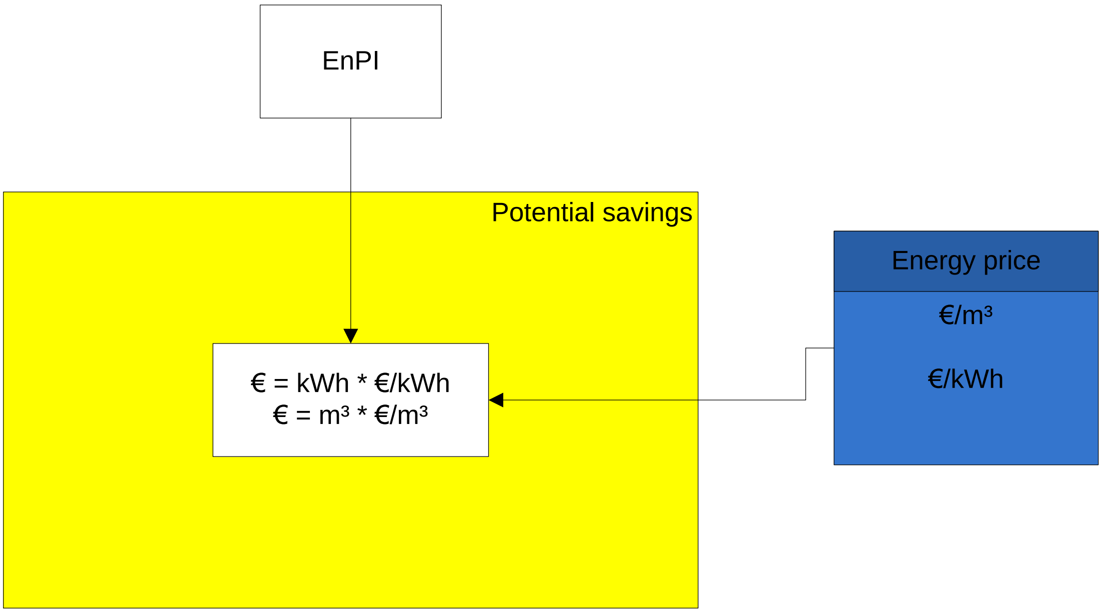


Fig. 38: calculation of potential savings

6.4 Getting a report out of the algorithm

This section discusses the use of the interface to receive a holiday anomaly report. Before the necessary proceedings with the interface are explained, a general overview of the whole algorithm from dataset to report is visualised in Fig. 39. This overview gives the first introduction to the use of the interface. Because the interface for the electricity algorithm will differ from the interface of the gas algorithm. Once the differences are explained the interfaces are visualised. Finally, the generated report of the algorithm is explained.

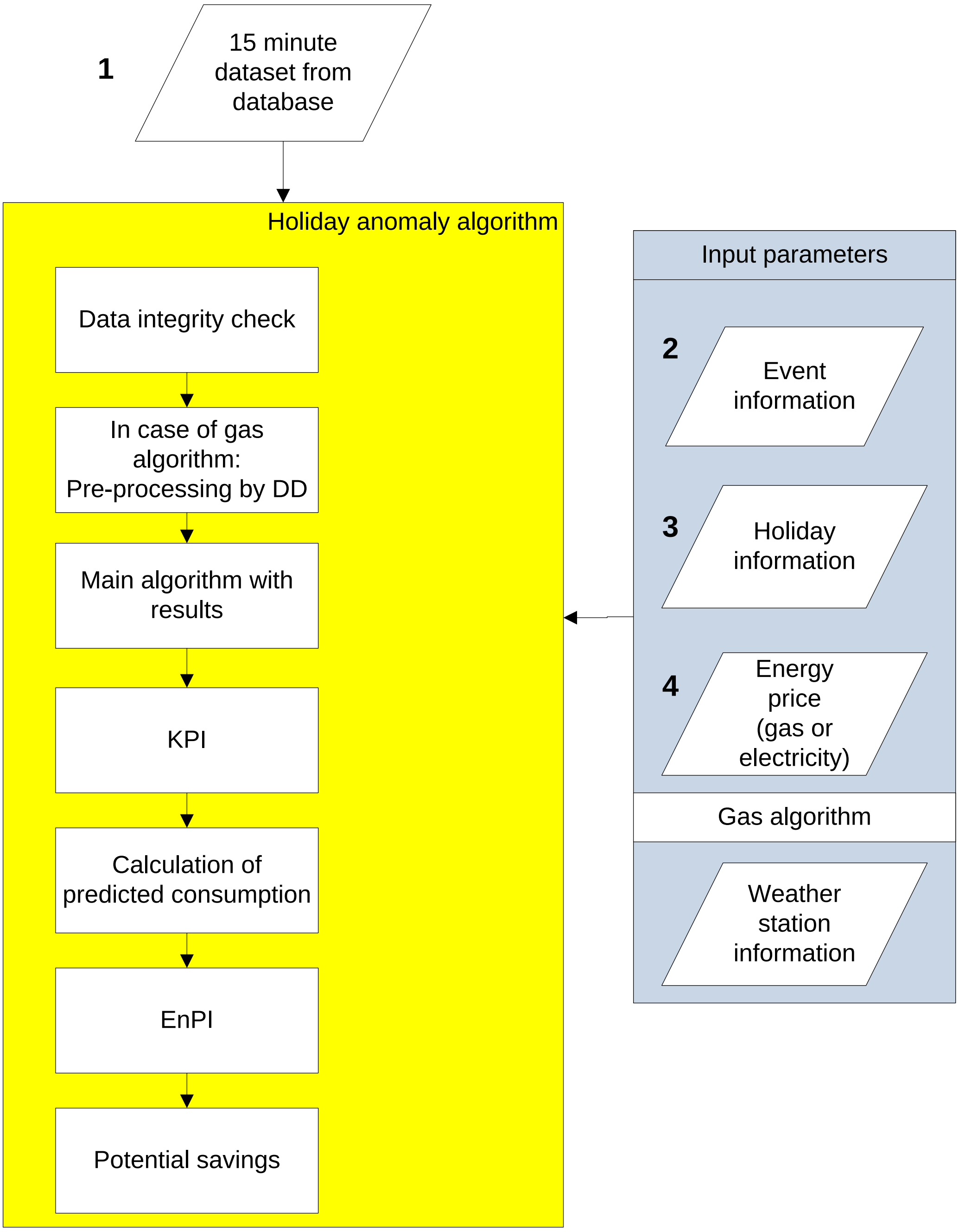


Fig. 39: roadmap holiday anomaly algorithm

In case of the electricity algorithm there are 4 parameter-sets necessary (numbers in Fig. 39) to start the algorithm. One extra parameter is necessary for the gas algorithm, since the data is normalised by DD. Furthermore, there are 2 differences between the gas and electricity algorithm. Firstly, the gas data is normalised by DD in the pre-processing step. Secondly, the predicted consumption is calculated differently. In case of the electricity algorithm the median consumption on closed days is used as predicted value, while the gas algorithm uses a decision tree to predict the consumption. Finally, it is important that the user of the algorithm receives an easy to understand report. Furthermore, it must be possible for any kind of user to analyse the results. Therefore, the final report is downloaded as an excel file from the algorithms website 'www.energydatascan.com'.

Fig. 40 and Fig. 42 shows how the input parameters are configured on the website.

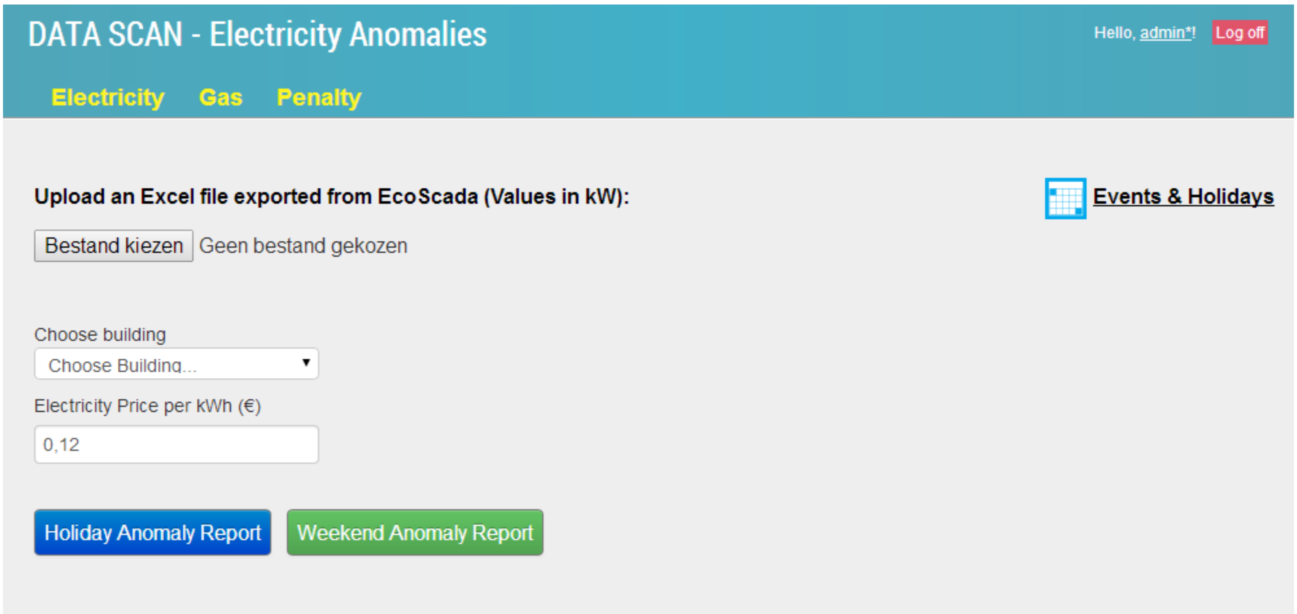


Fig. 40: configuration electricity algorithm

Within this figure it is not possible to configure the holidays and events. Therefore, one must select 'Events & Holidays' on the right side of the figure. Fig. 41 shows the screen which is visualised by clicking on the 'Events & Holidays' button.

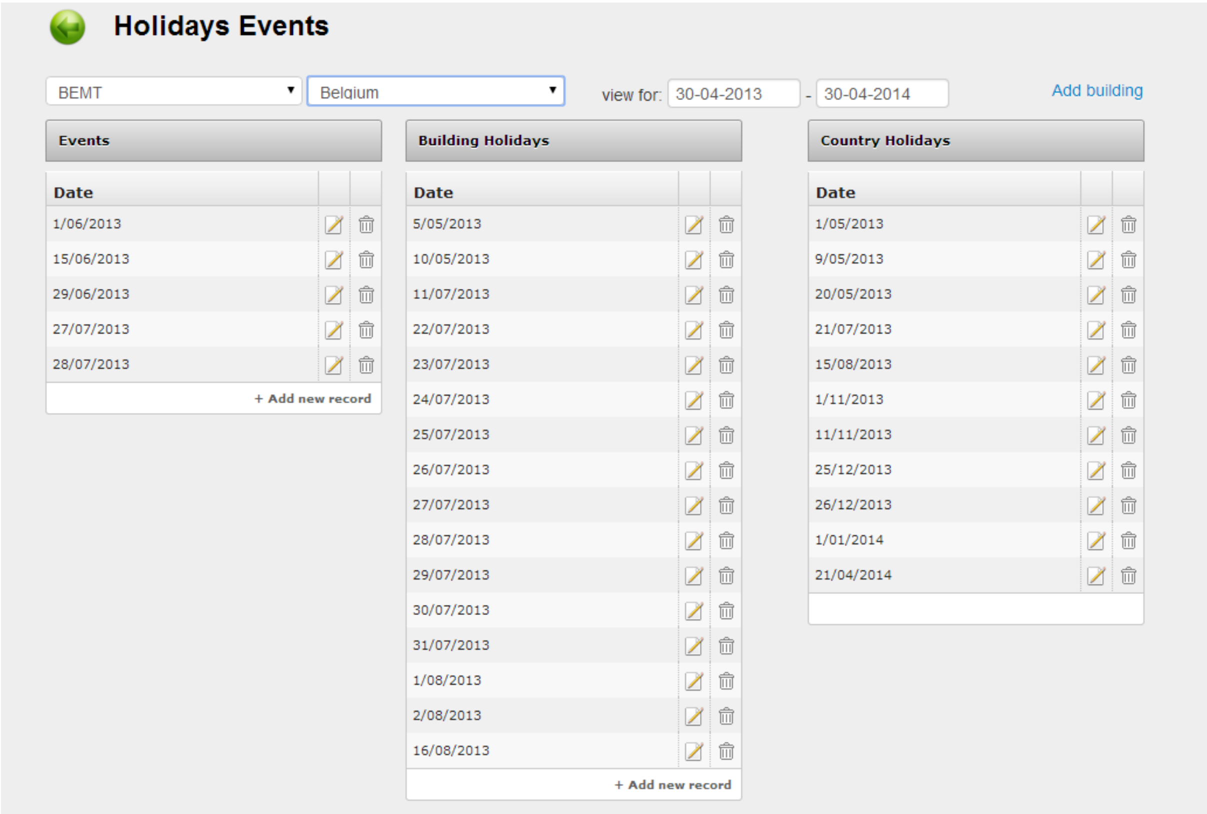


Fig. 41: configuration of events and holidays

The use of the electricity algorithm is explained by using Fig. 40 and Fig. 41. As the overview described, 4 parameter-sets are necessary to generate an holiday anomaly report for electricity. Due to the interface it is possible to generate a report based on 3 parameter-sets, because parameter-sets 2 and 3 of Fig. 39 are combined to one parameter-set in the application. Although, this is only possible if all the holidays and events are available for a specific building. Fig. 41 shows how the configuration is done in case that the data is not available in the beginning. Firstly, one needs to select the building and the country of the building. Once these parameters are configured, it is possible to add or remove holidays or events as showed in the figure. This data will be linked to the building. By selecting a building in the interface (Fig. 40), the event and holiday information becomes available for the algorithm. This is the reason why only 3 parameters are required to generate a report. Furthermore, the 2 other parameters are:

- select a local dataset with 15 minute values in kW;
- define the electricity price.

In case of the holiday anomaly report for gas, one extra parameter is needed since the gas data must be normalised by DD. Therefore, the selection of the weather station is integrated in the interface as 'Choose Weather Station'.

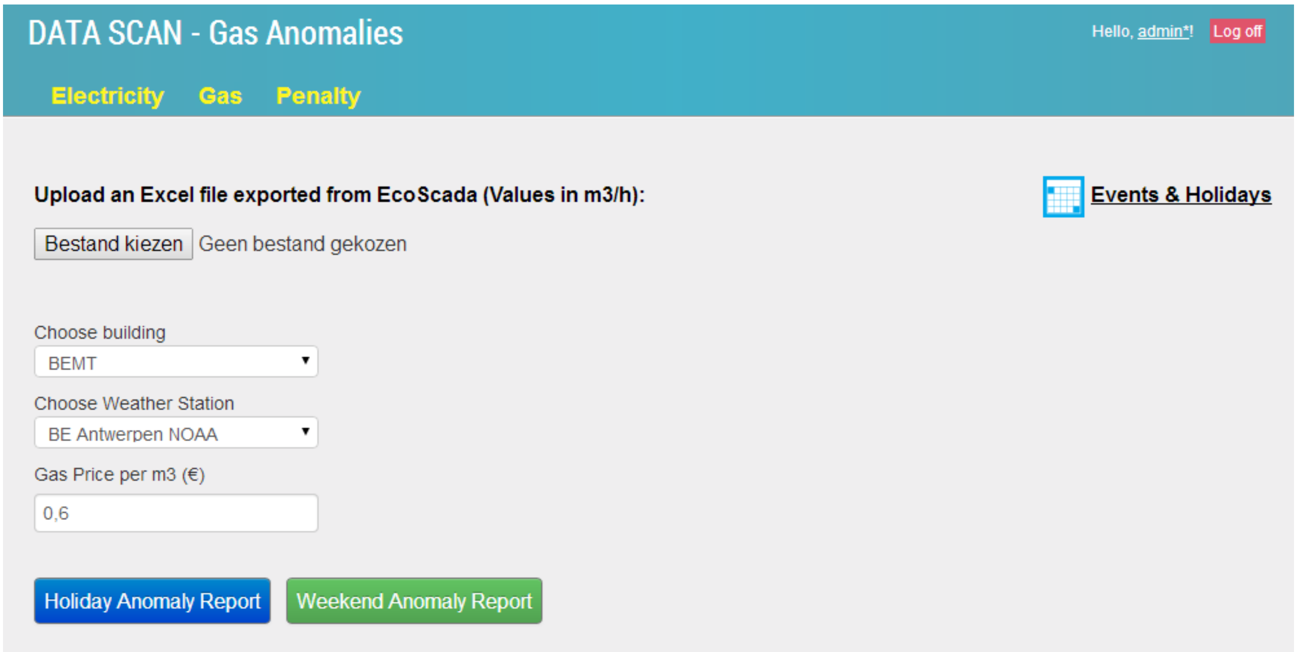


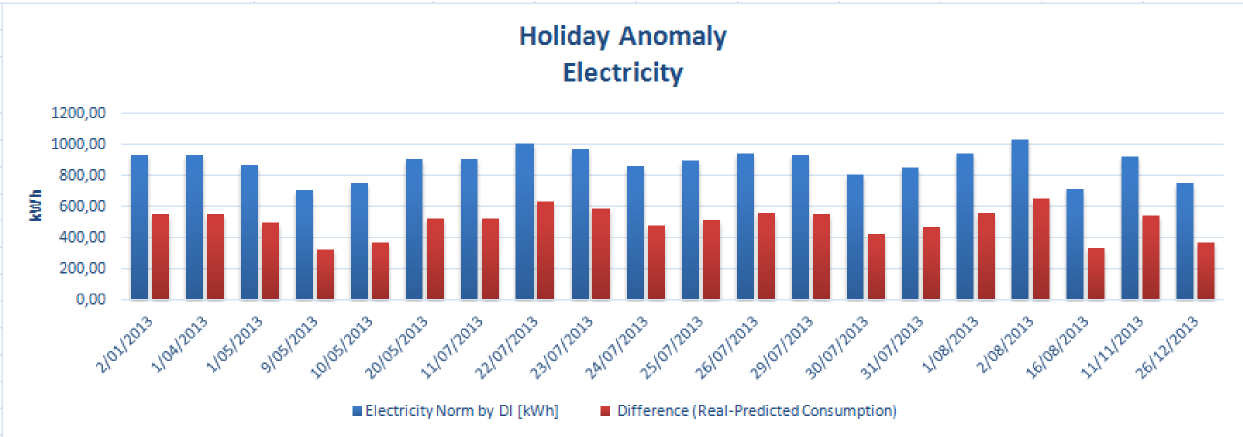
Fig. 42: configuration gas algorithm

Once the input parameters are configured, the holiday anomaly report may be generated by clicking on the blue button. The report will be generated and downloaded as an excel file. Fig. 43 shows what the report looks like

BEMT

Day	Electricity Norm by DI [kWh]	Difference (Real-Predicted Consumption)
02/Jan/13	933,20	552,91
01/Apr/13	933,35	553,06
01/May/13	874,72	494,43
09/May/13	706,12	325,83
10/May/13	755,23	374,94
20/May/13	904,75	524,46
11/Jul/13	903,80	523,51

22/Jul/13	1010,90	630,61
23/Jul/13	973,12	592,83
24/Jul/13	858,02	477,73
25/Jul/13	901,50	521,21
26/Jul/13	942,35	562,06
29/Jul/13	931,85	551,56
30/Jul/13	803,05	422,76
31/Jul/13	851,67	471,38
01/Aug/13	940,88	560,59
02/Aug/13	1031,75	651,46
16/Aug/13	712,88	332,59
11/Nov/13	928,50	548,21
26/Dec/13	749,38	369,09
Sum	**17647,02**	**10041,22**



Price per kWh (Electricity)	0,12 €
Money lost	**1.204,95 €**

Fig. 43: electricity report

The left table shows all the holiday anomalies with the real consumption and with the difference between real and predicted consumption which also gives the lost energy per anomaly. The sum of all these values is the EnPI and is visualised as the last cell of that column. Furthermore, the potential savings are calculated like explained before and are visualised as money lost. Finally, this sheet consists of a bar graph which indicates the real consumption and the difference between the real and predicted consumption. Besides this report there are 2 more sheets. Firstly, the data sheet consists of the whole dataset in daily consumption values. This dataset also contains other information which comes from the data integration step. Secondly, the outlier sheet summarizes the data of all holiday anomalies.

6.5 Algorithm according to ISO 50001

The energy management part of the study of the basic principles provides a general overview of all ISO 50001 sections. In this section, it is explained how the holiday anomaly algorithm fits in ISO 50001. This is an important part since it was one of the requirements. The sections of ISO 50001 where this algorithm is appropriate for, could be found by analysing what it exactly does. The holiday anomaly algorithm provides information about:

- opportunities to save energy,
- installations of significant energy use on holidays (heating, airco, ventilation,..);
- baseline holiday consumption,
- evaluate the past energy consumption on holidays.

As defined in the introduction the algorithm always uses historical energy data. The results of the algorithm, the amount of money lost and the EnPI inform about the opportunities for the organisation. In case of holiday anomalies, the installation responsible for this anomaly is an installation of significant energy usage.

As a next step the roadmap of energy management is visualised. Based on this roadmap it is possible to select the different sections in which the holiday algorithm fits.

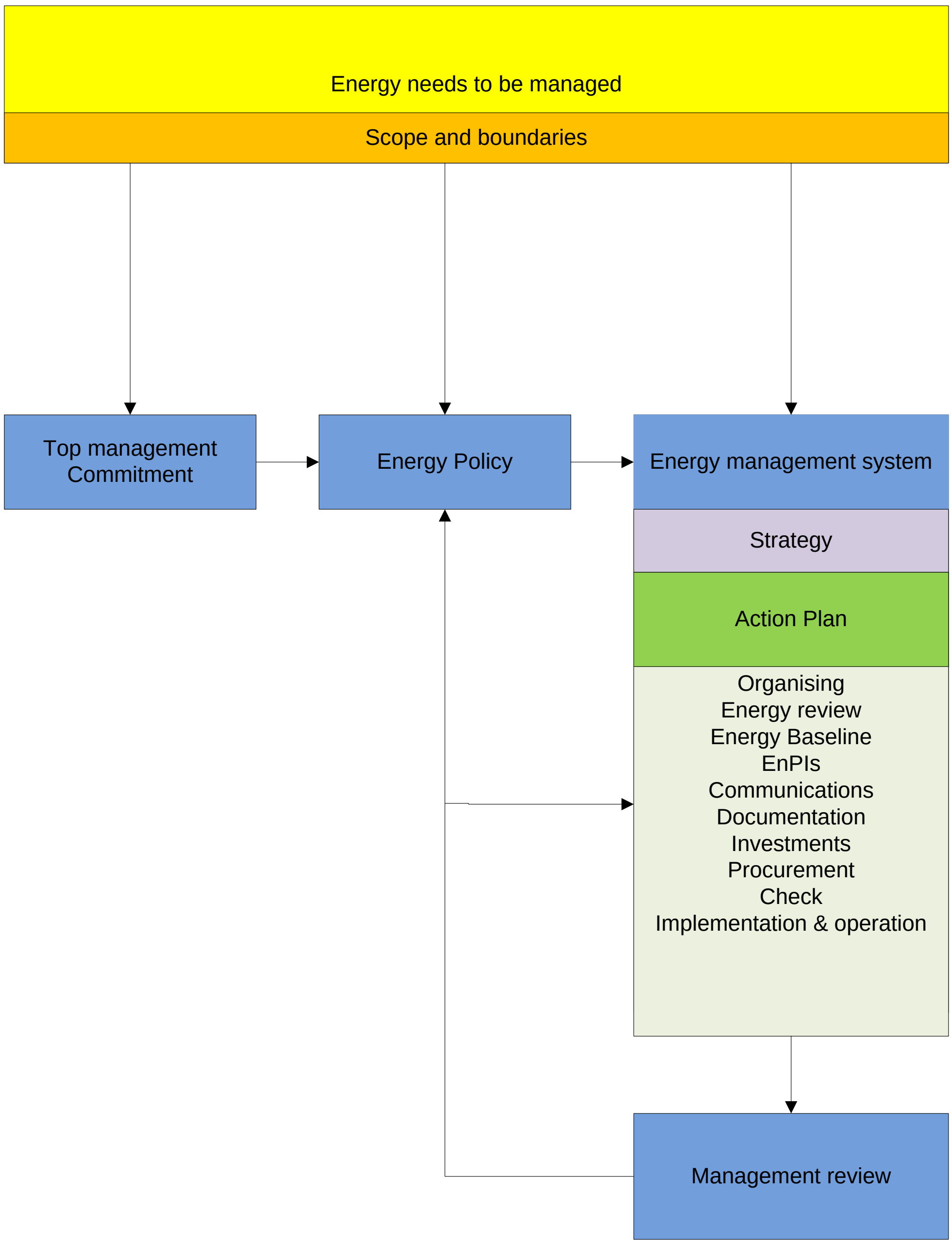


Fig. 44: roadmap for energy management

The algorithm is based on the analysis of data, therefore it does not fit within these sections:

- commitment,
- policy,
- strategy,
- management review.

The interesting sections for the algorithm are situated within the action plan. The sections which can be linked with the provided information are:

- energy review,
- energy baseline.

The holiday anomaly algorithm is linked to the energy review and energy baseline. The energy review discusses items as:

- analyse energy use based on data,
- identify areas of significant energy use,
- identify opportunities for improving energy performance.

Furthermore, the results of the algorithm could be used to set a baseline for the organisation. For instance, a building had 10 electricity anomalies in 2012 and 14 in 2013. Based on this information it is possible to assume that the baseline for holiday anomalies is 12 days. The organisation uses this kind of information to determine whether there is a problem on holidays or not.

7 Implementation techniques according to ISO 50001

The introduction explained that 2 control principles are investigated to find the best implementation method. The target to investigate 2 principles is not bad, because there are 3 principles to control something.

- automatically,
- manually,
- semi-automatically.

The investigated principles are linked to automatically and semi-automatically control. An automatic control system does not need information from an external system/interface. They have the intelligence to take decisions as needed. An example of this method is the real-time building occupancy algorithm. This algorithm will detect automatically if the building is occupied or not. If the building is not occupied it provides a shutdown of the working systems. A semi-automatic control system will automatically perform its control action to the linked system. But its decision is based on provided data from an external system/interface. For instance, a valve will be closed automatically if the temperature rises above 40°C. In this case an user-interface may be used to set the temperature. So a different temperature results in a different control action. Furthermore, investigating manual control actions is not relevant, since almost every building is equipped with BMSs which are also working semi-automatically.

Since there are 2 principles investigated, the first 2 sections are each linked to one implementation technique. Section 7.1.1 describes the principle of the real-time building occupancy algorithm to control BMSs on holidays. Section 7.2 explains the use of KPIs to control a semi-automatic system. In order to keep both sections separate, the selection of the best implementation technique is done in 7.3. Section 7.4 describes the development of the selected implementation technique. While section 7.5 analyses if the control technique fits within ISO 50001.

7.1 Real-time building occupancy algorithm

As the name already explains this method is based upon an algorithm. This algorithm will also use a previously described data mining technique. Based on the selected technique of section7.1.1, the algorithm is defined as an automatic control system.

7.1.1 Data mining technique real-time building occupancy algorithm

This algorithm must have the intelligence to provide a switch off of the systems that are not necessary on holidays. In comparison to the holiday anomaly algorithm this algorithm is totally different. Since it has to detect if a building is occupied or not and then it has to take a decision to shut the system down or not. The data used by the algorithm is limited, because it will work in real-time. Therefore, there is no need for specific data mining techniques. Straightforward principles and rule-based classification are the best techniques to create this algorithm.

7.1.2 Real-time building occupancy algorithm

The previous section discussed that this algorithm will use rule-based classification to interact with the BMS. Firstly, the algorithm must automatically detect if the building is used or not. There are several parameters which provide relevant information about the occupancy of a building. Since the algorithm will shut down either a gas consumer or an electricity consumer, it is not recommended to use these parameters for the occupancy test (gas and electricity aren't used to detect occupancy). But there are still other useful parameters:

- water readings,
- RFID-readings,
- door pulses.

RFID-readings and door pulses are directly linked with people entering and leaving a building. So the information out of this kind of readings is a measurement of the building's occupancy. On the other hand, the use of water readings is also relevant. Normally, there is no water usage when the building is not occupied. Once the building is occupied the water usage rises significantly. There are only 2 problems with the use of water readings. Firstly, the water usage is not measured for each building. Secondly, water leakages must be filtered out of the data. Fig. 45 shows an example of how water data can be used for occupancy control.

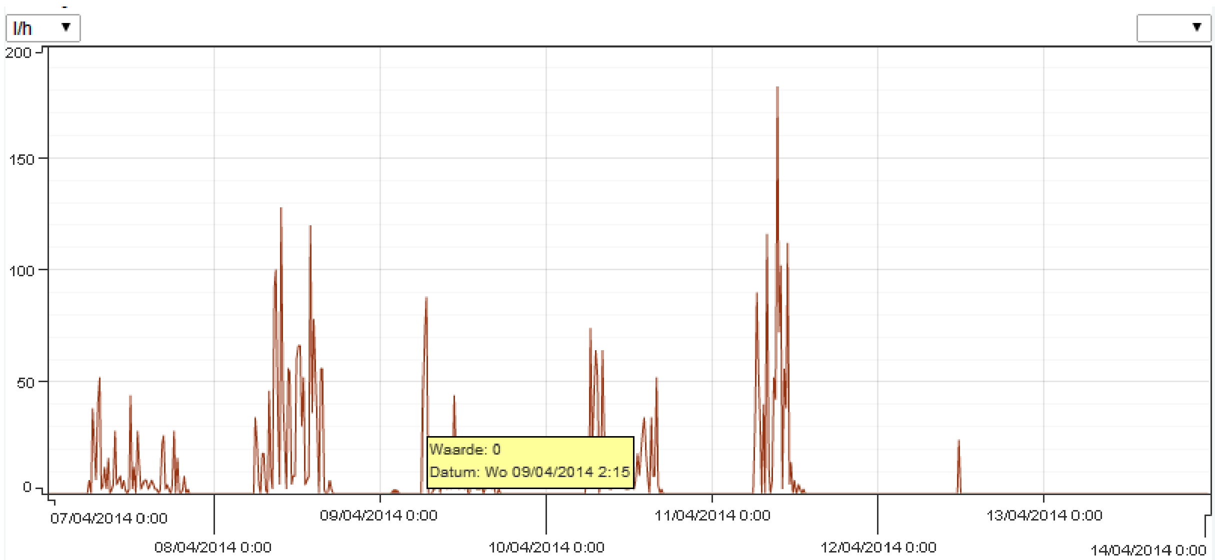


Fig. 45: example water usage

Firstly, it is clear that there are no leakages in this dataset, because here is no water consumption during the nights. Furthermore, once the building is occupied the water consumption rises significantly, as explained.

The occupancy rule is easy for door pulses, motion sensors or RFID-readings. Once a person enters a building the original reading is changed to a new value. From that moment on the building is used by a person. For instance, if the original reading has not changed until 10 o'clock, the algorithm provides a switch off action. In case of water readings, test data will be used to detect the occupancy of the building. For instance, a decision is required at 10 o'clock. The algorithm will use test data from the night. Assume that it will determine the average water usage between 3 and 5 a.m. This test value will be compared with the average water usage within the opening hours (7.30 and 9.30 a.m.). If there is a significant difference between both values, the building is occupied. Otherwise the building is not occupied and a switch off is recommended. Such an analysis is done at several intervals. The selected values (hours) of the example aren't used for real. Normally, the test hours are selected in consultation with the organisation. Fig. 46 shows the principle of this algorithm.

Implementation
• Automation of the process
Shutdown when: $\left[\frac{\sum ID}{Stop\_uur - Start\_uur}\right] < \frac{Baseline\_ID}{Stop\_uur - Start\_uur}$ With: ID = InDicator, the value which is measured (Water, RFID, pulses,..)   Start_uur = start time of the measurement   Stop_uur = stop time of the measurement   Baseline_ID = Baseline value of the InDicator on a normal day

Fig. 46: principle of real-time control

As already explained the principle is a comparison between 2 values. The left function visualises the test during opening hours. This value is compared with the baseline value. The baseline value is calculated during the closed hours. Furthermore, the selected ID will differ from building to building. But normally each building is equipped with one of the introduced indicators. The denominators of each function represent the time interval of the analysis. In case of the previous example the time interval during opening hours was from 7.30 a.m. until 9.30 a.m., while the time interval during closed hours ranges from 3 a.m. until 5 a.m.

7.2 KPI-control technique

This section discusses what is needed to tackle holiday anomalies when the KPI control method is selected. Since a KPI is a management tool, the behaviour of the KPI is observed by the persons responsible for the KPI. Therefore, each KPI is visualised. Such a visualisation may be an energy dashboard or a visualisation on a webpage. But only a KPI is not enough since this method is based on a semi-automatic control system. Therefore, the energy representative must have access to a configuration interface. The representative will use this interface to provide the BMS with information about holidays. Once the holidays are configured the platform to which the interface is linked knows when the building is closed. This platform communicates with the control system which is linked to the BMS. Once the control system has received information from the platform about the holidays, the system will force the systems to 0 (switch off) on holidays. Each of these control actions is generated at midnight and downloaded to the BMS several minutes after midnight. So the system won't start-up in the morning, since it is forced to zero in the night. If the representative does not configure the building holidays, the KPI is not reached and the representative will get a penalty.

7.3 Selection of the best implementation technique

The way both methods are linked to the BMS is the same. Therefore, the communication methodology is explained in the last chapter. There are also differences between both methods:

- Awareness,
- user interface,
- moment of control.

The KPI control technique requires an energy representative who configures the holidays into the system. Furthermore, the visualisation of the KPI and someone who is engaged to follow everything up, results in an increased energy awareness. This may result in other savings and is seen as an important parameter within ISO 50001. Moreover, the representative can configure the building holidays on a web based user interface. So a configuration is possible from everywhere where an internet connection is available. At the same time, it is possible to follow-up the KPIs or consumption profiles which is an advantage against the real-time building occupancy algorithm. The control data is downloaded each night to the control system. This means that the BMS does not start-up for a short time and this method can save more energy. Because the real-time building occupancy algorithm follows the BMS-instructions until the algorithm detects that the building is not occupied. At that moment it is possible that the building is already pre-heated, the amount of gas used for the pre-heating is wasted because the building is not occupied. Furthermore, the real-time building occupancy algorithm may work as a stand-alone application. So, it does not teach the energy team or representative something about energy awareness. Finally, Porta Capena already has a framework to visualise KPIs and to control systems with the use of calendars. Therefore, the KPI control method is recommended. The real-time building occupancy algorithm is a proposal which might be developed in the future.

7.4 Development of the implementation technique

As the previous section decided, the KPI-control technique will be developed in order to tackle holiday anomalies. The main strength of this tool is that it forces the energy representative to achieve his energy targets, KPIs. Before the visualisation is explained, it is analysed how a target for holiday anomalies is developed.

7.4.1 Build-up of a holiday anomaly target

An energy target is always defined in consultation with the organisation. Since the organisation know best how their systems work and behave. The results (KPI & EnPI) of the holiday anomaly algorithm are used as a starting point for the target definition. Although, it is also possible to use the baseline value, because the number of anomalies represents the size of the problem.

The Juran trilogy shows that feedback is given after an improvement. Although, there are still lessons to learn after the first improvement [41]. Therefore, it is not relevant to assume the first KPI as 0. It is better to give the organisation some space to learn the new implemented tools. A good proposal to the organisation is a KPI and EnPI which is 25% of the original KPI. For instance, the BEMT building had 28 holidays in 2013, 20 of these holidays were anomalies. The proposed KPI for BEMT will be: $KPI_{target} = 0.25 * KPI_{original} = 0.25 * 20 = 5$. BEMT is allowed to have 5 holiday anomalies in the first year after the improvements. The EnPI is calculated in the same way as the target KPI. In this case the EnPI will be: $EnPI_{target} = EnPI_{original} * 0.25$. The same principle is used for the upcoming years after the first improvement. The KPI target for upcoming years is calculated with the same formula. If there are 6 anomalies in 2014, the KPI for 2015 will be $KPI_{target} = 0.25 * KPI_{original} = 0.25 * 6 = 1.5 = 2$. Since the KPI is visualised in days it is necessary that the KPI_{target} is rounded, this value is always rounded towards the closest upper value. Finally, the minimum KPI value that can be used is 1. Because it is not relevant to give the organisation a penalty if they only have made one mistake.

7.4.2 Visualisation of the target

The EnPI and KPI are the 2 requirements which must be visualised on the energy screen (dashboard or website). Furthermore, it is recommended to show if there already have been holiday anomalies or not. Because it must be possible to follow-up the targets and the possible changes. A visualisation of these targets can be:

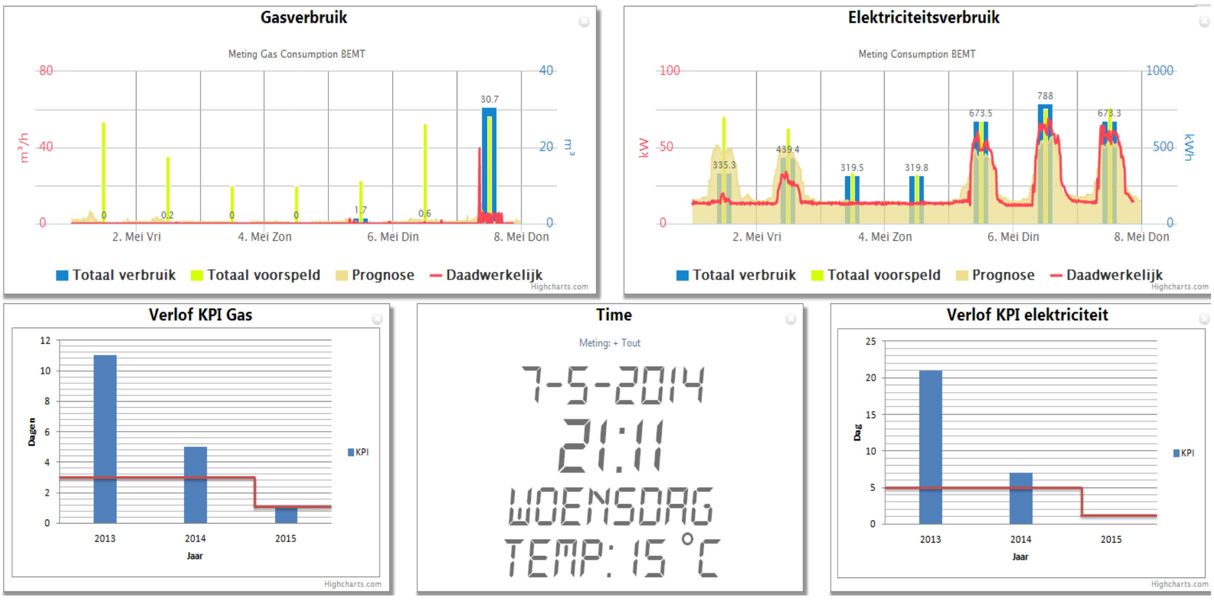


Fig. 47: target proposal

This figure is an example of what an energy screen may look like. The 2 figures above visualise the gas and electricity consumption from one week. The left and right bottom figures visualises the KPIs for gas and electricity. The holiday anomalies for each year are shown by the blue bars. The red line represents the target KPI for that year. As explained in section 7.4.1the target KPI will decrease in time, therefore the red line is not straight. Additionally to this proposal figure, it is possible to add data of the EnPI together with the EnPI target. This is possible in the same bar graph by adding a secondary axis. Otherwise, an extra figure is required.

7.5 Control actions according to ISO 50001

A general description of the ISO 50001 standard is given in the study of the basic principles. The general overview is visualised in the previous chapter. Therefore, it is not visualised again in this section. But the whole visualisation is not necessary, since the implementation of the control system and energy screen has an effect on the behaviour of systems. Therefore, the implementation technique can only fit within the action plan of the standard. Fig. 48 shows the action plan of the standard.

Fig. 48: action plan

The KPI-control plays a role in:

- operational control,
- awareness,
- evaluation of the present energy consumption on holidays.

The KPI-control technique is used to switch off installations on holidays, therefore it plays a role in the operational control of the building. Furthermore, the energy screens used to visualise the targets, makes people aware of their energy usage. Finally, each target is followed-up against the past anomalies. Therefore, the follow-up evaluates the present energy consumption. These parameters are linked with the standard in the following sections:

- check,
- implementation and operation.

The implementation & operation part of the standard discusses the operational control of different systems. Since the KPI-control technique is implemented to reduce the holiday anomalies, it is seen as a part of the operational control. The energy screen which is used to follow everything up fits within the check section, because such a screen visualises the targets and the screen makes an easy analysis of the results possible. Finally, Table 10 shows which sections (number & title) are linked to the holiday anomaly algorithm and the KPI-control method.

Table 10: tools within ISO 50001

Tools:	Holiday anomaly algorithm	KPI-control
Within ISO50001:	4.4.3 Energy review 4.4.4 Energy baseline	4.5.5 Operational control 4.6.1 Measurement, monitoring and analysis
Data sources:	Ecoscada, weather station,..	

The tools in this figure are introduced and developed within this thesis. Although, Porta Capena has more than 2 tools, which fit within ISO 50001. Therefore, the second appendix gives an overview of these tools and their role within ISO 50001.

8 Super controller for BMS

The tools to tackle the holiday anomalies are explained in the previous chapter. But it did not explain how the tools communicate with a BMS. Therefore, this chapter will introduce the BMS of BEMT. The BEMT building is used as a practical case. Furthermore, based on the BMS of BEMT several communication options are analysed and introduced. Finally, the conclusion selects the best implementation method according to the requirements.

8.1 The BMS of BEMT

The BEMT building is from 1999 and has a floor surface of 5800m² [42]. Because of this the contractor of BEMT decided to install a BMS, which has control over the HVAC system, the access to the different rooms and the lighting installation. Because of the size of the installations it was not relevant to control everything manually. Therefore, a visualisation on a desktop pc was implemented to control and follow-up the parameters of the BMS.

The installed BMS is from TAC Vista which is a product of the Schneider Electric group. In general TAC Vista consists of several products which are combined to become a total integrated system. Fig. 49 shows the topology of the TAC Vista product group.

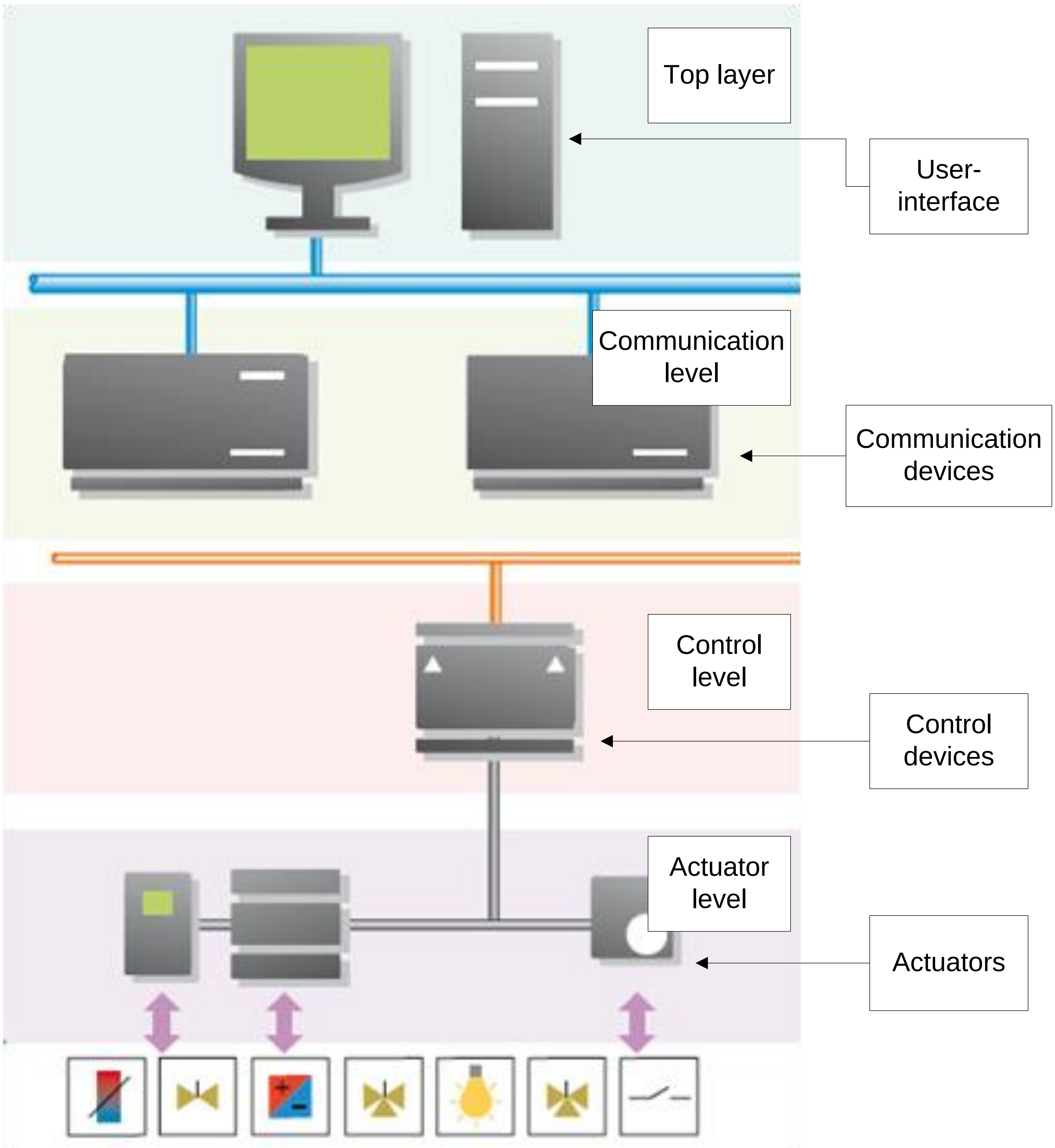


Fig. 49: TAC Vista topology
Source: TAC Vista & I/NET gebouwautomatiseringsoplossingen [43]

The top of the topology is established by the user-interface. Behind the interface runs the TAC Vista workstation software which is recommended when a desktop pc is used to control the BMS. Between the top layer and the communication devices there is a communication level which is used for remote control or for other industrial networks. This level is necessary since the TAC Vista workstation has an Ethernet based protocol and the control devices in the field communicate by LonWorks[4].

[4] LonWorks (Local Operating Networks) is a networking solution for the Building Automation. It is an open peer-peer protocol which means that it does not work with the master-slave principle. Furthermore, it has different physical communication layers (Ethernet, twisted pair, etc.)[44].

Therefore the communication devices translate one protocol (from the top layer) to the language of another protocol (LonWorks).

The control level is linked to programmable control devices. These intelligent devices are used to control actuators in the field and to analyse sensor readings. Depending on the size of the BMS it is possible to configure a network with the programmable controllers. Configuring a network on the level of control devices will increase the structure of the installation (more devices are required, the network will expand). Although, there will always be one controller that is responsible for his group and that takes the most important decisions, this device is also called a master.

The actuator level represents the sensors and actuators in the field. This level has no intelligence at all, since it consists of valves, temperature sensors, pumps, etc. [45], [43]. The TAC Vista product family has products for each level. This means that it is possible to build a BMS with only TAC Vista products.

The BMS of BEMT uses the same topology as explained in the previous paragraphs. Furthermore, the section introduced the possibility to define groups on the control level. The BMS of BEMT is an example of this approach. Fig. 50 shows the reason why such a groups are useful.

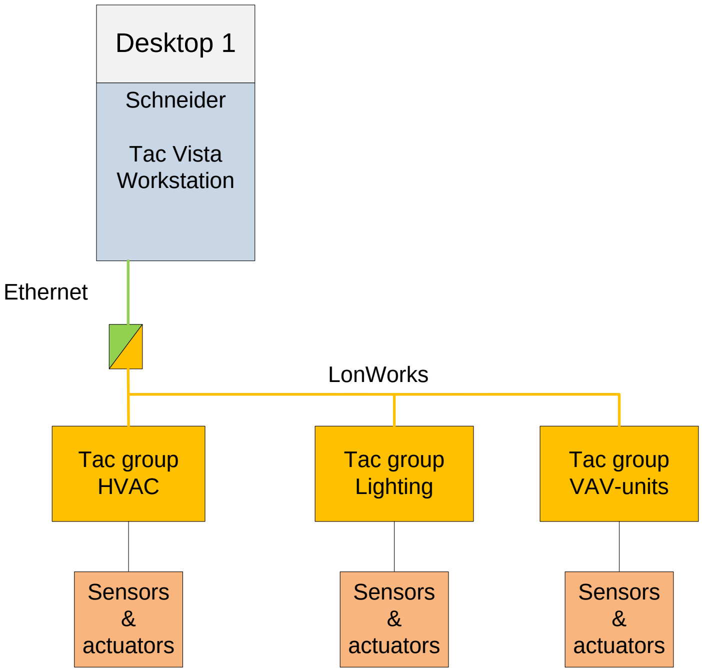


Fig. 50: BMS configuration BEMT

As the figure shows there is a TAC Vista control group for each type of installation. Furthermore, the same topology as in Fig. 49 can be derived from this scheme.

8.2 Communication principle

Before a super controller is selected, it is important to understand the communication between the implementation technique of 7.4 and the BMS. Without this explanation it would be difficult to understand that the communication principle doesn't provide any issues, since 2 systems want to control the same installation. In order to prevent this, the generated control action is the result of an 'and' relation between the BMS and Porta Capena. If the BMS doesn't send a control action to a system, it is impossible for Porta Capena to start-up an installation. On the other hand, it is possible for Porta Capena to switch off an installation when the BMS sends a control action. Furthermore, Port Capena sends a control action if the digital output is 0. This approach results in a failure proof solution, because if the system of Porta Capena doesn't react the digital output is 0. Therefore, this principle is the most appropriate to communicate with different installations.

8.3 Super controller options

The visualisation of the KPI on energy screens is established by Porta Capena. But Porta Capena cannot directly control the BMS from their web service. Therefore, this section introduces several options to establish a communication between Porta Capena and the BMS. The main target of the implementation is to switch off systems when possible without losing comfort. This kind of approach will save energy and money. As explained within the KPI-control technique, holidays are programmed by using a calendar. Based on this calendar rules are made to control the installations by the BMS. This means that information of the BMS, is not necessary. Therefore, one way communication (from Porta Capena to the BMS) is enough to switch off the desired installation. In case of the real-time building occupancy algorithm of 7.1, in 2- way communication is required.

Section 8.1 discussed the installed BMS of BEMT and its topology. The bottom layer (sensor and actuator level) of the topology is not applicable for communication. Because this level does not have any kind of intelligence. Thereby, there are only 3 layers applicable for the communication.

8.3.1 Top layer

Communication between Porta Capena and the BMS at this level is established by the use of OPC [46]. OPC is the standard for the secure and reliable exchange of data in industry. Most suppliers of industrial data acquisition or control devices (PLCs, BMS) are designed to work with the OPC standard. The BMS from TAC has also the opportunity to work with OPC. The advantage of this standard is that an abstraction layer (COM/DCOM[5]) is used to pass data between devices without being aware of each other's internal data language [48]. In practice this method uses the server-client principle. The server, Porta Capena guarantees the communication with a device, BMS, while the client, BMS, takes care of the communication with the

[5] This Microsoft Windows product enables software components to communicate in a variety of programming languages. Furthermore, DCOM is an extension of COM [47], [48].

application. OPC only defines the link between the server and client but does not interact with the Server-to-device and Client-to-application communications [49], since these are different for each installation. This method can be used on one device but it is more common that it is a communication between 2 devices, since the client and server are not always installed on the same device. Fig. 51 shows how it can be implemented for BEMT.

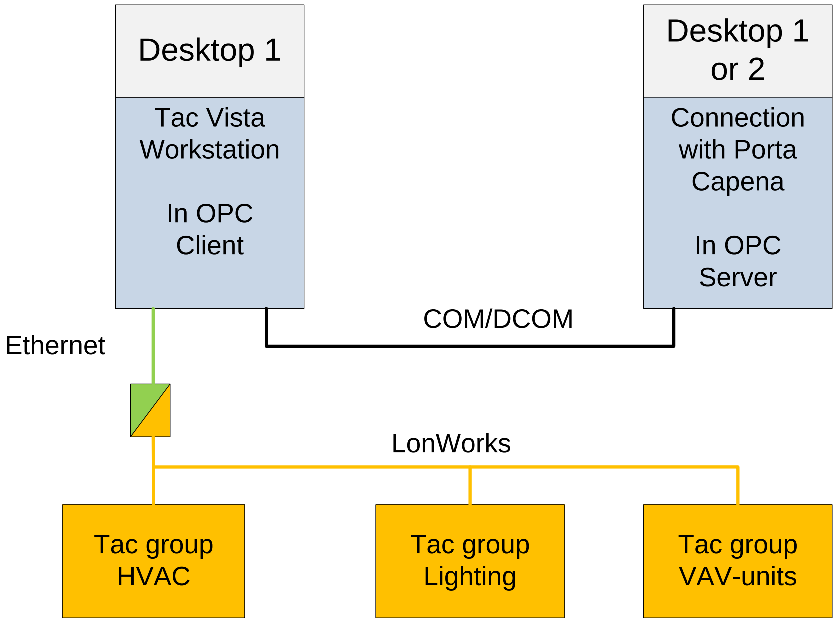


Fig. 51: OPC for BEMT

In comparison with the introduction of this section OPC has more features than needed. With the implementation of OPC it is possible to use, write and read commands. So it also possible to get data from the BMS.

As explained OPC is a standard which stands alone. Therefore, it is sold as a separate product in the form of licenses. The price of a license depends on the vendor name of the installation. License prices between €500 and €2000 are possible. This is only the license price so the price to program everything is additional.

8.3.2 Communication level

As explained in section 8.1the communication level makes it possible that different industrial networks communicate with each other. In case of BEMT this layer already converts the Ethernet protocol of Schneider into LonWorks. But the layer provides also a framework for other industrial networks. The only condition is that a network converting device is implemented, a proxy. The TAC Vista product family consists of several proxies. But there are also companies who are specialised in proxies, Anybus is one of these companies. Basically there are 2 requirements with this method to establish a communication link. Firstly, an industrial network must be chosen where Porta Capena can communicate with. Secondly, a proxy must be available which can translate the chosen language into LonWorks. Porta Capena uses controlling devices which have the possibility to communicate over Modbus.

This industrial network has several communication types but Porta Capena can communicate with the 2 most common types, Modbus RS 485(RTU) and Modbus TCP/IP. Furthermore, Modbus is situated at the application layer of the OSI model which means that the communication is established by client/server [50]. This principle is based on requests (issued by the client) and responses (issued by the server) [51]. Finally, the different communication types and the easy to use client/server principle are 2 reasons why Modbus is frequently used in industry.

The next step is to find a proxy which can translate the Modbus protocol into LonWorks. Within the product family of TAC Vista there are 3 communication modules available.

- webserver Xenta 511-b,
- programmable webserver Xenta 731;
- gateway Xenta 913.

All these devices have the functionality to connect a Modbus network. Since only the proxy is needed and not the web functionality, the best choice is the Xenta 511-b. This product will cost around €2000, while the cost for programming the connection and establishing the data transfer is additional. Anybus has also a proxy which can work with Modbus and LonWorks, the Anybus communicator LonWorks: LonWorks <> Modbus RTU (product name) [52]. This device works only if it is configured. Therefore, a pc-based configuration tool is necessary. This links data between the devices installed on a LonWorks net and devices installed on a Modbus network. This pc-based configuration software is an extra cost to the package. The cost price of the communicator (proxy) is €765 and the price of the software is €109, this results in a total price of €874. The programming of the data transfer is also an additional cost.

This proxy has more features than necessary for the application. The communicator has the possibility to transfer 256 input signals and to transfer 256 output signals. As explained before, the output signals from Porta Capena to the BMS are necessary, but there is no need for 256 signals. Because the installations that are responsible for holiday anomalies are limited, definitely not more than 10. Finally, Fig. 52 shows the principle of the implementation on BEMT.

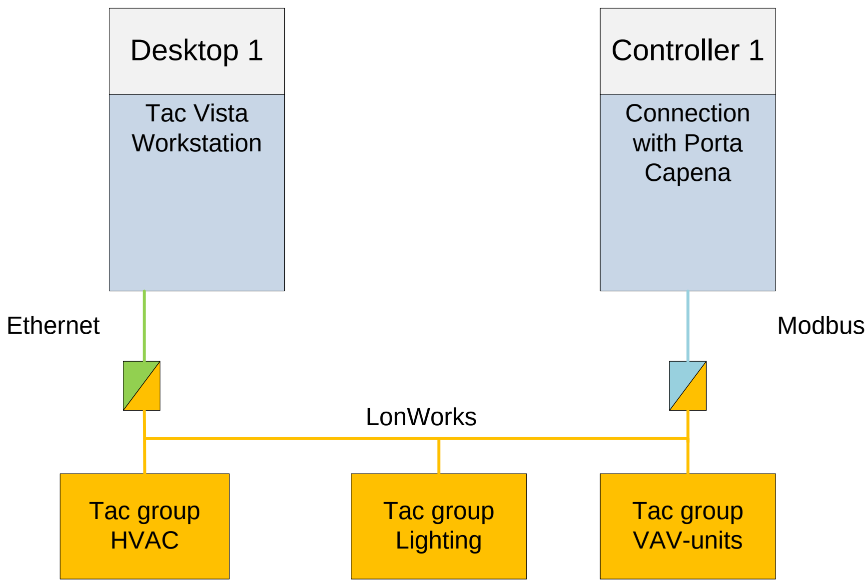


Fig. 52: communication via proxy

8.3.3 Control level

This layer is situated at the lowest level within the BMS topology where communication is possible. Therefore, the provided solution is not complicated and returns to the basics of communication. Because the communication is established by the transfer of bits. Since a programmable controller is made for communication with actuators and sensors, it can deal with digital and analogue signals. In comparison to analogue signals, digital inputs/outputs are more recommended. For instance: a ventilation system is still working on a holiday but the building is not occupied. So it is possible to shut down this system. But there is no need to reduce the ventilation capacity to 50% for example.

Communication between devices is possible by transferring digital signals. Each digital signal then has a specific meaning. As in the other examples it is not a problem to communicate with devices which are connected to another network. Because the intelligence of the signals is configured in the software. Thereby, the signals are meaningless for the hardware controllers but not for the software. So the controller will think that it is communicating with a sensor or actuator. In this case Porta Capena works with outputs and provides signals to the outputs while the BMS needs an input device to analyse the signals from Porta Capena. Therefore, Porta Capena can use their standard controlling device. This is also used when the HVAC algorithm is implemented (extra information see appendix). As already highlighted the BMS needs an extra input device. The number of inputs or outputs depends on the number of control actions that are necessary. In this thesis the holiday anomaly algorithm is implemented together with the HVAC start-up algorithm. Furthermore, a segmentation of the HVAC circuits was necessary which resulted in 10 signals. So, Schneider recommended a Xenta 412 input device with 10 inputs. In practice the installation looks like Fig. 53.

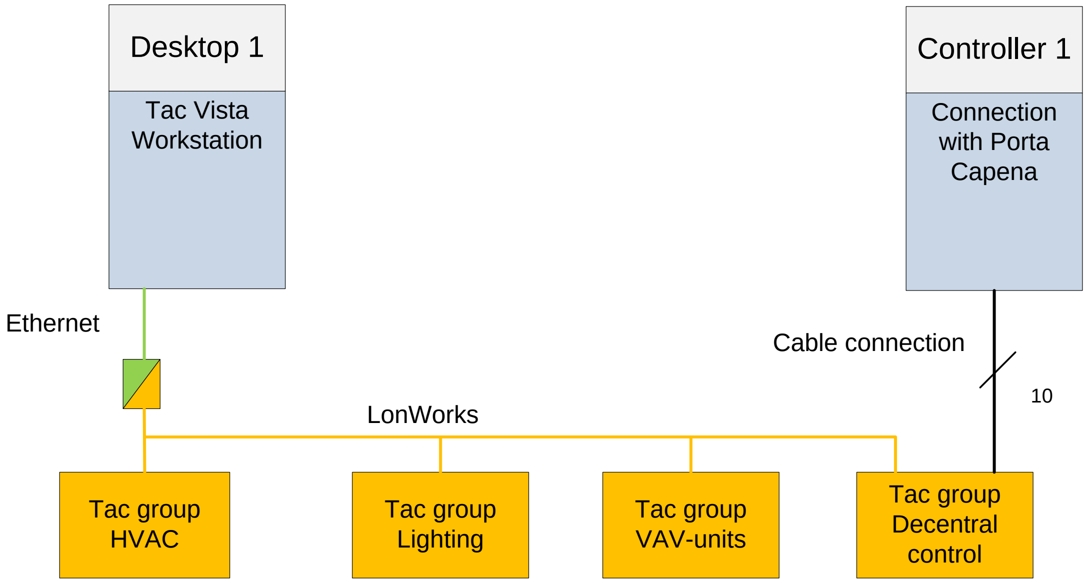


Fig. 53: communication with digital signals

This method fulfils the requirement (sending signals from Porta Capena to the BMS) defined in the introduction of this chapter. Although, it does not have extra features like the other proposals. Therefore, the hardware module (Xenta 412) is not expensive, it costs €455. Like the other proposals the programming of the communication signals into each device is an additional cost.

8.4 Conclusion

The introduction of this thesis highlighted several requirements for the communicator between Porta Capena and the BMS. Therefore, the requirements are repeated.

- sending signals in one direction,
- robust solution,
- failure proof installation,
- high quality/price ratio.

The first requirement was a prerequisite for the proposed solutions, so this is fulfilled by the 3 methods. Although, the communication method with the digital signals does not have extra features like the 2 others.

Secondly, it must be a robust solution which minimizes the communication problems and which does not crash without any reason. Based on experience Porta Capena knows that OPC sometimes has problems with decimal signals. Even a proxy must be programmed but this will be easier than with OPC. Although it is possible that errors are generated. The digital signal technique will not have this kind of issues. This method does not need a third device to translate the communication signals, which makes this method more robust.

Finally, it is possible to compare the price of each method with the provided quality. The level of quality depends on how good each method fulfils the requirements. Therefore, Table 11 shows the price of each method.

Table 11: prices of communication methods

	Super Controller for BMS					
	Top layer		Proxy		Digital signals	
price	500 à 2000	euro	874	euro	455	euro
	Programming is additional					

Each of these prices is based on a BMS of Schneider. If another BMS is used the prices may change. The communication method on the lowest layer (control level) of the topology is the cheapest and fulfils all requirements. Therefore, this method is the most appropriate for communication in one direction.

A problem may occur when communication in 2 directions (real-time building occupancy algorithm) is required. The selected super controller is established by easy to use programmable controllers. Gathering other information is then only possible by implementing another programmable device, which also needs additional programming steps. On the other hand, the 2 other methods do not need an extra device but just some extra programming steps to configure what kind of data is needed.

9 Conclusion

Two tools have been developed within this master's thesis. The first tool detects holiday anomalies automatically while the second tool provides a solution for the end-user. The principle to use these tools is as follows: if a public building owner detects a problem with the energy consumption, he can ask Porta Capena for an energy data scan, the holiday anomaly algorithm is one of the tools within the data scan. If the energy data scan discovers holiday anomalies and the customer wants to solve this issue, it is recommended to use the second tool, the implementation technique. This means that it is possible to use the holiday anomaly algorithm separately from the implementation technique. But it is not relevant to use the implementation technique separately from the holiday anomaly algorithm. Because, it is only acceptable to install this technique if there are holiday anomalies.

9.1 Holiday anomaly algorithm

Fig. 54 shows a summary of the algorithm. Based upon this summary, it is possible to discuss the most important parts of the algorithm.

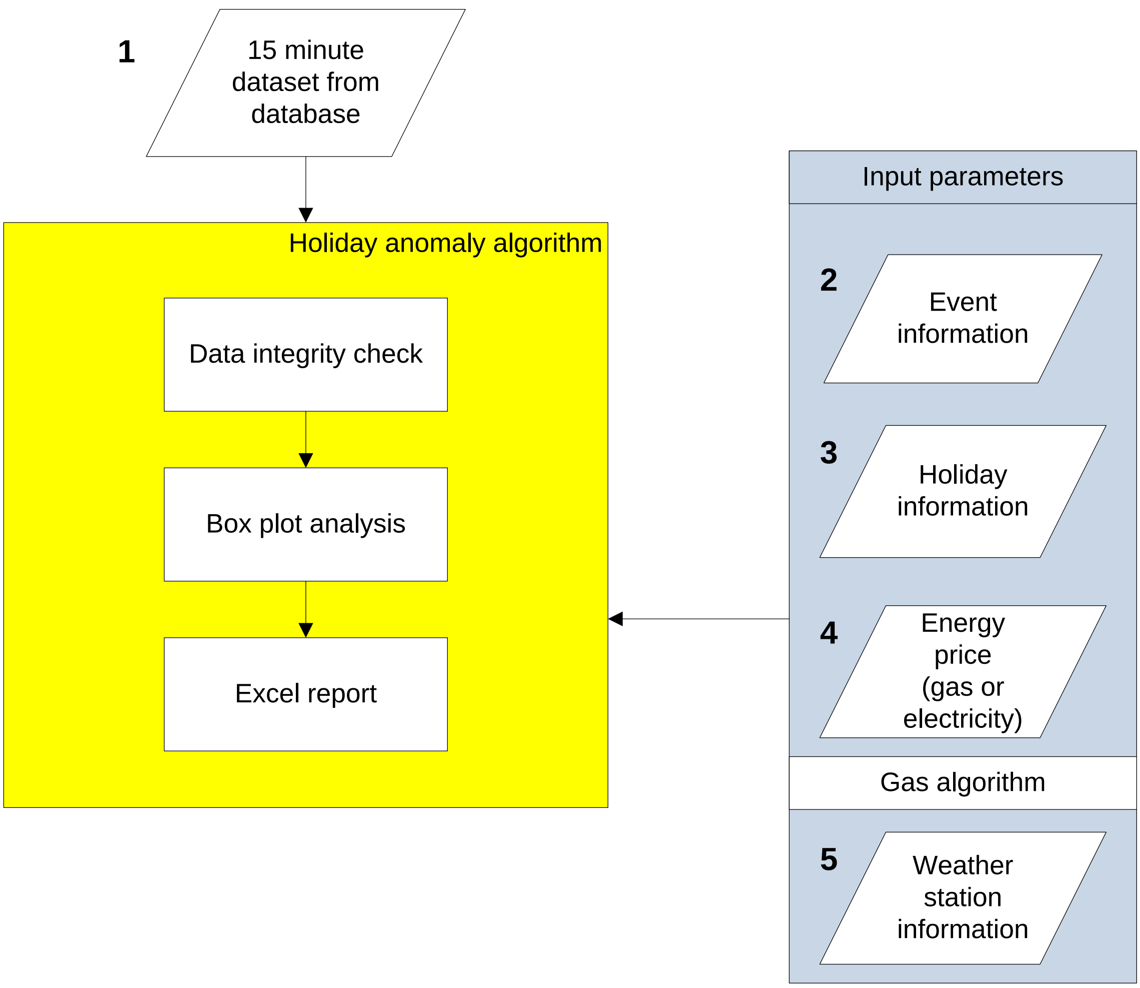


Fig. 54: algorithm summary

The user must provide a dataset (1) with consumption values on a 15 minute time base. Such a dataset consists of a timestamp and the consumption value in kW for electricity and in m^3/h for gas. If the user can provide a dataset in the required format, he can use the algorithm without any help of Porta Capena. Otherwise, it is recommended that Porta Capena converts this data into the correct format in their database. If the data is available on Porta Capena's database, it is possible to use EcoSCADA or the data collector to run the algorithm. It is recommended to use the data collector to extract data because it consists of several pre-processing steps. Although for a quick analysis one can better use EcoSCADA or its own data, because of the faster data extraction. There is not a large difference between that has been pre-processed data or has not been pre-processed, since the algorithm directly checks the data integrity.

Other parameter-sets for an holiday anomaly analysis are configured in the application. It is possible to add a new building in the application. The user defines the events(2) and holidays (3) for this building. The algorithm uses the name of the building to integrate extra data (2) & (3) for its analysis. Furthermore, it is possible to set the energy price (4) and to select a weather station close to the building (5) in case of a gas analysis.

The box plot is selected as the most appropriate data mining technique to detect holiday anomalies. Because it is independent from the data distribution, it is possible that there are no anomalies and the technique visualises the outlier boundary. The other analysed techniques couldn't fulfil all these requirements. Although, a test with a cluster analysis on the electricity data of BEMT has shown that the results are almost the same as with the box plot analysis.

Finally, this application is available on a standard web browser and can be used if the user has a license (login and password). This means that everyone with an internet connection can use the application and that there are no constraints for the installed operating system. Therefore, the report is generated as an Excel file, since almost everyone has Excel installed on his or her computer.

9.2 Implementation technique and controller

This thesis investigated 2 implementation techniques to tackle the holiday anomaly issue, an automatic control system and a semi-automatic control system. The automatic control system does not need any kind of user input, based upon rules based classification it will compare test data which is measured during the night against measured data during opening hours. The comparison result indicates if the building is occupied or not. If the building is not occupied, the control system will shut down energy consuming systems which are not necessary at that moment. Since this technique decides during opening hours if a switch off is required, it does not save all the lost energy. Furthermore, the system does not make people aware of their energy usage. Therefore, it is better to use the semi-automatic control system. This system uses a web based user-interface to define the building holidays, events and closed days. This means that the energy representative is responsible for this kind of configurations. Due to the configuration of the holidays, the control system knows when there is a holiday and provides a switch off signal during the night. So any kind of lost energy consumption is avoided. Furthermore, the user-interface is combined with an energy dashboard. This dashboard is used to visualise the targets (KPIs and EnPIs) for the representative and to make employees aware of their energy usage.

Besides the described implementation technique, a controller is needed. This controller will communicate with the installed BMS, because the BMS controls all the installed systems. The controller will provide a signal to the BMS when a switch off of a system is required. Several controllers are analysed in chapter 8, each of these solutions can be installed. But the selected solution depends on the project. Therefore, it is also possible that other implementation techniques are more appropriate than the provided solutions in this thesis. If only the holiday anomaly algorithm and the HVAC algorithm are installed, I/O communication between the controller and the BMS is the most robust and easy to use solution. Finally, it is recommended to install the HVAC algorithm in combination with the holiday anomaly algorithm, because the risk of a frozen installation is unacceptable against the amount of money saved by shutting down the heating system. Therefore, the intelligence of the HVAC algorithm is used to avoid such situations.

Bibliography

[1] "What is the EU doing about climate change? - European Commission." [Online]. Available: http://ec.europa.eu/clima/policies/brief/eu/index_en.htm. [Accessed: 03-Mar-2014].
[2] "EMAS and ENERGY management." European Commission, Jul-2013.
[3] "Energy management systems - Requirements with guidance for use (ISO/FDIS 50001:2011)." 12-Jan-2011.
[4] "An introduction to energy management." Carbon Trust.
[5] "Energy management." Carb.
[6] V. Vesma, "Energy Management principles and practice." BSI.
[7] "Een geleidelijke opwarming | climatechange.be." [Online]. Available: http://www.klimaat.be/spip.php?article176. [Accessed: 17-Jul-2013].
[8] "The Causes of Global Warming." [Online]. Available: http://www.climatehotmap.org/about/global-warming-causes.html. [Accessed: 17-Jul-2013].
[9] "Europe 2020 – EU-wide headline targets for economic growth - European Commission." [Online]. Available: http://ec.europa.eu/europe2020/targets/eu-targets/index_en.htm. [Accessed: 17-Jul-2013].
[10] "Energie trends 2012." ECN,Energie-Nederland en Netbeheer Nederland.
[11] G. Pepermans, "De distributie van elektriciteit in Vlaanderen." KU Leuven, oktober-2002.
[12] "Emissiehandel - Milieuloket." [Online]. Available: http://www.milieuloket.nl/9353000/1/j9vvhurbs7rzkq9/vhurdyxqxfmg. [Accessed: 19-Jul-2013].
[13] "DOE eGuide for ISO 50001." [Online]. Available: https://ecenter.ee.doe.gov/EM/SPM/Pages/home.aspx. [Accessed: 06-Aug-2013].
[14] "Creating an awareness campaign." Carbon Trust, Dec-2013.
[15] "Creating an energy awareness program." U.S. Department of Energy, Energy Efficiency and Renewable Energy.
[16] "Dimming CFLs and LEDs." [Online]. Available: http://www.lutron.com/en-US/Education-Training/Pages/LCE/DimmingCFLsandLEDs.aspx. [Accessed: 24-Jul-2013].
[17] "GROHE - Waterbesparende kranen - Waterbesparende producten - Voor je badkamer." [Online]. Available: http://www.grohe.com/be-nl/3916/badkamer/waterbesparende-producten/waterbesparende-kranen/. [Accessed: 24-Jul-2013].
[18] J., Micheline, Jian Han, Kamber, Pei, *Data Mining Concepts and Techniques*, 3rd ed. Morgan Kaufmann.
[19] "What is transactional data? - Definition from WhatIs.com." [Online]. Available: http://whatis.techtarget.com/definition/transactional-data. [Accessed: 20-May-2014].
[20] T. H., Charles E., Ronald L., Clifford Cormen, Leiserson, Rivest, Stein, *Introduction to Algorithms*, Second. .
[21] P., Froukje, Atze Jacobs, van Dijken, Boerstra, "Prestatie-eisen ventilatie in klaslokalen," p. 4, Aug-2007.
[22] "Degree Days - An Introduction." [Online]. Available: http://www.degreedays.net/introduction. [Accessed: 18-Aug-2013].
[23] "Degree Days - Handle with Care!" [Online]. Available: http://www.energylens.com/articles/degree-days. [Accessed: 18-Aug-2013].
[24] "Building Management Systems." Australian Government, Oct-2010.

[25] "Energiebesparing met gebouwbeheersystemen (GBS)." Agentschap NL, 25-May-2010.
[26] A. Al-Hammad, "Building Management System BMS." College of Environmental Design.
[27] V., Andy, Jing, Michael Masayna, Koronios, Gao, Gendron, "Data quality and KPIs: a link to be established." .
[28] "Guide to key performance indicators." PriceWaterhouseCoopers, 2007.
[29] Porta Capena, "Ecoscada." .
[30] "Eandis - Klant - Meteropneming - Organisatie meteropneming." [Online]. Available: http://www.eandis.be/eandis/klant/k_organisatie_meteropneming_P.htm. [Accessed: 20-Mar-2014].
[31] "Tijdzone - Wikipedia." [Online]. Available: http://nl.wikipedia.org/wiki/Tijdzone. [Accessed: 25-Mar-2014].
[32] "Zomertijd - Wikipedia." [Online]. Available: http://nl.wikipedia.org/wiki/Zomertijd. [Accessed: 25-Mar-2014].
[33] "Math Planet - Standard deviation and normal distribution." [Online]. Available: http://www.mathplanet.com/education/algebra-2/quadratic-functions-and-inequalities/standard-deviation-and-normal-distribution. [Accessed: 03-Apr-2014].
[34] "Statit Support: Normal vs. Non-Normal Distributed Data�Comparing Results." [Online]. Available: http://www.statit.com/support/quality_practice_tips/normalvsnon-normal.shtml. [Accessed: 03-Apr-2014].
[35] "Normal distribution - Wikipedia, the free encyclopedia." [Online]. Available: http://en.wikipedia.org/wiki/Normal_distribution. [Accessed: 24-May-2014].
[36] "Statistics: Power from Data! Box and whisker plots." [Online]. Available: http://www.statcan.gc.ca/edu/power-pouvoir/ch12/5214889-eng.htm. [Accessed: 11-Apr-2014].
[37] "K-Means Introduction." [Online]. Available: http://www.hypertextbookshop.com/dataminingbook/public_version/contents/chapters/chapter004/section002/green/page001.html. [Accessed: 12-Apr-2014].
[38] "Clustering - K-means." [Online]. Available: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html. [Accessed: 12-Apr-2014].
[39] "Euclidean distance - Wikipedia, the free encyclopedia." [Online]. Available: http://en.wikipedia.org/wiki/Euclidean_distance. [Accessed: 12-Apr-2014].
[40] "Classification and Regression Trees." stat.cmu.edu, 06-Nov-2009.
[41] "MSI." [Online]. Available: http://msi6.com/MSI6/QualityZone/QzoneJuranTrilogy.aspx. [Accessed: 20-Apr-2014].
[42] "BEMT." [Online]. Available: http://www.bemt.be/bemt.htm. [Accessed: 22-Apr-2014].
[43] "TAC Vista & I/Net Gebouwautomatiseringsoplossingen." Schneider Electric.
[44] "LonWorks Protocol Overview." [Online]. Available: http://www.rtaautomation.com/lonworks/. [Accessed: 22-Apr-2014].
[45] "TAC Vista - Integration, intelligence and TCP/IP deliver a smarter Building Energy Management System - Products overview - Schneider Electric." [Online]. Available: http://www.schneider-electric.com/products/ww/en/1200-building-management-system/1210-building-management-systems/2312-tac-vista/. [Accessed: 22-Apr-2014].
[46] "What is OPC? | OPC Foundation." [Online]. Available: https://opcfoundation.org/about/what-is-opc/. [Accessed: 22-Apr-2014].

[47]"COM: Component Object Model Technologies." [Online]. Available: https://www.microsoft.com/com/default.mspx. [Accessed: 22-Apr-2014].
[48]"Introduction to OPC - National Instruments." [Online]. Available: http://www.ni.com/white-paper/7451/en/. [Accessed: 22-Apr-2014].
[49]"Introduction to OPC for Building Automation." [Online]. Available: http://www.automation.com/library/articles-white-papers/building-automation/introduction-to-opc-for-building-automation. [Accessed: 22-Apr-2014].
[50]"Modbus Application Protocol." Modbus organisation.
[51]"Understanding the Modbus Protocol." [Online]. Available: http://jamod.sourceforge.net/kbase/protocol.html. [Accessed: 23-Apr-2014].
[52]"LonWorks to Modbus RTU Serial Gateway - Anybus Communicator Specific." [Online]. Available: http://www.hms.se/products/products.asp?PID=405&ProductType=Anybus%20Communicator. [Accessed: 23-Apr-2014].

Annex list

I. Closed day anomaly algorithm

This algorithm is introduced in the holiday anomaly algorithm according to the ISO 50001 chapter. The holiday anomaly algorithm uses this algorithm to exclude outliers in the reference data (closed day data), since this action will result in a higher quality of the algorithm. Furthermore, this algorithm is also used as a standalone algorithm to detect outliers on closed days.

In comparison to the holiday anomaly algorithm, this algorithm only uses data of closed days. Therefore, the box plot data mining technique is used to find these outliers. In comparison to the holiday anomaly algorithm this detection technique is less complex. Because only the five number summary is used to detect outliers. Fig. 55 shows the functioning of the algorithm.

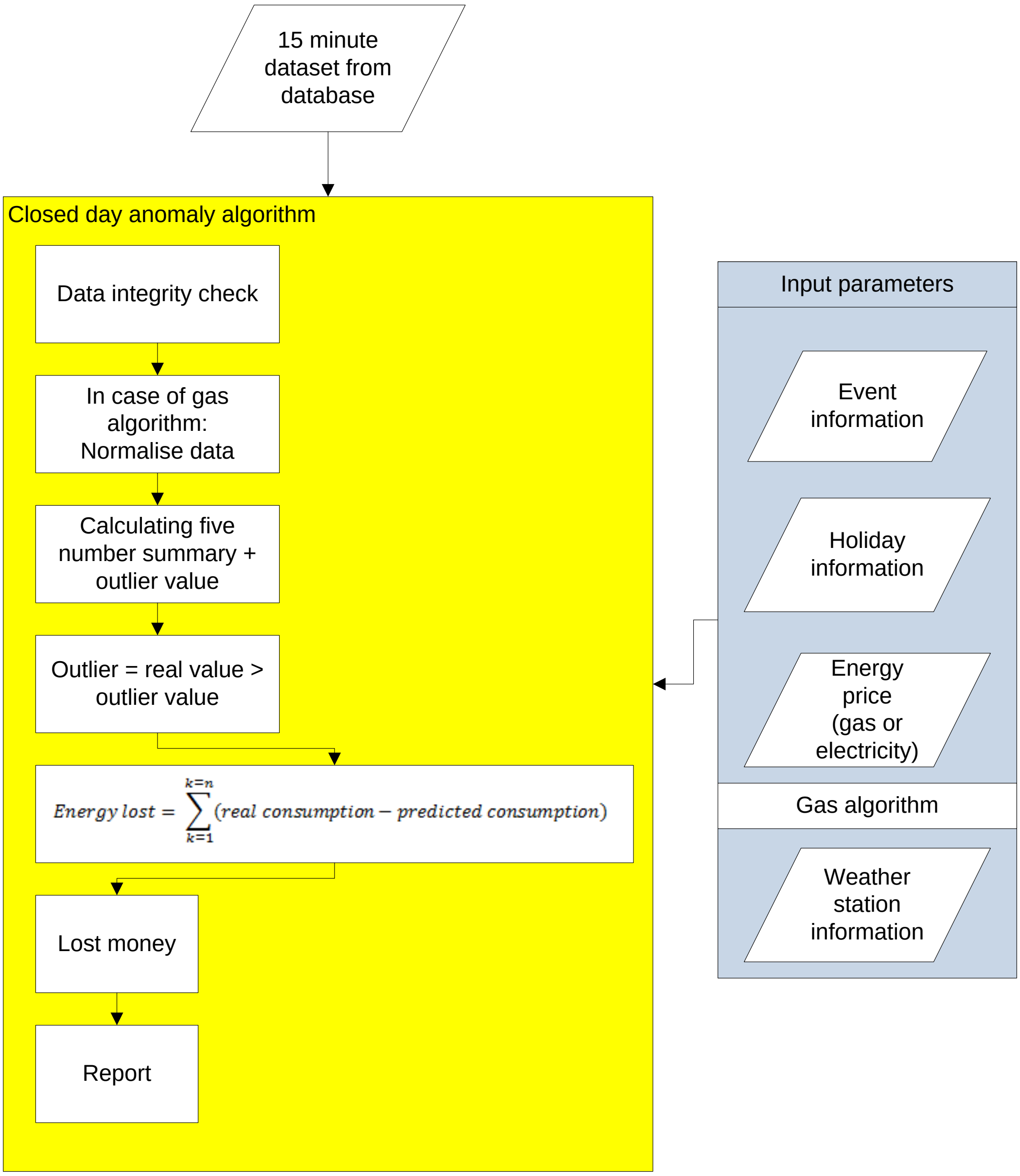


Fig. 55: holiday anomaly algorithm

Several algorithm blocks are the same as in the holiday anomaly algorithm. Data integrity is always a critical parameter within data mining. Therefore, this is always the first pre-processing step. Furthermore, the provided gas data is also normalised by DD. Once the data is pre-processed, it is possible to calculate the outlier value. This value is based on the result of the five number summary, $outlier\ value = Q3 + (Q3 - Q1) * 1{,}5$. Each value that is higher than the outlier value is categorised as an outlier. These outliers are used to calculate the opportunities, the amount of energy lost and money lost. Finally, all the results are visualised in a report. This report is similar to the report of the holiday anomaly report. Firstly, there is a table with all the detailed information of each day. Secondly, there is a bar graph which visualises the values of the table. Because this algorithm focuses on closed days and calculates the opportunities, extra information is necessary. This information is shown in the blue rectangle. Normally, the event information and holiday information is integrated in the algorithm if the building name is specified. Furthermore, the energy price is used to calculate the amount of money which is lost and the information of weather stations is used by the gas algorithm to normalise the consumption values.

The holiday anomaly algorithm does not use the whole closed day algorithm. The holiday algorithm must only know which closed days are outliers. Because these days must be excluded from the reference dataset. Therefore, the first 4 blocks are implemented in the holiday algorithm. The other blocks are only relevant if the algorithm is used as a standalone application.

II. Tools for ISO 50001

Within this master's thesis there are 2 tools introduced which fit within the standard ISO 50001. But Porta Capena has more than 2 tools which are applicable within ISO5001, all these tools are introduced in this chapter.

Tools:	Energy data scan	Flexible Interactive Energy Screens	HVAC control	Energy savings scan	Energy alarms	Data integrity	Ecoscada reporting
Real time:	●	●	●		●	●	
Periodic:	●			●	●	●	●
Within ISO50001:	4.4.3 Energy review 4.4.4 Energy baseline	4.5.2 Awareness	4.5.5 Operational control	4.6.1 Monitoring, measurement and analysis	4.6.4 Nonconformities	4.6.5 Control of records	4.7.2 Input to management review
Data sources:	Ecoscada, weather stations, Indexis,...						

Fig. 56: Tools within ISO 50001

The energy data scan:
The energy data scan consists of a set of tools which are developed to analyse the data of public buildings in a couple of minutes. The objective of these tools is to detect if the building has a specific energy problem or not. The data scan provides at this moment the next tools for electricity and gas:

- holiday anomaly algorithm;
- closed day anomaly algorithm;
- penalty analysis (electricity);
- HVAC start point analysis (simulation) (gas);
- temperature KPI-check (general).

The holiday and closed day anomaly algorithm are already explained in the previous sections. The other tools will be explained now. The electricity penalty analysis focuses on the contract with the electricity broker. Within almost each contract there is a maximum capacity value on 15 minute basis defined. If a building consumes more than x % of the defined value, the electricity price raises with for instance 1 €cent/kWh for the consumption of that month. An example will make this principle clear. BEMT has a defined value of 98 kWh. If they consume 20% more than 98 kWh, they receive a penalty. Assume that BEMT consumed in March 130 kWh during 15 minutes at one time. This value is higher than the penalty boundary of 117.6 kWh. So BEMT receives for that month a penalty of 1€cent/kWh. If there normal electricity price is 10 €cent/kWh, they should now pay 11€cent/kWh due to the penalty. This means if BEMT consumed 1000kWh in March that they lost 1000*0.11-1000*0.1 = €100. The algorithm is developed to detect this kind of

problems. Furthermore, it defines how much money the organisation lost due to the penalty.
The HVAC start point analysis defines the optimal starting point for a HVAC-system. Therefore, the algorithm uses several input parameters as pre-heat time, opening hours, set temperatures, etc. This algorithm is used to define the savings opportunities for the heating system. Finally, it is possible to verify the comfort parameter temperature of a specific room. This algorithm checks if the defined KPIs have been reached or not. For instance, it is defined that a room has a comfort temperature between 20°C and 23°C. The KPI for this room is as follows: less than 3 Kh(Kelvin hour) beneath 20°C and less than 5 Kh(Kelvin hour) above 23°C. Based on this KPI it is possible to check if complaints from employees are relevant or not.

As already explained with the holiday anomaly algorithm, this kind of tools are used to define the possible opportunities of a building in the future. Furthermore, the results will identify installations of significant energy use and the output results can be used for the definition of the baseline.

Flexible Interactive Energy Screens:

The action plan of the ISO 50001 standard is one of the important sections towards an EnMS. Within this section, it is determined which installations consume more energy than needed. Once the new implementations are determined by the organisation, employees must be trained and must be aware of their energy consumption, this is described in section 4.5.2 of the standard. As the energy management section of the study of the basic principles explained, employees might lose their energy awareness after a while. Therefore, Porta Capena provides several Flexible Interactive Energy Screens. These screens makes it possible for the customer to visualise any kind of data which is measured. Furthermore, the customer can decide how this data must be visualised. Fig. 57 shows an example of this flexibility.

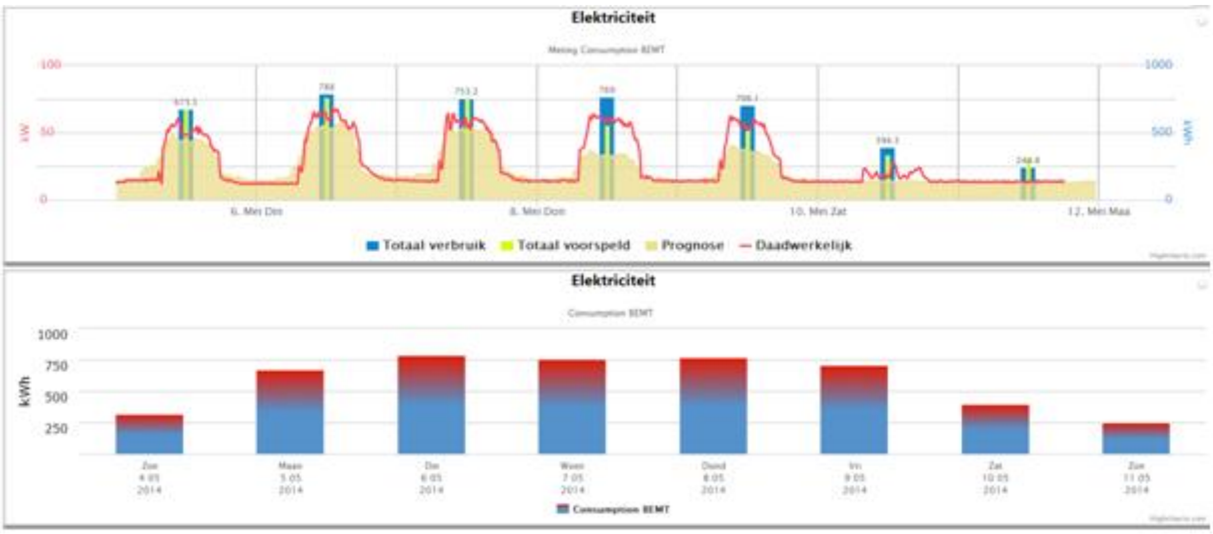


Fig. 57: flexible interactive energy screen

HVAC control:

This section describes the HVAC start point analysis (simulation) of the energy data scan, since a fine tuned version is used in control systems. The algorithm calculates the opportunity to save gas in the future. Therefore, the algorithm should know the new start-up moment of the heating system. This start-up moment is linked to the HVAC-control technique. By using this method the heating starts-up at the best moment for the installation. The algorithm uses weather predictions to determine this moment. For instance, a normal heating system uses the external temperature as a reference. Assume that it is 5°C in the morning and weather stations predicted 20°C at noon, the heating starts-up because the used temperature by the standard algorithm is 5°C. If the HVAC-control technique was used, the heating would not have been used that day. Because the algorithm would have known that the predicted temperature was 20°C. This is high enough because most heating algorithms turn off their system at an external temperature of 18°C. This HVAC-control technique is a tool which might be implemented in an upgraded version, therefore it fits within the operational control part of the standard. Fig. 58 shows a figure of the software tool.

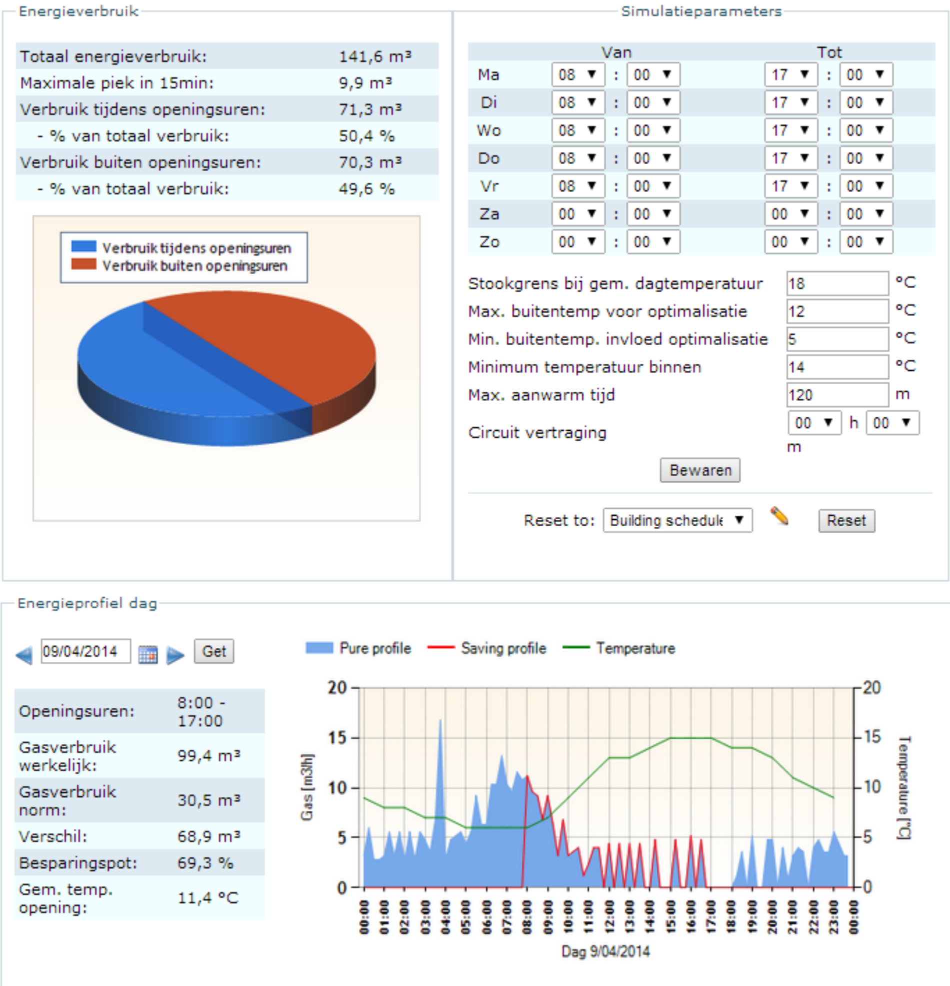


Fig. 58: HVAC-control

The settings necessary for the algorithm are shown at the right top of the screen. While the result of the algorithm is visualised below. The blue curve is the real gas consumption while the red line is the curve which the algorithm generated. If the technique was installed, the heating would have started-up at 7.30 o'clock. This action would have saved 66.3% ('Besparingspot' in the figure). This value visualises only the potential savings for one day. Fig. 59 shows the potential savings for a well defined period, from 04-2014 until now.

M.Start	O.Start	Warranty	Prognose	Pot. Savings
04/2014	05/2014	5%	15%	42%

Fig. 59: potential savings

Energy saving scan, energy alarms and data integrity:

The organisation must be sure that the action plan is working, so energy performance must be measured at planned intervals. Therefore, one of the last sections in the standard is checking. Firstly, one needs to be sure that the installed data loggers are still operating and providing data to the database. Porta Capena uses maintenance report to check if all the data loggers are still online. It is possible to set an alarm by email if a data logger goes offline. Furthermore, if the measured data is available, the objectives should be analysed against the reference period. Therefore, savings.EcoSCADA.com provides a platform which is used to check the objectives against the reference period. Finally, it is also possible to set several alarms. Form the moment that a threshold is reached, an alarm signal will be generated. If necessary such an alarm results into an email to the customer.

EcoSCADA reporting:

The final part of the ISO 50001 standard discusses the management review. But the time available to discuss the topics of an EnMS is limited. Therefore, the energy representative needs a couple of reports that visualise the essence. For this reason there are several reports developed.

- benchmark,
- year report
- building signature
- data quality

The benchmark report provides the top management an overview of the buildings position. This kind of benchmarks might be internal benchmarks as well as external benchmarks. Internal benchmarks refer to a comparison between the energy consumption now and in the past, so the benchmark focuses on one building. While the external benchmark focuses on the comparison with other similar buildings. Furthermore, the year report is used to give an overview of the energy consumption of the past year. Based on this kind of reports it is possible to set a baseline. On the other hand the building signature report visualises the energy objective, the reference period and analyses if the objective has been reached or not. Finally, the data quality report is more appropriate for the energy representative because it does not provide information about objectives or other targets. But it is useful when there has been a problem with the logging of the data.

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Data mining for Energy Management according to ISO 50001

Richting: **master in de industriële wetenschappen: energie-elektrotechniek**
Jaar: **2014**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Fosse, Jeroen

Datum: **5/06/2014**