A data-driven approach to reduce the energy consumption of buildings

van Rijn, F.R.

Award date:
2019

A DATA-DRIVEN APPROACH TO REDUCE THE ENERGY CONSUMPTION OF BUILDINGS

STUDENT NAME:	Folmer van Rijn
STUDENT NUMBER:	0918073
COURSE NAME:	Master project
PROGRAM:	Business Information Systems
SECTION:	Information Systems IE&IS
DEPARTMENT:	Industrial Engineering & Innovation Sciences

SUPERVISOR:	Prof. dr. Ir. U. Kaymak
SECOND READER:	dr. M. Firat
DATE OF SUBMISSION:	1 March 2019

ABSTRACT

Energy consumption by buildings in developed countries has increased substantially in the past years and it is predicted that it will increase even further. Most of the consumed energy is used for heating, ventilation and air conditioning (HVAC) to reach thermal comfort. Reducing the energy consumption by HVAC control will have a significant effect on the total energy consumption.

This study presents a data-driven approach to reduce the energy consumption while maintaining thermal comfort. The data used in this thesis is obtained from a utility building that is located in The Hague and consisted of weather-related data, HVAC data, room specific data and data of the general occupancy.

Two models have been constructed to predict the near future room temperature and another to predict the energy consumption on room level. Different modeling techniques have been used, whereof the MLP-Neural Network was selected as the most suitable. Evaluation of the models indicates that the room temperature model was of good quality, whereas the energy consumption model was far from acceptable due to inadequate data.

The temperature model is used in combination with the Model Predictive Control strategy for the selection of control settings. Simulation was used for the validation of the approach and indicated that the control technique was able to select the proper control settings under certain circumstances. It can therefore be concluded that under the right circumstances the approach can reach the desired state of reducing energy consumption while maintaining thermal comfort.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AHU	Air Handling Unit
ANN	Artifical Neural Network
APMonitor	Advanced Process Monitor
APOPT	Advanced Process OPTimizer
ARIMA	Autoregressive Integrated Moving Average
ASHREA	American Society of Heating, Refrigerating and Air Conditioning Engineers
BaB	Branch and Bound
BAS	Building Automation System
BFGS	Broyden–Fletcher–Goldfarb–Shanno
CBR	Case-Based Reasoning
CPU	Central Processing Unit
CRISP-DM	Cross-Industry Standard Process for Data Mining
EPBD	Energy Performance of Buildings Directive
EU	European Union
FIS	Fuzzy Interference System
GA	Genetic Algorithm
HVAC	Heating, Ventilation and Air Conditioning
IES	Integrated Environmental Solutions
IPOPT	Interior Point OPTimizer
KNMI	Koninklijk Nederlands Meteorologisch Instituut
kNN	K – Nearest Neighbors
LP	Linear Programming
MA & ES	Moving Average and Exponential Smoothing
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MILP	Mixed integer linear programming
MINLP	Mixed Integer Non-Linear Programming
MIMO	Multi Input Multi Output
MISO	Multi Input Single Output
MOGA	Multi-Objective Genetic Algorithm or Modified Genetic Algorithm
MOPSO	Modified or Multi-Objective PSO
MPC	Model Predictive Control
MLP-NN	Multi-Layer Perceptron – Neural Network
NLP	Non-Linear Programming
NN	Neural Network
PID	Proportional–Integral–Derivative
PMV	Predicted Mean Vote
PPD	Predicted Percentage of Dissatisfaction
PPM	Parts Per Million
PSO	Particle Swarm Optimization
QP	Quadratic Programming
ReLU	Rectified Linear Unit
RFT	Random Forest Tree
RMSE	Root Mean Square Error
SEMMA	Sample, Explore, Modify, Model, and Assess
SPEA	Strength Pareto Evolutionary Algorithm
SPEA-LS	SPEA with Local Search
SVM	Support Vector Machine
SVR	Support Vector Regression
VAV	Variable Air Volume
WFS	Weather Forecasting Service
XGB	Extreme Gradient Booster

1. INTRODUCTION

Over the past years the global energy consumption has increased substantially. A research conducted by Pérez-Lombard, Ortiz, & Pout (2008) has shown that in this period the total energy consumption of buildings in developed countries has increased between 20 to 40 percent. For many years the industry and transport sectors were the biggest consumers, but the energy consumption of buildings has exceeded them. This is caused by an increase in population, demand for thermal comfort and operational hours of companies. Roughly 50 percent of the total energy consumption of buildings in developed countries is used for heating, ventilation and air conditioning (HVAC) (Pérez-Lombard, Ortiz, & Pout, 2008).

The energy used for thermal comfort in buildings is a significant portion of the total energy consumption. Reducing it will strongly influence the total energy consumption. It is important to understand that thermal comfort plays a vital role, since it affects the human well-being and their productivity (Herkel, Knapp, & Pfafferott, 2005). Therefore, a solution that reduces energy should not decrease thermal comfort. This may raise the question whether reducing the energy consumption on HVAC while maintaining thermal comfort is even possible. In a research on this topic it has been estimated that over 39 percent of the energy that is used for thermal comfort is wasted on conditioning of unoccupied rooms, on overheating or undercooling, air leakage and inefficient appliances (Meyers, Williams, & Matthews, 2010).

Knowing that a substantial amount of energy is wasted, it may raise the question why the reduction of energy consumption by HVAC control has not been an important aspect for many companies and building owners to reduce expenses. The reason may be that the overall costs of energy are usualy only a small fraction of the total expenses for a company or building owner. Reducing the energy consumption is therefore usually not a first priority. In order to reduce the energy consumption of buildings it will be important to create awareness on the amount of energy used by a building and its negative effect on the environment. Besides awareness, legislation is required in order to ensure improvements. In a research about creating awareness, conducted by Wong & Krüger (2017), it was stated that a labeling system that indicates the energy consumption of a building would increase awarenes. Variations of this system have been adopted by multiple countries around the globe. For instance the Energy Performance of Buildings Directive (EPBD) in the European Union (EU) has introduced such a labeling system. This labeling system indicates the energy consumption of a building based on a rating scale of letters from A to G, where A is good and G is bad (Bull, Chang, & Fleming, 2012). The labeling system itself does not reduce the energy consumption, but by creating public awareness and by adding legislations the system does work. The EU enforced regulation by the implementation of Directive 2010/31/EU and as a result of this tightening of national

regulations in EU countries there are already examples of buildings in which a reduction of energy consumption up to 60 percent has been accomplished (Wong & Krüger, 2017).

1.1. PROBLEM STATEMENT

As stated in the introduction, buildings are one of the biggest energy consumers, particularly the energy that is used for thermal comfort. Moreover, a substantial part of the energy that is used for thermal comfort is even wasted on the conditioning of unoccupied rooms. So, cutting the energy that is spent unnecessarily could lead to a substantial reduction in energy consumption.

Mechanisms to reduce energy consumption of HVAC systems have been studied for quite some time, but so far there are not many systems implemented in buildings that are capable of reducing the energy consumption without compromising thermal comfort. Most buildings are still using on/off or PID control to operate their systems (Afram & Janabi-Sharifi, 2014). These control methods work on a reactive principle, where systems are activated when the measured condition does not satisfy the desired condition. The reason is that the control of a HVAC system is a quite unique and challenging process (Afram & Janabi-Sharifi, 2014). There are multiple factors that can have an influence on the thermal condition of a building. Four aspects that play a vital role are:

- the physical property of the building,
- the installed equipment of the HVAC systems,
- the outdoor environment and meteorological factors,
- presence/behavior of occupants in the building (Kwok & Lee, 2011).

Of these four aspects, the behavior of occupants is by far the most complex aspect. Human behavior is very unpredictable, and humans tend to do everything in their power to reach personal thermal comfort (Kwok & Lee, 2011). The combination of these factors makes the problem non-linear and hard to cope with by traditional solutions (Afram & Janabi-Sharifi, 2014).

A one size fits all solution for all buildings has not yet been discovered, due to the complexity and uniqueness of the problem for each building. Besides that, a standardized approach on the implementation of an advanced HVAC control system is also missing, so at the current state an implementation requires extensive expert knowledge and each implementation is a challenge on itself. This increases the costs significantly and hence increases the time for return on investment. Besides that, the energy reduction potential is hard to predict, due to the complexity of the problem. This prevents companies or building owners to invest in these types of improvements, due to the long return on investment timespan and the uncertainties that come along with it. Alternatives to reduce energy consumption are less expensive and the time for return on investment is easier to predict.

1.2. **RESEARCH QUESTION**

The importance of energy reduction increases the support for the development of new initiatives. The energy consumption related to climate control within buildings is one of the highest contributors to the total energy consumption. Reducing this could strongly influence the total energy consumption.

The overall complexity that comes along with the reduction of energy consumption of HVAC systems without sacrificing thermal comfort has been a threshold for quite some time. Improvements in data storage, computing and communication devices over the past years, made it possible to adopt technologies as machine learning into the construction of “smart” HVAC systems (Afram & Janabi-Sharifi, 2014). The amount of research on this topic has already increased, but so far there still are not many facilities around the globe that are equipped with a sophisticated system that is able to reduce energy consumption while maintaining thermal comfort by using near future knowledge.

A recent published review on the implementation of advanced control systems published by Serale et al. (2018) stated that implementation of advanced HVAC control systems using MPC are still lacking. There are only a sparse amount of papers that focus on this research topic, for instance da Costa Sousa & Kaymak (2001) published a paper on Model predictive control using fuzzy decision functions that included the optimization of a HVAC system using MPC. Although, further research on the implementation of such advanced contol system is required to reduce the complexity and the required level of know-how for implementation (Serale, Fiorentini, Capozzoli, Bernardini, & Bemporad, 2018).

This research tries to fill this gap by presenting a data-driven approach for the implementation of a HVAC control systems using advanced technology in a large utility building with a high thermal capacity. Based on this the following research question is formulated:

Can a data-driven approach be constructed for the implementation of an advanced HVAC control system that reduces the energy consumption and maintains thermal comfort in a large building with a high thermal capacity?

To answer the research question, four sub-questions have been formulated.

- **Which features are required to accurately model the energy consumption and temperature?**

This sub-question is answered by using both scientific literature and feature selection techniques. Previous conducted research in this field is used for preselecting the features. Afterwards, a feature selection technique is used to determine the final feature set that are used to predict the near future

temperature and energy consumption. The purpose of this sub-question related to the main research question is to establish the dataset that is used in this research.

- **Which modeling techniques are applicable to forecast energy consumption and temperature?**

This sub-question is answered by using scientific literature and comparing results from multiple modeling techniques. Scientific literature is used for the initial selection of modeling techniques. Afterwards, the models are constructed for the selected modeling techniques and the results from the models are compared to select the most suitable modeling technique. This sub-question is used for the selection of the modeling technique(s) that is used to answer the main research question.

- **Which control technique can be used to optimize HVAC control?**

This sub-question is also answered by using scientific literature. Literature on the different control techniques is studied and the most suitable technique is selected afterwards. An important prerequisite for the selected technique is that it can be used together with a cost-function. The main research question requires a control technique, which is selected via this sub-question.

- **Which optimization technique can be used to find local optima to the cost-function?**

The selection of the optimization technique is based on scientific literature. Previous conducted research in the field has already shown which techniques are viable. The main research question requires an optimization technique which optimizes the cost-function that describes the goal of reducing energy consumption while maintaining thermal comfort.

1.3. **METHODOLOGY**

This research is divided into a data mining part for the construction of the predictive models and a control optimization part for the control strategy. The control optimization part of this research is an extension on the data mining part and uses the results of the data mining part. Therefore, a data mining approach is adopted in this thesis and is extended with some additional steps that describe the control optimization part. Within the field of data mining approaches there are multiple methods that can be used to establish a structured approach to execute the research. The two most common methods in this field are Cross-Industry Standard Process for Data Mining (CRISP-DM) and Sample, Explore, Modify, Model, and Assess (SEMMA). A poll that has been executed for multiple years among data mining analysts has shown that CRISP-DM is the most used methodology for years (Harper & Pickett, 2006). SEMMA is the second most used standardized methodology but comes third on the list, since companies tend to either use CRISP-DM or develop their own methodology (Nadali, Naghizadeh kakhky, & Nosratabadi, 2011). The CRISP-DM methodology is explained in detail in chapter 0.

1.3.1. **CRISP-DM**

The CRISP-DM methodology consists of 6 steps as displayed in Figure 1. The general process of CRISP-DM is iterative and provides only general order in which steps have to be executed. Each step is briefly described.

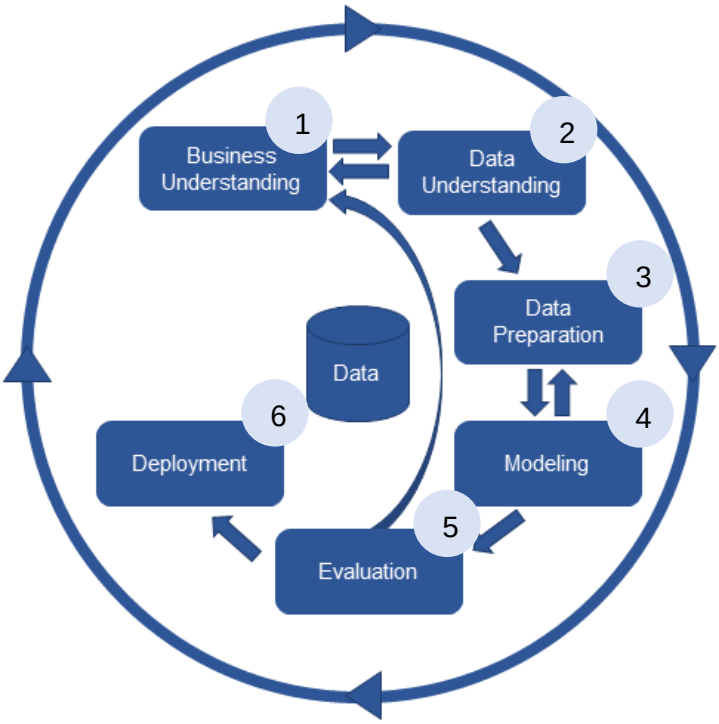


Figure 1 The steps of CRISP-DM.

1. In the *Business Understanding* step, the goal of the study and the requirements from the business perspective are defined. This step requires also to collect all the information that is required to properly execute the research and get a wide understanding of the field of research.
2. The *Data Understanding* step consists of collecting the data, identifying what is present in the data and which parts of the data are required for the research.
3. In the *Data Preparation* step the data that is selected in the previous step is “prepared” to be used for data mining techniques. This step is usually the most time-consuming step.
4. The *Modeling* step consists of the use of multiple modeling techniques on the prepared data to address the business needs that are described in the first step.
5. In the *Evaluation* step the models that are constructed in the previous step are evaluated on their quality. Common techniques that are used to assess the model on their quality are accuracy and generalization. Although, there are more techniques that can be used besides the aforementioned. It depends on the type of model that is constructed which techniques should be used.

6. The *Deployment* step depends on the purpose of the analysis. This could vary from presenting the results of the analysis in a paper, to the implementation in a system

(Maaskant, 2016).

1.4. **SCOPE**

This research project presents a data-driven approach to optimize HVAC control to reduce energy consumption while maintaining thermal comfort. The approach is designed specifically for utility buildings that have an advanced Building Automation System (BAS) to control the temperature within the entire building. Besides that, the existence of historical data about the configuration of the BAS is a prerequisite. The approach is specifically for the optimization of temperature control by using the adjustable heat sources that are available in a room.

2. LITERATURE STUDY

The literature study is structured according to the general process of HVAC optimization using a data-driven MPC approach from a technological perspective (displayed in Figure 2). It consists of multiple steps where each step is a process on itself. The resulting quality of each step has a direct influence on the quality of the subsequent steps. Each numbered step of the process is extensively elaborated and corresponds to the section number in the literature study chapter. After completing the whole process, results from previous conducted researches are briefly discussed, together with a joint conclusion.

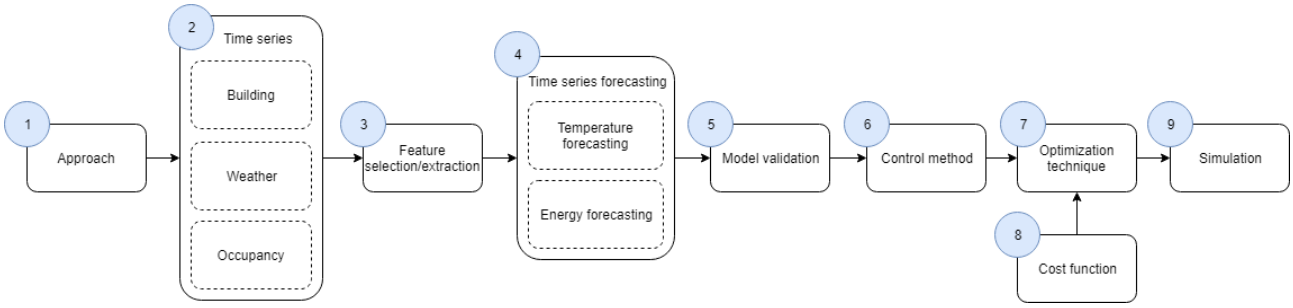


Figure 2 Process HVAC optimization using a data-driven approach from a technological perspective.

A brief description of each section is given below:

1. The *Approach* section describes on a top level the different types of approaches that can be used for HVAC optimization.
2. The *Time Series* section describes the process of analyzing time series data and the relation between the different data sources.
3. The *Feature Selection/Extraction* section describes the techniques to separate the relevant features from the irrelevant features and how features with strong recognition ability can be constructed.
4. The *Time Series Forecasting* section elaborates on the different techniques used for forecasting. Advantages and disadvantages of each of the techniques are listed together with results from previous conducted researches.
5. The *Model Validation* section describes the different techniques that are used to evaluate the quality of a model.
6. The C*ontrol Method* section elaborates on the different types of control techniques that can be used to control the HVAC systems. The control technique Model Predictive Control (MPC) is described in depth.
7. The *Optimization Technique* section describes the advantages and disadvantages of the different optimization techniques that can be used to minimize one or multiple cost functions.
8. The *Cost Function* section describes the goal of the function and lists some of the most commonly used cost functions.

2.1. **Classification of Data-driven approaches**

The optimization of HVAC control using a data-driven approach can be classified by ASHRAE (American Society of Heating, Refrigerating and Air Conditioning Engineers) in three broad groups as displayed in Figure 3 (American Society of Heating, 2013).

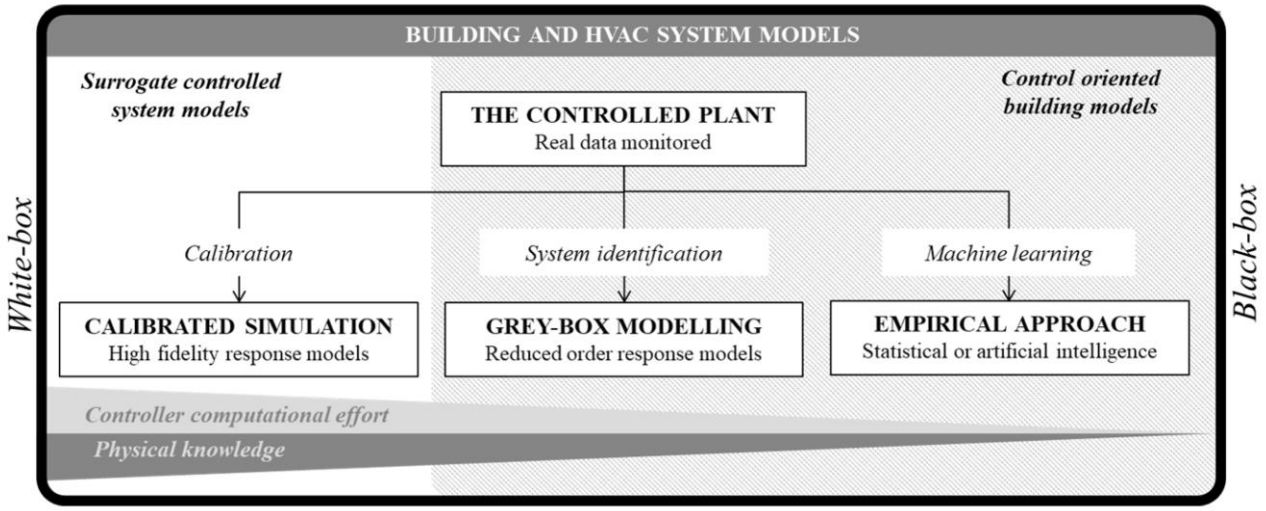


Figure 3 Different modeling approaches according to ASHREA (Serale, Fiorentini, Capozzoli, Bernardini, & Bemporad, 2018).

- Calibrated simulation models, sometimes referred to as white-box models, are models based on the physical principle to calculate the thermal dynamics and energy behavior of the whole building level or sublevel components (Magoules & Zhao, 2016). The construction of these models requires a substantial amount of knowledge on the physical properties of the building to correctly model the dynamics of the building. The downside of white-box modeling is that the complexity and the size of the calibrated simulation lead to optimization problems that exceed the computation timeframes required in a practical control application (Serale, Fiorentini, Capozzoli, Bernardini, & Bemporad, 2018).
- Empirical approach, sometimes referred to as black-box models, is exactly the opposite of white-box models. The construction of this type of models requires barely any physical knowledge and is purely based on the historical behavior of the building. Disturbances as climate and occupancy behavior on the thermal condition of the building are encapsulated in the dataset. Parameters are selected from the dataset and used to construct a model to predict the future state. The downside is that it can only produce reliable predictions within the range that is covered in the dataset. This approach can only be used when a sufficient amount of data is available.
- Grey-box is everything between white- and black-box modeling. Grey model prediction constructs a model based on selected parameters, but also includes the dynamics of the

physical system in the model. The construction of these models requires slightly more physical knowledge of the building. The downside of grey-box is the amount of information that is used to train the model, so proper selection is critical to prevent the model from under/over-fitting (Bacher & Madsen, 2011). For the construction of a grey-model it is common to use the Resistance-Capacitance (R-C) network (R-C networks are explained in the next paragraph) to describe the thermal process dynamics of a building (zone) (Serale, Fiorentini, Capozzoli, Bernardini, & Bemporad, 2018).

A review published by Serale et al., (2018) stated that roughly 50 percent of the papers use a grey-box approach.

Resistance-Capacitance (R-C) network

An R-C network models the transfer of indoor and outdoor heat of a certain area within a building. The area can be the size of a single wall, or that of an entire building. The size determines the level of abstraction in which the model is constructed. There are numerous aspects in play and describing each aspect separately would make the model become impossible to read and would lead to a state-space that could not be solved. So, the goal of the R-C model is to capture the physical relation that governs the indoor and outdoor energy interaction as precisely as possible, while keeping the state-space formulation of the model to a minimum so that it can be solved efficiently (Bueno, Norford, Pigeon, & Britter, 2012). A typical example of an R-C network model is displayed in Figure 4.

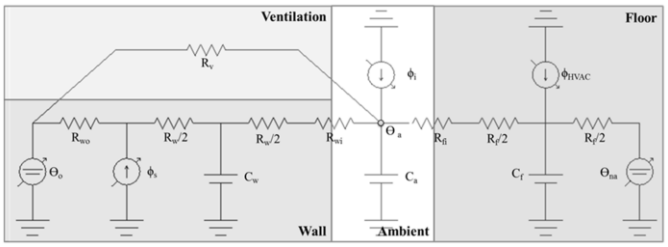


Figure 4 A typical example of an R-C network for MPC applications obtained from Serale et al., (2018).

2.2. **Time series**

A time series is a dataset of a time ordered sequence of observations, often with equal intervals between them. An import step before fitting a model is to understand and obtain the structure and the underlying pattern of the series (Deb, Zhang, Yang, & Lee Eang, 2017). The general approach to obtain the structure and understand the underlying pattern is to decompose the time series into systematic and non-systematic components.

The systematic components in a time series are the consistent and recurring patterns that are present in the series. The components can be divided into level, trend and seasonality (Brownlee , 2017).

- *Level* is the average value of the series.
- *Trend* is the general movement of the variables in the series.
- *Seasonality* is the repeating cycles of fluctuations in the series.

The Non-systematic components are the uncertainties in the series that cannot be modeled and are called the *residual* or *noise* that is present in the series.

Stationarity

For the statistical analysis of time series, it is often required that the time series is stationary. A time series is stationary when the mean, variance and autocorrelation are all constant over time (Hyndman & Athanasopoulos, 2013). A time series that contains trends or seasonality is not stationary. The trend and seasonality will affect the values of the time series over time and hence affect the mean, variance and autocorrelation. An often-made mistake is to confuse cyclic behavior with seasonality. Cyclic behavior is a recurring pattern that does not have a fixed length. A time series that has cyclic behavior but does not contain any trend or seasonality is still stationary. In general, stationarity is very uncommon in most real-life datasets (Thomson, 1994). When a non-stationary series contains a stable long-run trend, it may be possible to make the series stationary by using de-trending. A series becomes trend-stationary when it can be made stationary by using de-trending. A series that contains seasonality besides trends will not be stationary even after de-trending. Differencing can be used to stabilize the mean and variance by eliminating or reducing the trend or seasonality (Nau, 2017).

Data providers

In a data-driven approach to optimize HVAC control there are multiple data sources from which data can be extracted. The main data source contains the data of the HVAC system and building. Data sources with specific data about the weather or occupancy rate are additions that can be used to improve the quality of the models.

HVAC data

Data of the HVAC system and buildings can be obtained from either a simulation tool that generates data or historic data that can be extracted from the BAS of a building. Real data is generally preferred since it captures all factors that have an influence on the results (Deb, Zhang, Yang, & Lee Eang, 2017). When real data is not present simulation tools can be used to generate the data. Simulation tools require a large amount of detailed information of the building to produce accurate data. This process of gathering all the detailed information and modeling is usually a time-consuming process (Kwak, Seo, Jang, & Huh, 2013) (Zhou, Wang, Xu, & Xiao, 2008) (Li, Meng, Cai, Yoshino, & Mochida, 2009). A large amount of systems is available on the market that can perform building simulations. Some of the most popular are: DOE-2, eQUEST, IES, EnergyPlus and ECOTECT (Deb, Zhang, Yang, & Lee Eang, 2017). These simulation tools produce HVAC data based on the preconfigured settings. The quality and accuracy of the produced data depends on how accurate the building is modeled in the tool.

Weather data

Weather data is commonly used as an addition to enhance the accuracy of load prediction in buildings (Lazos, Sproul, & Kay, 2014). Weather data can be divided into two groups: historical data and expected near future. Historic weather data is usually accessible as a public service from government institutes, whereas expected near future weather is obtained from different sources.

A paper published by Lazos, Sproul, & Kay (2014) researched weather forecasting techniques. The paper stated that weather forecasting data can be obtained by using a model to predict the weather based on the current state or can be extracted from third-party providers. Predicting the near future weather by using an onsite weather station has shown to be more accurate than extracting it from a third-party provider. The construction of such a model can be a time-consuming process. For this reason, most commercial buildings use data extracted from third-parties, rather than using an onsite weather station in conjunction with prediction methods (Lazos, Sproul, & Kay, 2014).

A paper published by Sun, Wang, & Xiao (2013) researched the correlation coefficient of the weather variables (outside temperature, relative humidity, radiation and cloudiness) to the internal building temperature. In the paper they stated that dry bulb-temperature had the strongest correlation which was about 0.6. The weather variables relative humidity and radiation showed a correlation of about 0.4. The lowest correlation was the cloudiness with a correlation of about 0.08 (Sun, Wang, & Xiao, 2013). The results from this analysis can deviate for other buildings since the location of the building and the building materials have an influence. For instance, a building with more windows is likely to gain more heat from radiation than a building with fewer windows.

Occupancy data

Another data source that can be used to improve the accuracy for load prediction in buildings is the occupancy rate (Kwok & Lee, 2011). Besides that, occupancy pattern recognition can be an optimization technique on its own. For instance, most HVAC systems are programmed to keep the temperature of a room always within acceptable ranges and ventilates prior to maximum occupancy. But some rooms, like conference rooms, are not always used or only used by a small group of people. Conditioning based on the presence of occupants would be more efficient. This method requires a system that identifies whether a room is occupied or not and uses this information to control the temperature in the room. The disadvantage is that, when an occupant enters the room, the room temperature possibly is not within the desired temperature ranges. A more advanced alternative predicts whether a room becomes occupied in the near future and conditions the room slightly in advance. Erickson & Cerpa, (2010) presented a model that conditioned a room based on near future occupancy prediction. They also adjusted the ventilation speed in rooms based on the number of occupants inside it to keep the carbon-dioxide level within acceptable ranges (Erickson & Cerpa, 2010). The downside of this technique is that it requires advanced technology that can

measure the number of occupants in a room in real time. Also, when a room becomes unexpectedly occupied, the room has not been pre-conditioned yet.

2.3. **Feature selection and extraction**

A high dimensional dataset is a set that consists of a high number of features and becomes hard to compute. The combination of HVAC, weather and occupancy data has the characteristics of such a dataset. Reducing the number of features has the following advantages:

- The machine learning algorithm trains faster
- Reduces the complexity of the model
- Improves the accuracy, when the correct features are selected
- Reduces the chance of overfitting the model (Kaushik, 2016) (Cai, Luo, Wang, & Yang, 2018).

There are several techniques that can be used to reduce the dimension size. The most commonly used technique is feature selection, whereas feature extraction is optional and may not always reduce the dimension size.

- Feature selection is the process of separating the relevant features from the irrelevant features of the high dimensional space. This process results in a sub-set of the initial data set where only the relevant features remain.
- Feature extraction (or feature construction) is the process of transforming data to features with a strong recognition ability. This technique can be used to reduce the dimensional space, when multiple features are combined to a single new feature. But this

An alternative, when expert knowledge is present, is to select the features without using automated feature selection or extraction techniques (Guyon & Elisseeff, 2003). For instance, when previous conducted research has already studied the recognition ability of features, this can be used to select the features. Besides that, it is not uncommon to use a combination of these techniques. The initial technique is used to make a preselection of the feature, which is usually done by a more lightweight technique, followed by a technique that constructs the final feature set based on the preselection.

Feature selection

There are many feature selection methods that have been suggested to solve the variable selection problem. The feature selection techniques can be categorized into: Filter, Wrapper and Embedded. Another aspect, which plays a role in each category, is the type of data.

Filter

The filter method is a lightweight selection method with a low computational cost and complexity (Guyon & Elisseeff, 2003) (Yu & Liu, 2003) and it is independent of any machine learning algorithm (Bekkerman, El-Yaniv, Tishby, & Winter, 2003) (Forman, 2003). This feature selection method usually presents a ranking instead of an explicit best feature subset. A cut off point for the ranking

can be chosen by way of cross-validation. Formulas that measure the correlation, dependency or distance between features are commonly used (Cai, Luo, Wang, & Yang, 2018). The most common formula to calculate the correlation is the Pearson correlation coefficient (Rodriguez-Lujan, Huerta, Elkan, & Cruz, 2010).

Wrapper

The wrapper method selects a subset of the features, constructs a model based on that data and evaluates the quality by measuring the accuracy (John, Kohavi, & Pfleger, 1994) (Kohavi & John, 1997) (Blum & Langley, 1997). Based on the results, the wrapper adds or removes a feature and repeats the process. This iterative process is usually very computational expensive since it needs to evaluate a cross-validation scheme for each iteration (Lal, Chapelle, Weston, & Elisseef, 2006). The method of selecting the features can be categorized into: forward selection methods, backward elimination methods and nested methods (Lal, Chapelle, Weston, & Elisseef, 2006).

- The forward selection method starts with an empty feature space and keeps adding features until all features are selected or when the quality doesn't increase anymore. In reality, this means that for each iteration all features are first compared, and the best feature is selected afterwards (Sutter & Kalivas, 1993).
- The backward elimination method works according to the opposite of the forward selection method. It starts with all features and removes a feature each round for as long as there are features available and the quality is not decreasing. The only difference is that once a feature is removed it cannot return and it becomes uncertain whether the final sub-set will also be the optimal sub-set (Sutter & Kalivas, 1993). Moreover, when there are many features this strategy becomes infeasible.
- Nested methods can add and remove features from its selection at each iteration (Lal, Chapelle, Weston, & Elisseef, 2006).

Embedded

The embedded method is a combination of the wrapper and filter methods (Kaushik, 2016). The embedded model selects the features for the next iteration during the process of training and validates the model with the previous selected features (Cai, Luo, Wang, & Yang, 2018). Two of the most popular techniques are the Lasso and RIDGE since they have an inbuilt penalty function for overfitting (Kaushik, 2016).

Comparison

The filter method is usually preferred over the other techniques. This is due to its usability with different classifiers, the low calculation time and their simplicity (Guyon, 2003; Yu and Liu, 2003). The downside is that just combining the best features does not generally lead to the best overall results (Cover, 1974; Cover and Thomas, 1991; Jain et al., 2000). The wrapper method generally outperforms the filter method on accuracy due to this downside (Zhang, Li, Wang, & Zhang, 2013),

but tends to over-fit (Hu, Gao, Zhao, Zhang, & Wang, 2018). The embedded method is likely to overfit as well. Only when there is enough training data available the embedded method is likely to perform better than the filter method (Lal, Chapelle, Weston, & Elisseef, 2006). Finding the optimal number of features can therefore be a challenging task when the number of features in a dataset increases. A paper published by Chen, Wilbik, van Loon, Boer, & Kaymak (2018) presented a method that uses the mutual information of features to determine the optimal number of features.

Feature extraction

Feature extraction is a technique that can sometimes be used to reduce the dimensional space. The primary goal of feature extraction is to convert raw data into a set of useful features (Guyon & Elisseeff, 2006). This could reduce the dimensional space when multiple features are combined to a new feature, but the dimensional space may also remain the same or even increase. The extraction of useful features from data often requires human expertise to know what could lead to useful features. The downside of this technique is that useful information can be lost when new features are constructed. It is sometimes advised to validate whether the new feature improves the performance compared to the original feature. As stated, this technique requires more human expertise, but it can be complemented with automated feature construction methods. Examples of Automated Techniques are: Standardization, Normalization, Signal enhancement, Extraction of local features, Linear and non-linear space embedding, Non-linear expansion and Feature discretization (Guyon & Elisseeff, 2006).

2.4. **Time series forecasting**

Time series forecasting is a method for predicting events through a sequence of time. The forecasting methods can be categorised into: statistical methods, machine learning methods and physical methods. A hybrid version that combines methods is not uncommon, since techniques can work to complement each other and hence, increase the quality of the forecast (Lazos, Sproul, & Kay, 2014).

- Statistical methods try to find the linear relationship between variables and use this to predict the future (Srivastava, 2015).
- Machine learning methods try to simulate the non-linear and non-stationary univariate or multivariate dependencies through network (Lazos, Sproul, & Kay, 2014).
- Physical methods are based on attempting to produce a mathematical model that captures the physical process and use this to predict the future state (Lazos, Sproul, & Kay, 2014).

Selecting a technique can be a difficult task, since there are many techniques available. So far not a single best solution exists or a one size fits all solution. Each technique may have it advantages and disadvantages compared to others. A possible strategy would be to find a similar scenario in

which a technique was used successfully. Below are four scenarios in which different forecasting techniques were used to predict the energy consumption or electric load.

1. A feedback Artifical Neural Network (ANN) has been used by González & Zamarreño, (2005) for the short term energy load prediction of a building. The performance was measured by using ten datasets with real data. On each of the data sets, a model was constucted and validated. It scored an average Mean Absolute Percentage Error (MAPE) of 1.945, with a maximum of 2.88 (González & Zamarreño, 2005).

2. Autoregressive Integrated Moving Average (ARIMA) has been used by Abdel-Aal & Al-Garni, (1997) to forecast the monthly electric consumption in a building. They used 5 years of data to train the model and validated it on the 6^{th} year. They stated that the average percentage error for the model was 3.8 percent (Abdel-Aal & Al-Garni, 1997).

3. A paper published by Oğcu, Demirel, & Zaim, (2012) compared the Artifical Neural Network (ANN) and the Support Vector Machine (SVM) on forecasting the electricity consumption in Turkey. They used two years of monthly electric energy consumption data to test the models and that resulted in a MAPE for ANN and SVM of 3.9 percent and 3.3 percent respectively (Oğcu, Demirel, & Zaim, 2012).

4. Liu et al., (2010) proposed a short-term electric load forecasting method by using sliding window fuzzy time series. They used slide window fuzzy time series first to train a trend predictor, and then used this trend predictor to forecast. The model was validated by using the electric load of four days and it resulted in a maximum MAPE of 7.74 percent (Liu, Bai, & Fang, 2010).

These are just a couple of scenarios in which different techniques were succesfullly used to forecast the electric consumption or load. A comprehensive review was published by Deb et al., (2017), which compared the nine most common forecasting techniques for buildings, which are listed below.

- Artifical Neural Network (ANN)
- Autoregressive Integrated Moving Average (ARIMA)
- Support Vector Machine (SVM)
- Case-Based Reasoning (CBR)
- Fuzzy Interference System (FIS)
- Grey-box modeling
- Moving Average and Exponential Smoothing (MA & ES)
- k – Nearest Neighbor (kNN)

- Hybrids

In total they reviewed 166 different papers, where one of the techniques was used to forecast energy consumption or load in buildings. They concluded that each of these time series forecasting techniques has been succesfully applied in the prediction of energy consumption in buildings. Based on the reviewed papers, they published a qualitative comparison on the different techniques, which is displayed in Figure 5.

Model	Advantages	Disadvantages
ANN	1. Ability to precisely map input and output relationships	1. Depends on initialization of weight values
	2. Performance well for non-linear time series	2. Problem of the local minima
	3. More general ad flexible	3. Overfitting and difficult to generalize
ARIMA	1. Uses lag and shift of historical data	1. Model identification is difficult
	2. Regression model with a moving average (improves accuracy)	2. Not suitable for long-term predictions
	3. Provides confidence intervals on predictions with reliability	3. Does not fully capture the non-linear patterns of the series
SVM	1. Good for fitting and generalization	1. lack of transparency of results
	2. Performs well for long-term time series	2. Finding optimum parameters can be a computational burden as
	3. Use of kernel function introduces nonlinearity and deals with arbitrarily structured data	number of parameters and size of dataset increases
CBR	1. Similar to human cognitive processes	1. Needs introduction of new aspects, e.g. case representation for time series processing
	2. Doesn't need to find rules between parameters of the problem	2. Needs huge data
Fuzzy	1. Close to human experience via membership functions and rules	1. Temporal patterns are defined by rigid regions, hard to adjust with noise
	2. Good for solving uncertainties in load forecasting	2. High computational complexity and lacks stability
Grey	1. Capable of predicting with limited data and incomplete information	1. Inadequate in recognizing random components
	2. Easy to compute and calculate	2. Problem with conventional approach of model validation
MA & ES	1. Simplicity in calculations	1. Poor results compared to sophisticated techniques
	2. The use of low number of observations	2. Not suitable for long-term and non-linear predictions
	3. Transparency in approach	
NN	1. Simple process with no explicit training step required	1. Function is often approximated only locally
	2. Intuitive and ease of implementation	2. Challenging to compute exact number of nearest neighbors
Hybrid	1. Complimentary combinations of different machine learning methods	1. High model complexity
	2. Robust for complex problems and often improves performance	2. Computationally intensive
		3. Often difficult to identify with methods to combine

Figure 5 Summary of qualitative comparison for the 9 major time series forecasting techniques (Deb, Zhang, Yang, & Lee Eang, 2017).

This qualitative comparison can be used as a guideline for selecting a forecasting technique. It can be noted that, even with these guidelines, there are still many uncertainties, which makes it difficult to tell which technique would be the most appropriate for each case. Besides that, even if a technique has the odds in its favor, there is no guarantee for success.

2.5. **Model quality evaluation methods**

Modeling evaluation techniques are used as a measurement to determine the quality of a model from a certain perspective and to compare different classification models. There is a large variety of verification techniques available and each technique describes the quality from a different perspective.

The verification of time series is slightly different compared to traditional regression classification techniques. In time series, time plays an essential aspect since behavior can alter over time. Validating the performance on historical data may not present a valid representation of the performance and therefore only new data can be used to determine the performance of a model

(Hyndman R. J., Measuring forecast accuracy, 2014). The size of the training set is usually around eighty percent of the total data-set but only when a sufficient amount data is present. When there is a limited amount of data available, splitting it in test and train would even further reduce the amount of data that can be used to train the model. A solution to this problem is cross-validation that constructs multiple different train and test sets from the same data. There are multiple types of cross validation, one of them is time series cross validation. The concept of time series cross validation is to have multiple training sets, where each training set has one more observation than the previous set, as displayed in Figure 6.

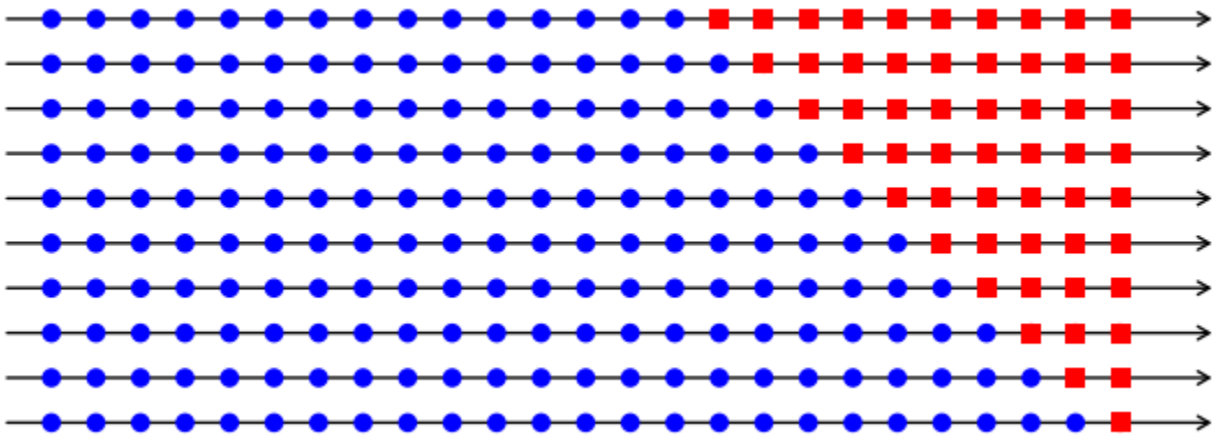


Figure 6 A graphical representation of time series cross-validation (Hyndman R. J., 2014).

In this figure the blue dots represent the observations of the training set, and the red squares represent the test observations. The accuracy is calculated by averaging the results of the individual sets. An alternative to this is the one-step forecast displayed in Figure 7, in which the structure of the train set is identical but the test set contains only the next observation.

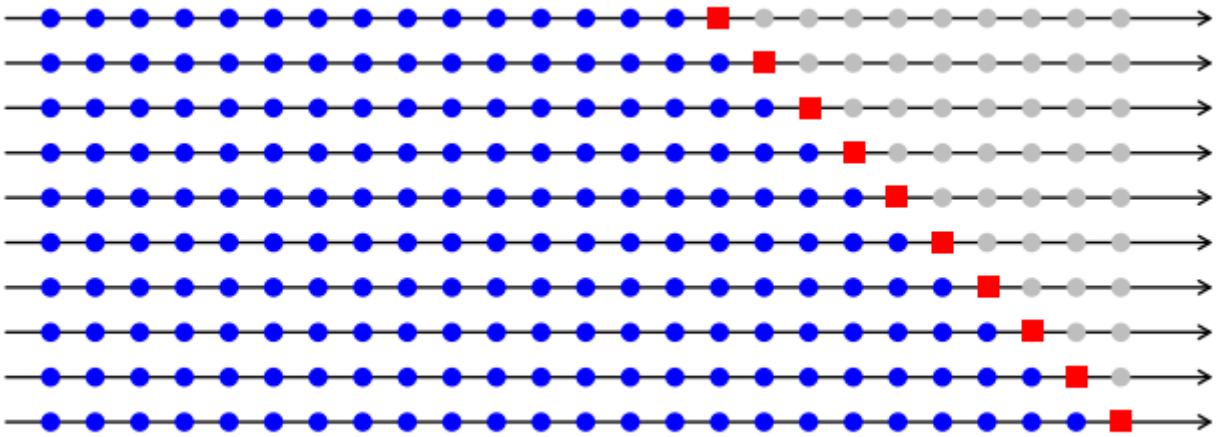


Figure 7 A graphical representation of the one-step time series cross-validation (Hyndman R. J., 2014).

In this figure the blue dots represent the observations of the training set, and the red squares represent the test observations. The forecasting accuracy is based on the average score of all sets.

Still, in any of the time-series cross validation approaches a minimum amount of data is required to produce reliable accuracies and models (Hyndman R. J., Measuring forecast accuracy, 2014).

For regression there is a variety of commonly used evaluation techniques (Hyndman & Koehler, 2006). The techniques can be divided into four categories: scale-dependent measures, measures based on percentage errors, measures based on relative errors and relative measures. The scale-dependent measure techniques are useful for comparing different models using the same dataset, but with the restriction that the data consists of a single scale. Whereas the measures based on percentage errors are scale-independent and are useful for comparing forecast performances across different data sets.

Scale-dependent

The most preferred techniques of the scale-dependent measure techniques are the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) (Hyndman R. J., Measuring forecast accuracy, 2014). The RMSE technique has the downside that it is sensitive to outliers. Therefore, using it for forecast accuracy evaluation is sometimes advocated against (Hyndman & Koehler, 2006). In the energy consumption predicting this formula is commonly used (Monfeta, Corsib, Choinièreb, & Ark, 2014) (Tso & Yau, 2007).

The RMSE measures the average magnitude of the error that uses a quadratic scoring rule. It takes the square root of the average of squared differences between prediction and observation.

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

The variable y_t in the formula denotes the observed value at time t, and the variable $\hat{y}_t$ is the forecasted value at time t. The n in the formula denotes the total amount of variables which are predicted over.

Another variant of the RMSE is the Coefficient of Variance Root Mean Square Error (CV-RMSE). This function determines the goodness of fit of the model and should be below 25% for energy prediction (Haberl, Culp, & Claridge, 2005). The CV-RMSE formula:

$$CV - RMSE = \frac{\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}}{\bar{y}}$$

In this formula the variables y_t and $\hat{y}_t$ remain the same as in the previous formula, and the variable $\bar{y}$ is the mean value of all the observed values.

The MEA formula measures the average magnitude of the errors, without considering their direction. It takes the average of the absolute difference between the predicted and observed value with an equal weight.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

The variables y_t and $\hat{y}_t$ remain the same as in the previous formulas.

Percentage error

The percentage error has the advantage of being scale-independent and is therefore frequently used to compare the performance of forecasting models on different datasets. The Mean Absolute Percentage Error (MAPE) is the most commonly used measure function (Hyndman R. J., Measuring forecast accuracy, 2014).

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

The disadvantage of the MAPE is that when y_t is zero, its response is undefined and when it is not zero but close to it, it returns an extreme value. Besides this a common mistake is to think that scale-independent can be used to compare Fahrenheit and Celsius but these are not measuring a quantity and therefore cannot be compared (Hyndman R. J., Measuring forecast accuracy, 2014).

2.6. **Control methods of HVAC systems**

The instructions for HVAC systems are selected using a HVAC control method. A list of control methods that are used for the control of HVAC systems are displayed in Figure 8.

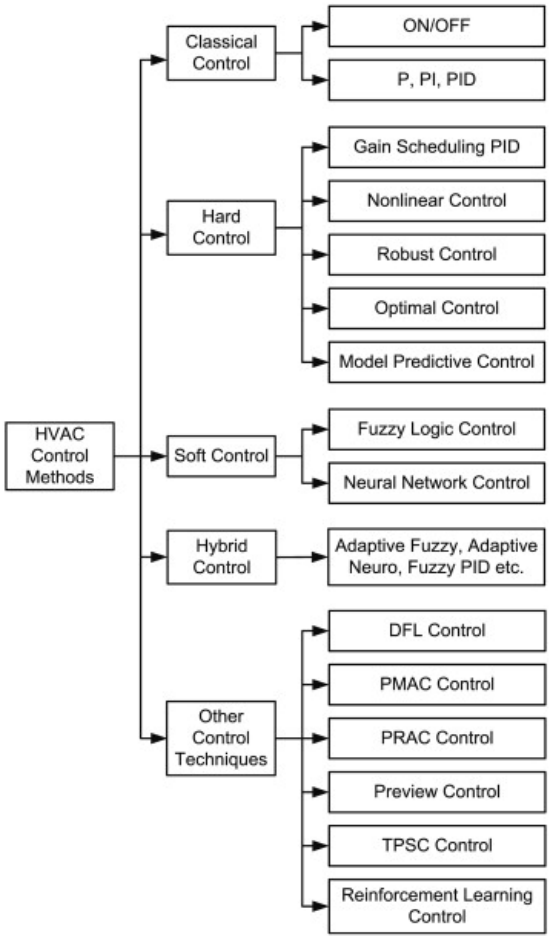


Figure 8 A structured display of all the different types of HVAC controllers.

These control methods are divided into classical control, hard control, soft control, hybrid control and other control techniques. The classical control techniques are the most commonly used to control HVAC systems (Afram & Janabi-Sharifi, 2014).

Model Predictive Control

One of the Hard control techniques is the Model Predictive Control (MPC). MPC is a control method in which the control objective is translated into an optimization problem over a finite horizon. The goal of the optimization problem is to minimize a certain cost function while staying within given boundaries. The boundaries can regulate the upper and lower limits of the controllable variables and outputs. Common controllable variables for HVAC systems that can be adjusted are: temperature, ventilation speed and pressure. The structure of the cost function depends on what should be optimized. Common goals of the cost function related to HVAC are: energy and cost savings, peak load shifting, steady-state improvements, better regulation and improved thermal comfort (Afram & Janabi-Sharifi, 2014) (Lazos, Sproul, & Kay, 2014).

The goal of improving thermal comfort can be hard to measure since the perfect condition is different for each individual. Two methods that can be used as thermal comfort indicators and can also be

used as constraints for the MPC are Predicted Mean Vote (PMV) and Predicted Percentage of Dissatisfaction (PPD) (Mirakhorli & Dong, 2016). The goal of the methods is to calculate the percentage of people that will feel thermal discomfort.

The general principle of MPC is represented in Figure 9. From this figure the x-axis is considered the time-axis that can be separated into 3 different time horizons: Prediction Horizon, control horizon and sample time.

- **Prediction horizon** (or optimization horizon) is the entire horizon over which the optimization problem is used. As displayed in Figure 9, this starts at time 0 and continues up until the point where the line of the predicted output stops.
- **Control horizon** is the horizon over which the control input can be adjusted. The control horizon is always between one sample time up to the entire control horizon. In the figure the control horizon is equal to the prediction horizon, since the control input has been changed in the last sample time of the prediction horizon. One of the reasons to take a smaller control horizon than the prediction horizon is to reduce the computational complexity of the optimization problem.
- **Sample time** (or control step) is the amount of time the control input remains identical.

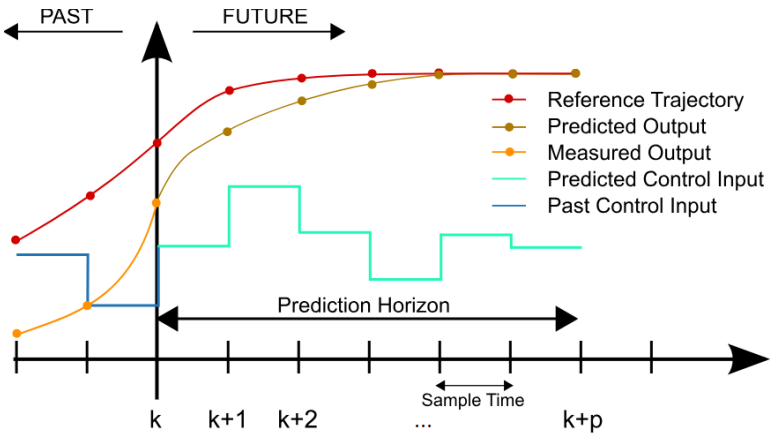


Figure 9 A graphical representation of the control method MPC (Behrendt, 2009).

The lines within the graph display each a different aspect:

- The **Reference Trajectory** is the target point that the measured/predicted output tries to follow as close as possible.
- The **Predicted Output** is the future expected state of the system, based on the selected control input.
- The **Measured Output** is the measured state of the system in the past. The past is sometimes used as part of the prediction of the future state. The display of the measured output is optional and therefore not always visible.

- The **Predicted Control Input** describes the future state of the control settings. The figure displays only a single predicted control input, but this can also be multiple control inputs.
- The **Past Control Input** displays the previously used control settings of the system. The display of the past control put is optional and therefore not always visible.

The length of the prediction horizon, control horizon and sample instant differ in each case. An aspect that plays a role in the selection of the sample time, is the time it takes before changes show effect. Durations commonly used in slow-moving processes of HVAC systems are: about 5 to 48 hours for the prediction horizon, 4 to 5 hours for the control horizon and 1 to 3 hours to the sample time (Afram & Janabi-Sharifi, 2014). Serale et al., (2018) published a more extensive review and presented the time span of the sampling time (Figure 10a) and prediction horizon (Figure 10b) of the papers they analyzed. They also included the number of optimization time-steps (Figure 10c) that have been witnessed. It can be noted that the number of steps is usually between 13 and 24.

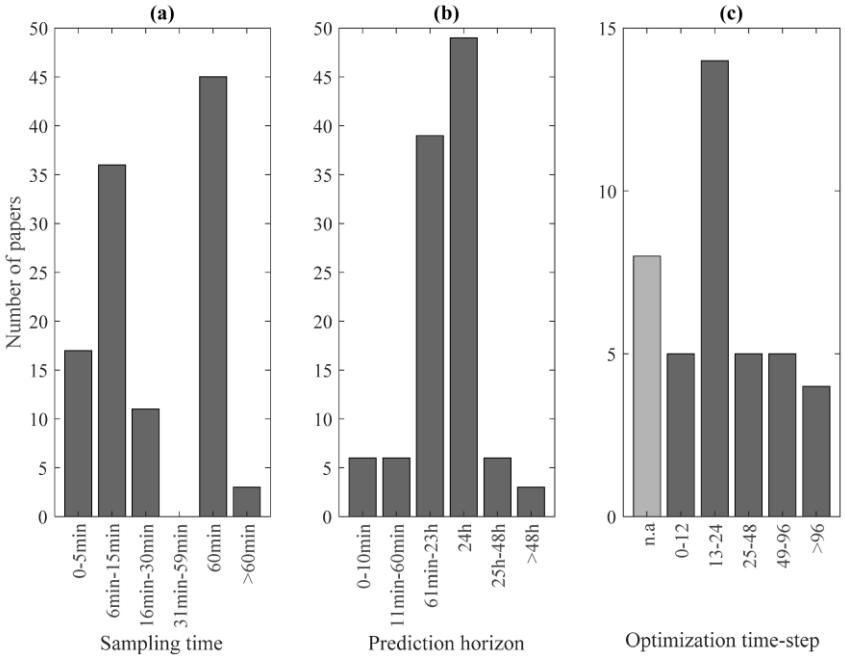


Figure 10 Frequency of the sample time (a), prediction horizon (b) and number of optimization steps (c) published by Serale et al., (2018).

It can be concluded that the average timespan of the control horizon over the two papers show similarities. The sample time on the other hand is slightly higher in the paper of Afram & Janabi-Sharifi, (2014). A possible reason is that the paper published by Afram & Janabi-Sharifi, (2014) may have only incorporated papers that use supervised control. Supervised control usually has a larger sample time.

It is important to understand that this whole process repeats itself after each sample time and only the control variables of the first sample time are used as input for the system. The remaining control

variables are discarded. The reason that the process repeats itself and uses only the first future state is to eliminate any un-modeled disturbances and modeling errors (Afram, Janabi-Sharifia, & FungaKaamr, 2017).

The Model Predictive control (MPC) controller offers many advantages over the other controllers (Afram & Janabi-Sharifi, 2014). One advantage is that it can handle hard constraints on control inputs and states (Qin & Badgwellb, 2003). The MPC has also become a standard for complex constrained multivariable control problems, such as HVAC (Beltran & Cerpa, 2014). For a long time, the MPC method was not used due to the high computational requirement that comes along with it (Mirakhorli & Dong, 2016). The reason for the high computational requirement is that HVAC optimization is a non-linear problem and an algorithm that guarantees global minima for nonlinear problems does not exist. Finding local optima comes with high computational cost and can be time consuming (Afram, Janabi-Sharifia, & FungaKaamr, 2017).

2.7. **Optimization techniques**

The goal of optimization is to find a set of numerical values that result in the best fit in an equation. For a linear problem this can be done by differentiating the equations with respect to each parameter in turn, setting the set of partial differential equations to zero and solving this set of equations (Swann, 1969). For more complicated problems, such as non-linear problems, finding a best solution can be difficult, since there is no analytical solution that is able to produce the optimal solution (Afram, Janabi-Sharifia, & FungaKaamr, 2017). To solve this problem dynamic programming and gradient methods have been proposed that can find local optima.

Afram et al., 2017 performed a review on eight optimization techniques that are often used together with ANN-MPC. The optimization techniques that were reviewed are:

- Genetic Algorithm (GA)
- Modified Genetic Algorithm or Multi-Objective Genetic Algorithm (MOGA)
- Newton-Raphson method
- Interior-point method
- Branch and Bound (BaB) method
- Particle Swarm Optimization (PSO) algorithm
- Modified or Multi-Objective PSO (MOPSO)
- Strength Pareto Evolutionary Algorithm (SPEA)
- SPEA with Local Search (SPEA-LS)

The most commonly used optimization technique of these eight is the GA. GA can solve both constrained and unconstrained optimization problems. The GA optimization method minimizes its

function without using the derivate and is not restricted to the estimation of uncorrelated parameters (Afram, Janabi-Sharifia, & FungaKaamr, 2017). The downside of the GA is that it is slow in converging complex problems, which makes it less efficient for use in real-time constructions.

The Newton-Raphson method is a quadratically convergent algorithm that has the advantage that it converges in a lower number of iterations. This makes it faster, more efficient and more suitable for real-time application (Afram, Janabi-Sharifia, & FungaKaamr, 2017).

The Interior-point can optimize linear and non-linear convex problems and guarantees an optimal solution.

The BaB-method can be used to optimize discrete systems. The method will always find the optimal solution, can implicitly deal with constraints and is not influenced by poor initialization (Ferreira, Ruano, Silva, & Conceicão, 2012).

The PSO is a stochastic optimization technique that is inspired by flock of birds. The PSO has a well-balanced mechanism to enhance and adapt to the global and local optimization problems and excels in solving single-objective optimization problems with a fast convergence (Kusiak, Li, & Tang, 2010).

A variation of the PSO method, MOPSO, can also solve multi-objective optimization problems. The MOPSO produces a set of non-dominated (Pareto optimal) solutions, since there is no optimal solution for a multi-optimization problem (Asadi, Gameiro da Silva, Antunes, Dias, & Glicksman, 2014). S-MOPSO is the combination of SPEA and MOPSO algorithm and is effective in finding an optimal solution for non-linear multi-objective models, since it combines the global solution search of SPEA with the local optimization of PSO (Kusiak, Xu, & Tang, 2011).

The firefly algorithm is excellent in finding global optima and, compared to the PSO and evolutionary strategy, is more effective, efficient, robust and uses less CPU time to converge (Zeng, Zhang, & Kusiak, 2015).

2.8. **Cost function**

The cost function is a mathematical formula that describes the objective, which the optimizer ties to minimize. An example objective of the cost function could be to reduce energy cost. But sometimes there are multiple objectives, like reducing energy consumption and improving thermal comfort. When there are multiple objectives a weight is attached to each of the objectives. The weight describes the importance of each of the objectives. The cost function serves to stabilize the systems when the cost function can be described by a Lyapunov function. In slow-dynamics systems the

stability of the cost function is not a problem, such that any kind of cost function can be used (Afram & Janabi-Sharifi, 2014). Temperature control is an example of a slow-dynamic and hence stability is not a problem. The cost function can take the form of tracking error, control effort, energy cost, demand cost, power consumption or a combination of these factors (Afram, Janabi-Sharifia, & FungaKaamr, 2017).

2.9. **Results**

The optimization of HVAC control using a data-driven approach is a complex process that has been studied for quite some time. Different approaches and techniques have been used to reach the goal of reducing energy. The level of reduction that has been witnessed deviates. According to a review published by Afram et al., (2017) the energy or cost reduction by using Model Predictive Control (MPC) is in the range of seven to fifty percent. These results are similar to those published by Serale et al., (2018), that combined the results of 82 papers of which roughly eighty percent of the results were obtained by using simulation tools. Lazos et al., (2014) on the other hand stated to witness even energy savings up to eighty percent, without violating the thermal comfort. This kind of energy reduction, up to eighty percent, can be considered very rare and do not give a fair indication of the expected energy reduction. The reduction between seven and fifty percent is also similar to the qualitive analysis of the expected potential energy reduction that is outlined by Oldewurtel, et al., (2010).

2.10. **Conclusion**

Methods to reduce energy consumption of buildings by improving the control of HVAC systems have been studied for quite some time and resulted in numerous cases in which it was possible to achieve a substantial reduction in the energy consumption. The approach to reach the goal deviates between the papers, and each paper uses a different set of techniques. Although most of them have in common that they construct an implementation specifically for a single location. A detailed description of data-driven approaches that can maintain thermal comfort and reduce the energy consumption are still quite rare. A recent review published by Serale et al., (2018) supports this statement, and states that a detailed description for the implementation of advanced HVAC control systems using MPC is still lacking. A structured approach would reduce the effort and know-how to construct a advanced controller system.

3. METHODOLOGY

The general structure in which this research is conducted consists of nine individual steps as displayed in figure 1, where section A refers to the Data Mining part and section B refers to the Control optimization part.

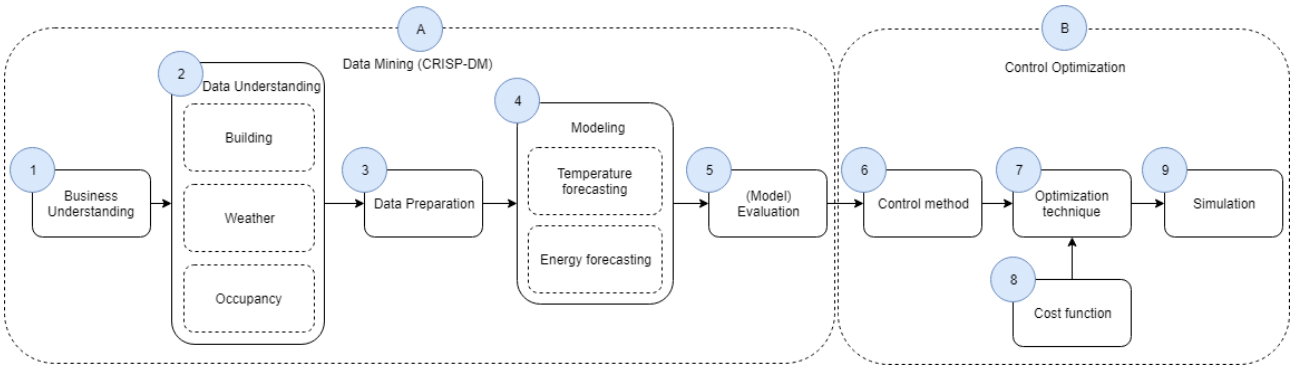


Figure 11 General structure of the research setup.

As mentioned in Methodology paragraph of chapter 1, this research uses the CRISP-DM methodology for the data mining part of the research, since it is the most commonly used methodology within data mining sector. The level in which the methodology CRISP-DM is adopted, can deviate based on the level of complexity. Phases that are quite straight forward are explained on phase level, whereas those that are complex are divided into generic/specialized tasks.

Step 1) The first phase of CRISP-DM, Business Understanding, is performed with a literature study and obtaining information from field experts to understand the current situation of the building. The goal is to get a general understanding of the field in which this research is performed and the building. Temperature control within utility buildings is slightly different from ordinary systems that are used within residential buildings, so it is important to fully understand the control systems and HVAC systems that are used within utility buildings. Afterwards, scientific literature is used to understand what has already been studied within research which has already been performed in the field of HVAC optimization. In addition, this is also used to find the research gaps that are present.

Step 2) The second phase of CRISP-DM, Data Understanding, describes the steps that are taken to gather and evaluate the data from the different sources. This phase is divided into several generic tasks:

- Gathering data: The objective of the gathering data task is to understand which data types are available and relevant. Scientific literature and expert knowledge can provide answers to the question which data types are useful.

- Describing data: The next step is to give an in-depth description of the data. This should contain: the origin of the data, the format in which the data is delivered, which fields are present in the dataset and any other information that is important.
- Exploring data: The objective of this task is to get feeling with the data by studying it more closely. This is useful for the identification of the quality of the data and to identify which steps are required during the preparation phase.
- Verifying data quality: The goal is to determine whether the required data points are also present in the dataset and whether they are of a sufficient quality. In absence of data, or in case of bad quality, alternative sources could be consulted as replacement.

Step 3) The third phase of CRISP-DM, Data Preparation is, as mentioned before, usually the most time-consuming process and mistakes during this phase become even more time consuming to correct in later phases. Therefore, the data preparation phase is split up into specialized tasks. The goal of this phase is to transform the data from the data sources to a combined set that contains only data points that are flagged as useful for the modeling phase of the data and is used to answer the first sub-question of the research question. The specialized tasks are described for each of the generic tasks.

- Data consolidation: The goal of the first generic task is to construct a single dataset that only contains the required data points. This generic task is divided into three specialized tasks.
    - Collect data: The first specialized task is to extract the data from the datasets that have been selected in the previous phase.
    - Select data: The next step is to select the features from the datasets that may contain useful information. The features that are considered useless are removed from the set. These decisions are based on a combination of expert knowledge, scientific literature and obtained knowledge during the previous phase.
    - Integrate data: The last specialized task is to combine the selected data from the different datasets to a single dataset. An important aspect here is to ensure that the timeline is identical over the different datasets, when integrating it into a single set.
- Data cleaning: The goal of the data cleaning task is to remove and repair inconsistencies that are present in the data. Inconsistencies that remain in the dataset have a negative influence on the quality of the models. This generic task is divided into several specialized tasks and is executed by analyzing the data.
    - Impute missing values: The first specialized task is to locate missing data points and repair these by replacing them with realistic data or removing them from the dataset.
    - Reduce noise in data: The second specialized task is to locate and remove noise that is present in the dataset.

    - Eliminate inconsistencies: The last specialized task is to remove inconsistencies.
- Data transformation: The goal of this generic task is to improve the quality of the data. Tools can be used to contribute in the execution of the tasks.
    - Normalize data: The goal of this task is to get a unanimous scale in the data. This is essential for certain modeling techniques to guarantee stable convergence of weights and biases.
    - Discretize/aggregate data: The goal of discretization is to ensure that the features are within a finite state space. Aggregation is to combine two or more objects to a single object. Tools, together with expert knowledge, can be used to reach this goal.
    - Construct new attributes: The last specialized task is to find new features that can have a strong recognition ability with the predictor value. For the construction of new attributes feature extraction techniques can be used.
- Data reduction: The goal of this generic task is to create a sizable set containing relevant features that can be used by the modeling technique to construct the model. This generic task is divided into two specialized tasks. Both are executed by using tools.
    - Reduce the number of variables: The first specialized task is to reduce the number of variables. Scientific literature on HVAC optimization have already shown which variables are considered relevant. This information can be used to construct a preselection. Afterwards, feature selection techniques are used to select only those variables that are relevant.
    - Reduce the number of cases: The second specialized task is to select a timespan over which the data is taken. The size has a direct influence on the construction time of the model and the quality of the constructed model.

Step 4) The fourth phase of CRISP-DM, Modeling, describes the process of the model construction. In our research two models (temperature and energy) are constructed. This process is divided into several generic tasks:

- Selecting modeling techniques: The objective is to select one or multiple modeling techniques that are used to forecast energy consumption and temperature. Important aspects for the selection are the applicability of the technique for our problem. Scientific literature on the differences of modeling techniques and expert knowledge is used to select the proper technique. The results from this task is used to answer the second sub-question of our research.
- Designing test: As a pre-step to building and assessing the models is the selection of techniques used to split the data into training, (validation) and test sets.
- Building model: Modeling techniques usually have certain hyperparameters that can be adjusted for the construction of the model. These parameters have a direct effect on the

quality of the model. A method or technique must be selected to find the right parameter settings.

- Assessing model: The techniques that are used to assess the quality of the models are extracted from scientific literature. Previous conducted researches in this field have already shown which techniques are suitable to assess the quality of models.

Step 5) The fifth phase of CRISP-DM, Evaluation, is used to evaluate the results from the constructed models. The results determine whether the constructed models are sufficient enough to be used in the next phases. Insufficient results would mean reevaluation of the previous phases and might result in adjustments. Comparison with previous studies and expert knowledge is used to determine whether the overall quality is sufficient enough.

The sixth phase of CRISP-DM, Deployment, is not included in our case. The models are just a step of the entire chain, since they are used by the control system. The Deployment step is therefore been replaced with four different steps that are not part of the CRISP-DM methodology: Control method (step 6), Optimization technique (step 7), Cost function (step 8) and simulation (step 9). The control method uses an optimization technique to optimize the cost function. Since they are loosely coupled to each other they are divided into the three separate steps. The simulation step is the final step that is used to evaluate the quality of the control system.

Step 6) The Control method step is used for the construction of the control technique and is used to answer the third sub-question of the research question. The control technique step can be separated into selection and construction, although the selection task is executed even before the construction of the models of step 4. A control technique has a direct effect on the entire process and has therefore been selected upfront. The control technique that is used in this research is Model Predictive Control (MPC). The construction of this technique on the other hand is executed after the model construction phase, since it uses the models. MPC is a well-documented control technique that has been adopted in several frameworks and is available from online libraries.

Steps 7-8) The steps Optimization technique and Cost function are taken together due to their dependency on each other.

- Cost-function: The cost-function formulates the optimization problem and determines the importance of the included aspects. The construction of a cost-function is well documented in scientific literature.
- Optimization technique: The optimization technique is selected based on the flexibility regarding the use of different cost-functions. Discussions on the differences between

optimization techniques are widely available in scientific literature. The results from this task are used to answer the fourth sub-question of the research question.

Step 9) The last phase Simulation is used to verify the quality of the constructed control system. The developed control system is tested by presenting a variety of different situations that are covered in the dataset. The results from the simulation are afterwards compared with those that are stored in the dataset and based on these findings the quality of the constructed control system is determined.

4. **BACKGROUND**

4.1. **Building**

Figure 12 A photograph of the front side of the building taken from the sky.

The building used for this study, displayed in Figure 12, is located in The Hague in The Netherlands and is built in the years 1912 to 1917. The architectural style originates back to the Classicism of the 17th century and was designed by the architects D. Knuttel (1917) and H.J.M. Ruijssenaars (1994). The building was used in the early years by the former Ministry of Agriculture, Trade and Industry (Nowadays Ministry of Economic Affairs). In the years 1991-1994 the building was extended with an additional 6000 square meter to a total of 21 000 square meter. In 2014 the Ministry of Economics was transferred to a different location and the building was entirely renovated. The renovation project was a joint venture of the public and private sector. Facilicom was assigned to execute the renovation and exploitation of the building based on a DBFMO-contract (Design, Build, Finance, Maintain and Operate) for 30 years. The renovation was completed in September 2016 and is now occupied by several organizations:

- Central Planning Bureau,
- The Netherlands Environmental Assessment Agency,
- The Social and Cultural Planning Office,
- The Dutch Data Protection Authority,
- The Council for the living environment and infrastructure.

4.2. **Technical specification**

The temperature in the building is controlled by a combination of systems. There are heat/cooling producers and consumers.

Heat producers – The heat capacity is produced by a combination of various sources. These sources together are responsible for the entire heat supply of the building.

1. District heating (maximum capacity: 350 KW) – Hot water is generated from a centralized location by an external party. Buildings connected to this system can extract heat from the hot water flow and pay based on the amount of heat that they have extracted.
2. Heat pumps (maximum capacity: 440 KW) – The building is provided with two separate heat pumps that produces heat and cooling energy.
3. Variable Refrigerant Flow (VRF) (63 KW) – An advanced heat pump system that produces cooling based on the required amount of cooling energy. The byproduct of this process is heat that is used as heat supplier. This system was explicitly meant to produce cooling energy for the server room stationed in the basement of the building. Unfortunately, the server room was never realized due to cancelation. Resulting in a low heat energy production.

Cooling producers – The cooling load is produced by a combination of two sources.

1. Mono-well (maximum capacity: 417 KW) – Cooling energy that has been stored in the ground below the building during the cold seasons is extracted to serve as cooling.
2. Heat pumps (maximum capacity: 440 KW) – The building is provided with two separate heat pumps that produce heat and cooling energy.

The heat/cooling consumers – There are four HVAC systems that extract heat and cooling.

1. Air handling units (AHUs) – There are in total nine air handling units in the building that are responsible for the ventilation of the entire building, but also consume a significant proportion of the total heat/cooling load.
2. Radiators - There is only a small number of radiators present in the building and they are mainly located in the basement and staircases.
3. Convectors – Each room in the building has a separate convector that is used to reach the desired temperature locally. This system is only enabled when the room is occupied.
4. Floor-heating – Several locations in the buildings are equipped with under floor heating.

It’s worth mentioning that the AHUs are used as main source to regulate the air quality in the building but also influence the internal temperature of rooms. The other sources (radiators, convectors and floor-heating) are used in a specific area to regulate the temperature. The office part of the building is divided into four zones that each have a separate AHU. These AHUs cover the largest part of the

building and are together responsible for approximately 75 percent of the total airflow in the building. The remaining five AHUs cover the restaurant, seminar/work-foyers and atrium.

4.3. **HVAC control types**

In most (large) non-residential buildings a Building Automation System (BAS) is used to provide centralized management to control heating, ventilation, air conditioning, lighting, safety and security to achieve maximum efficiency and comfort (Mirakhorli & Dong, 2016). The controls used by the BAS can be divided into two groups; supervised control and local control.

- Supervised control is the high-level control of the system. It defines setpoints for local controllers to achieve cost-efficient thermal comfort without violating system constraints (Wang & Ma, 2008).
- Local control is the low-level control that controls each component of the system individually. The instructions are sent directly to a single system, which executes these actions (American Society of Heating, HVAC Applications Handbook, 2011).

4.4. **HVAC-systems**

The thermal condition in the building is controlled by a collaboration of several systems. A graphical representation of these systems is displayed in Figure 13. Each of the essential components is briefly discussed. There are some more components present, for instance filters that are less important and therefore are not displayed.

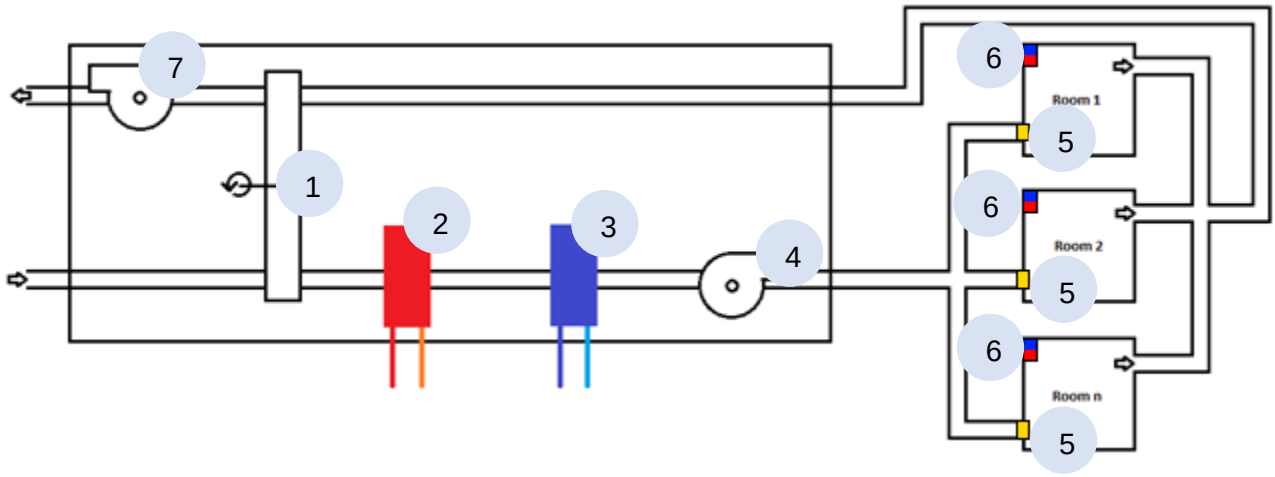


Figure 13 Graphical representation of the combined HVAC systems.

1. The *thermal wheel* is a heat recovery system that is used to reduce energy consumption. The thermal wheel is a large slow rotating wheel that transfers heat between the return and supply sections of the AHU. Air from both sections is directed through the thermal wheel that is made of a high thermal conductive material. Since the thermal wheel is spinning, the heat that is extracted from the return air is afterwards used to heat up the supply air. (During the summer the thermal wheel extracts cooling instead of heat from the return air.)

Alternatives for the wheel used to recover heat from the system are:

a. Plate heat exchanger – works in principle the same as the thermal wheel but does not have any moving parts. The return and supply air are both moving through a coil that exchanges heat during this process.
b. Recirculation – A certain amount of the return air is mixed with the supplied air to extract heat from the return air. This tactic is not often used, since used air is send back in the building may have a higher CO_2 level and odors.

2. The *heating coil* is used to increase the temperature of the air. A hot substance flows through the coils to increase the temperature of the coils. The air that is of a lower temperature passes through the coils and extracts the heat from the coils. Some of the commonly used substances are: water, refrigerant or steam.
3. The *cooling coil* works according to the same principle as the heating coil, only it cools down the air. It is important to note that the heating coil and the cooling coil should never be active at the same time.
4. The *supply fan* regulates the fresh air supply that is stored in the ventilation duct. The fan speed is operated by setting a static air pressure setpoint.
5. The *Variable Air Volume boxes (VAV-boxes)* are used to regulate the amount of fresh air that is extracted from the supply air duct. The amount of fresh air that is extracted from the supply duct is based on the measured air quality that is present in the room.
6. The *convector* is used to regulate the temperature of the room. The convector can be considered as a small AHU without thermal wheel that uses recirculation of internal air to increase or decrease the room temperature.
7. The *return fan* regulates the amount of air that is extracted from the building. The speed of the fan is based on the speed of the supply fan.

4.5. **Data**

The data that is used in this research is obtained from three different sources. Each of the sources is briefly discussed.

4.5.1. **HVAC/BAS**

The data of the HVAC system is obtained from the Building Automation System (BAS). The BAS is the overarching system that controls all the systems that are present and connected in the building. The BAS keeps, based on the configuration, the data of the HVAC system. This data can contain the historical configuration of the HVAC system, the energy usage, the temperature and more.

4.5.2. **Weather data**

The weather data can be divided into two categories; historical weather conditions and expected (near) future weather conditions. The historical weather conditions can be extracted from the KNMI

Data Centre or from the onsite weather station. The KNMI is the meteorological institute of the Dutch government and it provides a free to use service to extract all the weather data that was measured in the past by each of the weather stations. Data about the weather forecasts can be extracted from third party service providers or by a model that can predict future weather conditions based on the current state. There are multiple third-party providers who deliver these services, but the interval and period over which forecasts are given can deviate.

4.5.3. **Occupancy data**

The occupancy rate is based on data that is obtained from the system that operates the entry gates. All personnel and visitors are required to scan their card before entering or leaving the building. The gates open only when a legitimate card is used. This data is used to estimate the occupancy at a certain point of time.

5. THE DATASET

This chapter describes the entire process of the construction of the dataset. The general structure of this chapter corresponds with the second and third step of the CRISP-DM as described in the Methodology (1.3) chapter. The chapter starts with a general view in the data understanding paragraph, in which the data sources are explained together with the available data. Afterwards, in the data preparation paragraph the data preparation steps are explained in detail. Finishing with a brief conclusion on the final dataset that is used to model.

5.1. Data understanding

The general purpose of this chapter is to describe the data sources that are used together with an overview of associated data.

5.1.1. Data sources

The data that is used in this research is obtained from multiple sources, which can be divided into the categories internal and external data. The internal data represents the data that is stored on-site, whereas the external data represents the data that is obtained from an external provider.

Internal data sources

On the internal level there are primarily three sources that are used to extract data from:

- **Building Automation System (BAS):** A BAS can be compared to an advanced thermostat that is commonly used in residencies, where all the HVAC systems and sensors that are stationed in a building are connected to. This includes sensors from an on-site weather station when present. The BAS is the overarching system that monitors and controls all the individual systems that are equipped in the building and therefore also the systems that control the temperature in the building. The building in our case is equipped with Priva, a commonly used BAS in the Netherlands, and is extended with a History package that allows the BAS to store measured sensor values and setpoints of the connected building hardware systems. The History package stores the measured values and setpoint values with an interval of eight minutes. The values that are stored are measured values at that particular moment; fluctuations in between the eight minutes are not registered.
- **Entrance gates:** The entrance gates of the building are connected to the system that keeps track of movements in- and outside the building. Entering or leaving the building is only possible by passing through the gates with an identification pass. Each employee has its own identification pass, whereas visitors have to register themselves before entering the building. Hereby, the system knows at any moment of the day which persons are inside the building and therefore the occupancy rate of the building during the day is known.

- **Blueprints:** Blueprints of the building that were used during the construction and installation of all the systems can contain valuable information. For instance, blueprints of the rooms contain room specific information that can be used to compare rooms, whereas blueprints of the connection of systems help to understand the true meaning of measured values by sensors.

External data sources

The only external data source that is used is a weather forecasting service. The condition of the weather has a direct effect on the consumption of energy by a building. Near future changes in the weather could have a direct effect on future configurations of the systems that regulate the temperature in the building. There are multiple third-party weather service providers on the market that provide the service to forecast the weather for each location. They provide location specific weather predictions, by using the weather information from the weather stations that are stationed across the country.

5.1.2. Data structure

The data extracted from the described sources are structured into the layers: Building, Air Handling Unit and Room.

Building

The first layer, building, contains the more general data that is equal throughout the entire building. This consists of primarily weather-related information and some data on the occupancy rate of the building. The weather conditions are extracted from the on-site weather station that is connected to the BAS and an external weather service provider. The data obtained from the BAS is used to construct the models, whereas the data from the external weather service provider is used for future predictions. An on-site weather station tends to be more precise than external sources that obtain their information from weather stations located nearby. Weather forecasting on the other hand is not incorporated in the on-site weather station and has to be extracted from a third-party weather service provider. The entire list of parameters that can be extracted from the GBS, entrance gates and weather service provider is displayed in Appendix A.

Air Handling unit (AHU)

The second layer contains the data related to the AHUs that are present in the building. The building is equipped with multiple AHUs that each control an entire zone in the building. The primary task for an AHU is to distribute fresh air and remove used air, in order to control the air quality that is present in the building. The fresh air is obtained from outside the building and is first preconditioned to the desired temperature (temperature setpoint). Afterwards, it is stored in the air duct that is connected to all the rooms that are in the zone. Each room extracts fresh air from the air duct based on

measured air quality. The entire list of parameters that can be extracted from the GBS for each of the AHUs is displayed in Appendix B.

Room

The last layer, room, describes the data that is obtained from each room in the building. Each room is equipped with a simplified thermostat. This thermostat is initially used by the user to identify that he occupies the room. The thermostat notifies the BAS and the BAS will switch on the light and activate the convectors to condition the room to the desired temperature. The desired room temperature can be adjusted by the user on the thermostat. During the day, the BAS monitors the room closely and alters settings based on the measured condition. When for instance the air quality of the room decreases, the BAS will increase the flow of fresh air in the room by opening the VAV-box farther. Certain rooms have multiple VAV-boxes and convectors in order to comply to demand. In this case the settings are equal over the same systems. When the movement sensors in the room do not recognize any movement for a certain amount of time, the BAS changes the status back to unoccupied and switches the lights back to off and putting the convectors on hold. In Appendix C, the entire list of parameters that can be extracted from the GBS for each room can be found.

5.2. **Data preparation**

This section describes the data preparation steps according to the CRISP-DM framework as stated in the methodology.

5.2.1. **Data consolidation**

The parameters that are extracted from the data sources are described in detail in this paragraph. The layering structure that is used in the Data understanding (5.1) paragraph corresponds to the structure used in this chapter.

Building

From the entire list of available parameters displayed in Appendix A primarily four types are selected; Temperature, Wind, Humidity and Occupancy. There are a couple of reasons for excluding certain weather parameters:

- Only parameters that are available by both the BAS and the third-party provider can be used.
- Previous conducted research on the effect of weather on buildings has shown that primarily temperature, wind and humidity are major players in consumption and temperature changes.

The resulting selection of parameters is presented in Table 1.

#	Name	Description	Unit	Source
1	OUT-TEMP	Outside – Dry-bulb temperature	°C	BAS/WFS*
2	OUT-WVEL	Outside – Wind velocity	km/h	BAS/WFS*
3	OUT-WDIR	Outside – Wind direction	Deg	BAS/WFS*
4	OUT-HUMI	Outside – Humidity	% RH	BAS/WFS*
5	BD-OCCU	Building – Occupancy rate	persons present per hour	Entrance gates

Table 1 Selected parameters of the building layer.

*WFS = Weather Forecasting Service

AHU

From the entire list of parameters displayed in Appendix B the parameters are selected primarily based on following criteria.

- Only the parameters that have an influence on the room temperature

Parameters that do not comply with this criterion are excluded from the selection. The Supply air duct static pressure parameter is also excluded, since the data has shown that the value is always equal to the given setpoint. The deviations that were witnessed in the data were too small to have any direct effect. The setpoint of the Supply air duct is based on a predefined default scheme.
The resulting selection of parameters is presented in Table 2.

#	Name	Description	Unit	Source
1	AHU-TSET	AHU – Supply air temperature setpoint	°C	BAS
2	AHU-PSET	AHU – Supply air duct static pressure setpoint	Pa	BAS
3	AHU-STMP	AHU – Supply air temperature	°C	BAS

Table 2 Selected parameters of the AHUs layer.

Room

The selection of parameters for the room layer is in principle an extension of the rules that were applied in the AHU layer.

1. The parameter is used to calculate the energy consumption of the convectors.
2. The parameter is used in relation to the room temperature.
3. The parameter is used to compare the temperature behavior in different rooms.

Parameters that do not comply with these criterions are excluded from the selection.
The resulting selection of parameters is presented in Table 3.

#	Name	Description	Unit	Source
1	RM-MTMP	Room – Measured temperature	°C	BAS
2	RM-DTMP	Room – Desired temperature	°C	BAS
3	RM-SITU	Room – Situation	{off, stand-by, comfort}	BAS
4	RM-AIRQ	Room – Air quality	PPM (Parts Per Million)	BAS
5	RM-HEAT	Room – Heating	%	BAS
6	RM-COOL	Room – Cooling	%	BAS
7	RM-VENT	Room – Ventilation	%	BAS
8	RM-VAV	Room – Variable air volume	% open	BAS
9	RM-SURF	Room – surface	M^2	Blueprints
10	RM-TYPE	Room – type	Set of types listed in Table 4	Blueprints
11	RM-CAPA	Room – capacity (People)	N	Blueprints
12	RM-WIND	Room – Windows	N	Blueprints
13	RM-NCNV	Room – Number of convectors	N	Blueprints
14	RM-AHUZ	Room – AHU zone	N	Blueprint
15	RM-GKWZ	Room – GKW&CV zone	{NO, ZW}	Blueprint
16	RM-FLOR	Room – floor	N	Blueprint
17	RM-LOCA	Room – Location	{N, O, Z, W, G}	Blueprint
18	RM-NUMB	Room – Number	{text}	Blueprint

Table 3 Selected parameters of the room layer.

The parameter Room type consists of a finite set of room types, which is displayed in Table 4. The list contains the most common types of rooms that are present in the building. There were curtain types that occurred only once or twice in the building and therefore combined to the type Others.

#	Room type
1	Hall
2	Consolation space
3	Silence room
4	Conference room
5	Workspace
6	Others

Table 4 The types of rooms that are present in the building.

General

In a dynamic system, like the HVAC system, the current system state is significantly influenced by the previous state(s). This means that the values from previous states have a direct influence on the current state. Research conducted by (Kusiak, Li, & Tang, 2010) has shown that there is a clear correlation between the previous temperature and energy consumption with the current. They analyzed the correlation between the past five system states with the current state, and selected the past two states for the construction of the model. By this means, the past two states of the temperature and energy consumption are included in the selection of features. The notations “Nd” or “Nd2” correspond to the first and the second past stage respectively.

5.2.2. **Data cleaning**

Data cleaning is the next step and is separated into the tasks: impute missing values, reduce noise in data and eliminate inconsistencies as described in the methodology.

Impute missing values

There are gaps of missing data observed in the dataset. The size of the gaps varies from hours to several days. Investigation has shown that these gaps are a result of shut downs and not enough storage space leading to loss of data. These gaps do not affect a single value or table but the entire database. Due to this, imputing the missing values becomes an infeasible solution. To prevent inaccuracies, the dataset is divided into multiple datasets, in which the gaps are used as separators.

Besides these gaps, the length of the history varies over different data points. In the BAS each data point has to be configured separately to ensure that it is stored in the history package of Priva. This resulted in certain data points having a different starting date. These different lengths are primarily observed between rooms. Since the goal is to construct a uniform model for all rooms, these different lengths do not affect the overall result, since identical rooms will fill these gaps.

Reduce noise in data

There are incidents in the dataset in which sensors have observed unrealistic values like inhumane room temperatures and extreme PPM (CO_2) levels in an unoccupied room during the weekend while the building is closed. Besides these extreme high values, there are also incidents found when the actual values are not registered and replaced by zero. Although a temperature of zero is physically possible, a drop from twenty degrees to zero in a period of eight minutes is most likely impossible. Each incident is evaluated separately and based on the length in which obscure values were witnessed, the records were either removed or the entire room was excluded from the dataset. These obscure values are rarely unique and affect only a small proportion of the dataset.

Eliminate inconsistencies

As previously mentioned, values are only stored with an interval of eight minutes. However, there are two cycles, that both start four minutes after each other. This inconsistency is resolved by adding four minutes to one of the two cycles to ensure that they are equal in the entire dataset. These four minutes could have negative influence on the accuracy of the data. However, the data is averaged to an interval over a larger period and therefore the influence of the four minutes difference becomes negligible.

5.2.3. **Data transformation**

This paragraph describes the next specialized task as described in the methodology. This consists of the task's normalization, aggregate data and construction of new attributes.

Normalize data

The dataset is normalized to values between values of zero and one. Normalization of data is an important step for neural networks to increase the quality of predictions.

Aggregate data

The data from the BAS is stored in intervals of eight minutes, whereas the prediction horizon is set to 64 minutes. The prediction horizon is determined by analyzing the changes in temperature that are captured in the dataset. Results from multiple analyses have been used to determine the most suitable prediction horizon. A detailed description of the performed analyses can be found in Appendix E.

Transformation of the data based on the prediction horizon is performed by averaging all values within the interval. The average is selected since that would be the best representation of the measured condition between two intervals. with the exception for the predicted values. The predicted values are selected by taking the first measured value of a selected timespan. The main goal is to determine how the predicted value has changed over the timespan.

Construct new attributes

Feature extraction was an important and mandatory step for the prediction of the energy consumption. From the perspective of a room there are primarily two sources that consume energy, the AHU that is connected to the rooms and the convectors located in the rooms. Other aspects as electronic devices and light bulbs that consume and produce heat are excluded from the energy balance. These sources consume and produce only a marginal fraction of the total energy balance. Besides that, there are no sensors that measure the energy consumption or production. The energy used by the AHUs and convectors can be divided into electrical energy to run the systems and heating/cooling energy to alter the air temperature. The electrical energy used by these systems is

substantially higher than the energy used by the light bulbs and electronic devices, but compared to the energy that is extracted from the cooling and heating water it is still only a small fraction on the total energy balance. Just as the electronic devices and light bulbs, the electricity used by the AHUs and convectors in the building are not measured and therefore also excluded. The energy consumption that is measured and used is the energy that is obtained from the heating and cooling water that is used to alter the room temperature. Both the AHUs and convectors consume cooling and heating water to alter the air temperature. The energy consumed by these systems is not specifically measured. The energy usage is calculated by using the heat transfer formula.

The Heat transfer formula:

$Q = M * c * \Delta T$

In this formula Q refers to heat in Joules, c refers to the specific heat, M refers to the mass flow rate and ΔT refers to the difference of temperature. The specific heat is a constant, which is $4186\,\mathrm{J/(kg * K)}$ for water

The energy consumption of the AHUs is calculated by using the temperature of the water before entering and after leaving the AHU, and using the position of the valves that regulate the water flow. Afterwards, the energy consumption is translated on room level, by using the VAV-box openings for each room.

The energy consumption of the convectors is slightly more difficult, since there is not any data available on the entrée temperature and return temperature for each convector particular. There is only data available on the entrée temperature and return temperature of an entire sub-group, that contains roughly fifty percent of all the convectors and some floor heating that are present in the building. The valve regulating the flow for each convector on the other hand is available. Since the exact temperature differential is required, it is impossible to precisely calculate the energy usage for each convector. The closest estimation to measure the energy usage of the convectors is to use the temperature differential over the entire group as the temperature differential for each of the convectors. The average building temperature is quite stable in the entire building, due to the high thermal mass of the building. This means that this estimation will result in quite accurate measurements.

An in-depth description of the energy consumption calculation for the AHUs and convectors is found in Appendix D. It explains the cooling water distribution for a single side of the building. The opposite side of the building works exactly according the same principle. The distribution of heating water for the entire building is calculated in precisely the same way, and therefore not discussed in detail.

From the convectors the cooling and heating power is combined to a single feature, “Room – Convector Setpoint”. A convector can use only one of the two at the time since it is either cooling or heating the air. By combining it, it increases the value of the feature and also reduces the dimensional space.

5.2.4. **Data reduction**

Reduce number of variables

By combining all the features, as described in the Data consolidation (5.2.1), of the Building, AHU and room together, it results in a list of total thirty features. Using all these features for the construction of the model would substantially increase the computational complexity and therefore the process time. Besides that, it may even have a negative effect on the quality of the model, due to lack of correlation. To reduce the number of features a feature selection method has been used. Based on previous conducted research within this field a wrapper method using Extreme Gradient Booster (XGB) has been selected. The XGB is a quick and lightweight tool that contains a feature importance function to determine the quality of the features. A recursive feature elimination technique has been used together with the XGB to select the most suitable features. Multiple runs with each a different parameter setting of the XGB have been used to remove inconsistencies in the suggested feature set. From these runs, only the results were used where the R^2 score of the XGB model was above 0.7. The results for temperature prediction are displayed in Appendix F and the results for the energy prediction are displayed in Appendix G. Based on these results, the following features have been excluded for the temperature model:

- Room energy
- Room energy_Nd
- Room_energy_Nd2
- AHU_supply_temperature
- Room_nr_convectors
- Room_nr_vavs

For the energy model the following features have been excluded:

- Room_desired_temperature
- Room_nr_vavs

Reduce number of cases

Reducing the number of cases in the dataset is useful when the number of cases in the data is too large or data points from the past do not correctly represent the current state. Based on the given situation where outside weather conditions play a vital role in the prediction of the internal temperature and energy consumption, it is beneficial to use a dataset that contains all four seasons of a year, so preferably an entire year of data. The dataset used in this research project contains

roughly ten months of data (March to December), leading to a total dataset of over eight million historical records. The dataset originates from 2017, but the internal situation has remained identical and is therefore still representative for the current situation.

Due to the size of the dataset it was not feasible to use the entire dataset during the feature selection task and parameter optimization. Therefore, the dataset has been reduced to roughly three million data points. The entire dataset is used to construct the final models.

5.3. **Final dataset**

The final dataset that is used in the modeling phase of the CRASP-DM methodology is constructed by processing the data based on the described steps in the Data preparation (5.2) paragraph. The entire list of parameters that is used for the modeling phase is displayed in Appendix H for the temperature model and Appendix I for the energy model.

6. MODELING

This chapter describes the entire process of the construction of the models. The general structure of this chapter corresponds with the fourth step of the CRISP-DM as described in the Methodology (1.3) chapter. This chapter begins with a discussion on the selected modeling techniques. Afterwards, a description is presented on the setting in which the models are constructed. It finishes with a description on the selected evaluation techniques.

6.1. Modeling techniques

Three different modeling techniques from the Scikit-learn package (Pedregosa, et al., 2825-2830) have been used to predict the energy consumption and near future room temperature. The first modeling technique, a Linear Regressor, is used to set a baseline. The remaining two more advanced models, Multi-layer Perceptron Neural Network (MLP NN) and Random Forest Tree (RFT), should at least exceed this baseline to show that it is beneficial to use more advanced modeling techniques for this problem. The reasoning for the selection of these techniques and the techniques themselves are described below.

Linear Regressor

The linear regression modeling technique is one of the most basic modeling techniques. It constructs a single straight line based on the given variables. A single independent variable linear regression is displayed in Figure 14, where the red dots are representing the observation, the x-axis is the independent variable(s), the y-axis is the dependent variable and the blue line between the dots is the underlying relation. The dimension grows proportionally to the number of variables.

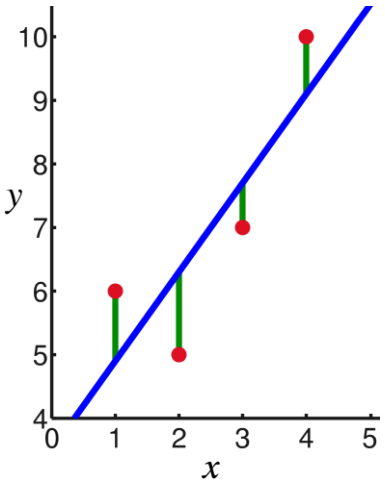


Figure 14 Simple representation of linear regression.

This modeling technique is selected for the simplicity of the technique itself. The main purpose is to use it as a baseline that makes it possible to determine whether it is beneficial to use more advanced techniques.

Multi-layer Perceptron Neural Network

The Multi-layer Perceptron Neural Network (MLP-NN) is a class of a feedforward artificial neural network. An MLP-NN consist of three layers of nodes: input layer, hidden layer and output layer, as displayed in Figure 15.

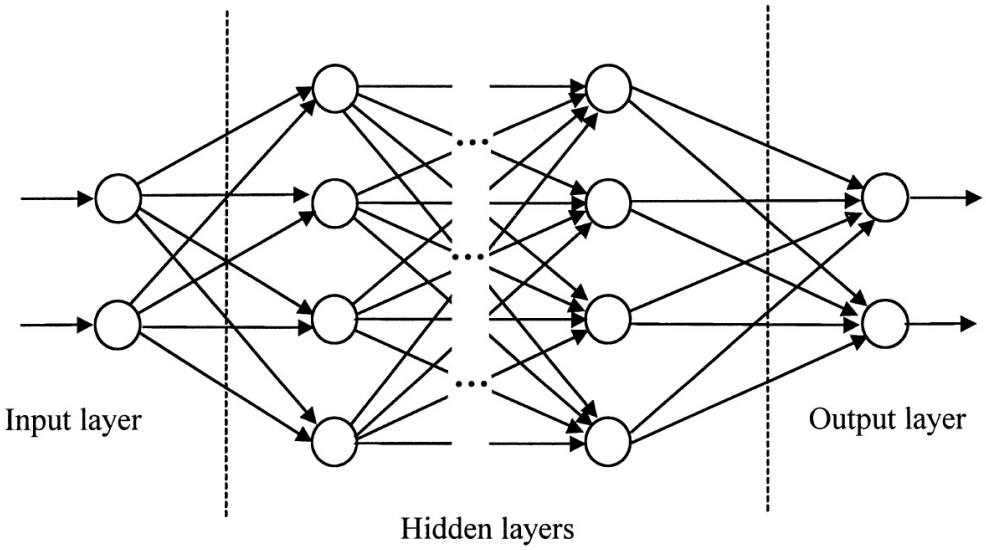


Figure 15 Layering structure of an MLP-NN.

The number of nodes for each layer can deviate.

- The number of nodes in the **input layer** is equal to the amount of input variables.
- The number of nodes in the **output layer** is equal to the amount of output variables (predicted values).
- The number of nodes in the **hidden layer** depends on the number of hidden layers and the number of nodes each layer has. Both the number of hidden layers and the number of nodes each layer contains is configurable.

Each node in a layer is connected to all the nodes of the next layer. For example, the input nodes in Figure 15 are only connected to each node of the first hidden layer. All nodes, except those of the input layer, are considered neurons that use an activation function. The general goal of the neuron is to calculate the weighted sum of all incoming values of the previous layer together with a bias and pass it to the activation function. A graphical representation of this process for a single neuron is displayed in Figure 16.

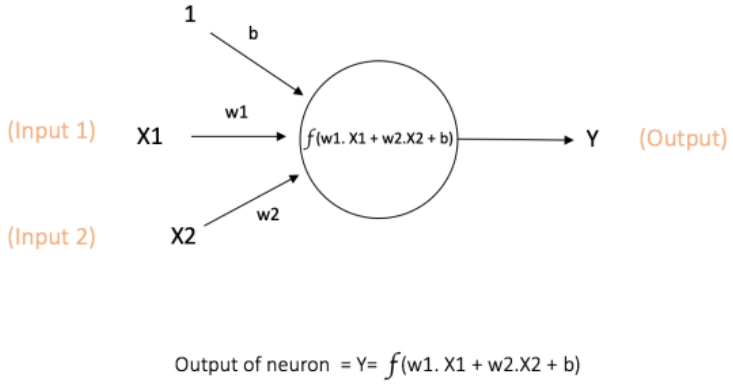


Figure 16 Graphical representation of a single neuron (Karn, 2016).

The characteristics of the activation function is that it is nonlinear and it has a single input value where a fixed mathematical operation is performed over and return the output value of the neuron. Some of the most common used activation functions are:

- **Sigmoid**: It takes any real-value as input and returns a value between 1 and 0.
  $\sigma(x) = 1 / (1 + \exp(-x))$
- **Tanh**: It takes any real-value as input as well and returns a value between -1 and 1
  $tanh(x) = 2\sigma(2x) - 1$
- **ReLU** (Rectified Linear Unit): It takes any real-value and thresholds it at 0.
  $f(x) = max(0, x)$

The bias is a predefined value that can be used to shift the activation function to the left or right. This can be a crucial aspect for the successfulness of the model.

This modeling technique MLP-NN is selected for multiple reasons:

- It is a relatively new modeling technique, but has shown to outperform other techniques in general.
- The MLP-NN is a modeling technique that is capable in handling non-linear problems.
- The MLP-NN performs really well with large datasets. The downside is that it also requires more data in general to perform well.

Random Forest Tree

The technique behind random forest tree is based on a concept of a decision tree. The concept of a decision tree is displayed in Figure 17.

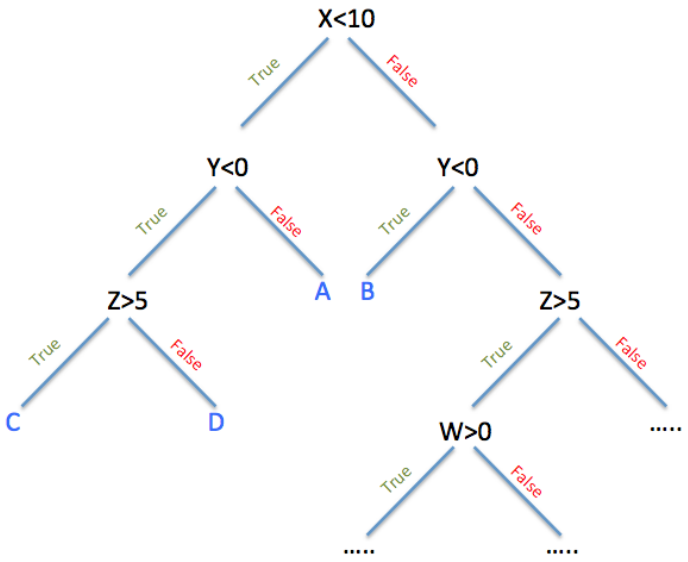


Figure 17 The concept of a decision tree (Cavaioni, 2017).

A decision tree is an upside-down tree with the root at the top. The goal is to predict an outcome based on a set of questions. It starts at the root of the tree and moves through the questions until it reaches an endpoint, which will determine the outcome. In a decision tree there are some commonly used terminologies.

- The black text, for example $Z > 5$, represents the conditions or internal nodes.
- The splits at each condition are referred as the edges; in the given example that is True and False for each condition.
- Each endpoint of the trees is considered the decision or the leaf of the tree. In the given figure noted as the blue capital letter (A, B, C, D).
- The depth refers to the length of the longest path from a root to a leaf.

Random forest tree is an extension on this principle, where it uses a collection of multiple decisions trees, as displayed in Figure 18.

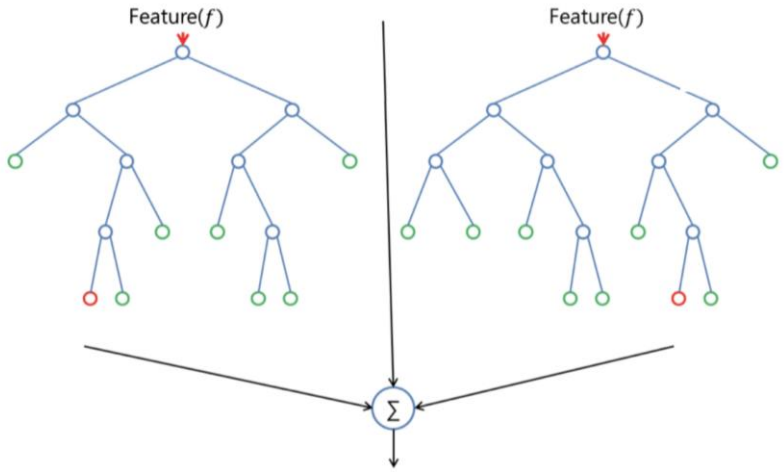


Figure 18 The concept of Random Forest Tree (Donges, 2018).

The general process for the construction of the trees is for random forest tree slightly different and can be described with the following steps (Lan, 2017):

1. The model selects randomly n number of features from the available features D. This number of selected features is usually much lower than the total number of available features.
2. Compute the best splitting points for each tree k using a splitting metric, split the node into daughter nodes and from this node reduce the number features D.
3. Repeat the first and second step until either the maximum tree depth has been reached or the splitting metric reaches some extrema.
4. Repeat the first, second and third step for each tree k that is present in the forest.
5. Calculate the votes for each predicted value and consider the highest vote as the final predicted value.

The primary difference with single decision trees is that the random forest splits on a random selected selection of multiple feature variables instead of a single feature variable. This results in a wide diversity that generally results in a better model. The downside is that the random forest tree needs to construct multi trees and therefore increases the computation time substantially.

The Random Forest Tree has been selected for the following reasons:

- It is one of the most used modeling techniques due to its simplicity.
- It is capable to produce good results even without hyper-parameter tuning.
- It is capable to make non-linear decisions by using only linear questions.
- It is less prone to overfit.

6.2. **Modeling setup**

Modeling techniques can be divided into the types Multi Input Single Output (MISO) and Multi Input Multi Output (MIMO). The difference is that MIMO modeling techniques can predict multi variables with a single model, where a MISO can only predict a single value. The advantage of a MIMO modeling technique is that it can learn from the relation between the predicted values and therefore may score better. The downside of MIMO is the difficulty of evaluating the quality of the model. The quality of the model is based on the prediction of both variables instead of one. A weighing system has to be used to balance the quality of both predictions, and the quality of the weighing system has a direct effect on the quality of predictions for both values.

A MISO structure has been used in our case study because of the following reasons:

- A MISO modeling technique has been used for feature selection and therefore can only determine the quality of the features for each predicted value separately.
- The MLP-NN of the three modeling techniques is the only one that allows MIMO modeling.
- The difficulty of obtaining a proper weighting system. Using an improper weighing system has devastating effects on the overall quality of the model.

Cross-validation

Another aspect is the partitioning of the dataset into a training, validation and test set. Using the same data for the construction and validation of the model would be a mistake that leads to overfitting the model. The model is already familiar with the test samples and therefore perfectly able to predict the correct answer. Afterwards when new unseen data is used the model would perform far less. It is therefore common to divide the data in a training, validation and test set. The **training** set is used to train the model, the **validation** set is used for tuning the hyperparameters of the modeling technique and the **test** set is used to validate the quality of the model. The downside of dividing the dataset into these three groups is that the model is not able to use all the data to train the model, resulting in model of a less quality. Cross-validation is a method that is used to partition the data into multiple different training, validation and test sets, in which each partition is evaluated separately and the overall quality of the model is determined on the performance of the different partitions. The higher the number of partitions, the more accurate the performance can be qualified. The downside is that it is a time-consuming process, since each partition requires the construction of a model. The models in this research project are validated using five partitions.

The selection of the data for the training, validation and test samples is based on the selected iterator. There are multiple commonly used Cross-validation iterators available. The data used in this research project can be categorized as a time series due to the ordered sequence of observations. In time series, the time aspect often determines the relevance of the data. The more recent the data,

the better it represents the actual state. Due to this, it is important that the quality of the model is determined based on the most recent data. Time series split is an iterator that achieves this. The chapter Model quality evaluation methods(2.5) contains an in-depth discussion on the time series split iterator. The data can be categorized as a time series split, however a shuffle split iterator is used. There are two primary reasons for the selection of the shuffle split iterator over the time series split.

- First, the quality of the model should be determined based on the prediction quality of all the seasons (winter, autumn, summer and spring) and therefore it is important that the training, validation and test set should include all the different seasons.
- Second, the situation inside the building has not been altered over the entire period of the data and therefore all the data points in the dataset are equally relevant.

6.3. **Hyperparameter optimization**

Most modeling techniques have adjustable parameters that can limit or control the learning behavior of the model and therefore have a direct effect on the quality of the model. The selection of the configuration of parameters is not a trivial process, in which following a roadmap leads to the most suitable selection of parameters. The general process of selecting the most suitable parameters is by constructing and validating multiple models with different parameters. Afterwards, the results are compared and the parameters of the model with the best results are selected. Since there is theoretically an infinite amount of possible variation of parameter, it is not feasible to try all variations. Therefore, it is common to use a parameter-grid that contains a selection of parameter configuration.

The used modeling techniques are MLP-NN, RTF and linear-regression, of which the MLP-NN and RFT have multiple adjustable parameters. For both models a parameter grid is used to determine the most suitable parameter configuration.

Multi-layer Perceptron Neural Network (MLP-NN)

The variation of parameters used for the MLP-NN is displayed in Table 5. Only the parameters in which multiple configurations have been tested are listed. The remaining parameters have been untouched and are set to a default value. The number of hidden layers is selected based on the following principles. An MLP-NN can process linear data without any hidden layers and needs at least a single hidden layer to process non-linear data. A neural network with over three hidden layers is considered deep. These types of neural networks are known to be hard to train, since adjusting the input weights often results in the solution getting stuck in local minimum (Larochelle, Bengio, Louradour, & Lamblin, 2009). Besides that, increasing the depth also increases the computation time. Therefore, only one and two hidden layers have been tested. The number of neurons for each hidden layer is selected by using the baseline: $Number\ of\ neurons = (input + output)/2$, where a

margin is included of +/-2. The activations are selected based on an article published by Sharma V, (2017). The article stated that the activation ReLU works most of the time as general approximator. Besides that, the activation Tanh is a very popular and widely used activation. Therefore, the activation ReLU and Tanh have both been selected.

Parameter	Settings
Number of hidden layers	1 and 2
Number of neurons	9 to 13
Activation	ReLU and Tanh
Learning rate	0.001 and 0.005

Table 5 Grid of parameters for the MLP-NN.

Random Forest Tree

The selection of parameters for the random forest is based on a random selection of parameters, resulting in the parameters displayed in Table 6.

Parameter	Settings
Number of trees	100
Max depth	None, 5, 10, 15
Min sample split	100, 1000
Min sample leaf	10, 100
Max features	Auto, 10

Table 6 Grid of parameters for the Random Forest Tree.

6.4. **Model assessment techniques**

The model assessment techniques that have been used for this thesis are RMSE, MAE and R^2. Each of the measurements has its advantages and combining them will give a good representation of the quality of the constructed models.

- The **MAE** has the advantage that it can be interpreted by non-experts and gives a clear indication how accurate the predictions are in general.
- The **RMSE** has the advantage that is gives a higher penalty for large errors. The degree of deviation will have a direct effect on the thermal comfort and energy consumption, and is therefore an important measurement.
- The $\boldsymbol{R^2}$ returns a value between 0 and 1 and can be used to determine the general quality of a model. The cut off point for a model is around 0.6, but a good model should at least be around 0.7-0.8.

7. MODEL EVALUATION

This chapter describes the process of the evaluation of the constructed models. This chapter corresponds with the fifth step of the CRISP-DM as described in the Methodology (1.3) chapter. This chapter evaluates the temperature and energy models separately; it begins with a brief discussion on the results of the Linear Regressor and afterwards discusses the advanced modeling techniques. The chapter finishes with a joint conclusion of both the temperature and energy models.

7.1. Temperature

As previously mentioned, a Linear Regressor with fivefold cross-validation is used to determine the baseline. The individual results of the five models are displayed in Table 20 of Appendix J, whereas the average results are displayed in Table 7.

Measure	Score
Fit time	30.97
Score time	1.22
Test MAE	0.29
Test R^2	0.90
Test RMSE	0.40
Train MAE	0.29
Train R^2	0.90
Train RMSE	0.40

Table 7 Quality results temperature model using Linear Regression.

Based on these results it can be observed that:

- The R^2 score of 0.90 is quite high, indicating that the Linear Regressor itself is already capable to accurately model the predicted behavior, it also means that there is a good relation between the predictors and the predicted value.
- As expected with a high R^2 score the mean absolute error is quite low. Based on the MAE it can be concluded that on average, the predicted value is 0.29 off the actual temperature. By comparing this to the average measured temperature in the building, which is 21.7, it has an offset of roughly 1.3 percent.
- The RMSE of 0.40 means only small fluctuations are witnessed in the predicted values.
- The simplicity of the model can also be observed in the time it takes to construct and evaluate the model.
- By comparing the scores between the training and test set it can be noted that they are equal. It is sometimes considered that a low difference is an indication that the model is not overfitting, since the model is equally capable to predict on seen and unseen (new) data.

The evaluation of MLP-Neural Network (MLP-NN) and Random Forest Tree (RFT) is based on the same principle, where fivefold cross-validation is used to determine the quality of the models. The individual results of the five models are displayed in Table 21 and Table 22 of Appendix J, whereas the average results are displayed in Table 8. Based on these results compared to baseline results of the Linear Regressor it can be stated that:

- Both the MLP-NN and RFT models perform above the baseline R^2 score, with both nearly a perfect score.
- The MAE and the RMSE scores in both cases are roughly three times as good as the baseline model. Resulting in an offset of about 0.5 percent
- As expected, the fit and score time is significantly larger, due to their complexity.

Measure	MLP-NN	RFT
Fit time	395.21	3329.63
Score time	3.10	223.15
Test MAE	0.10	0.08
Test R^2	0.98	0.99
Test RMSE	0.18	0.15
Train MAE	0.10	0.07
Train R^2	0.98	0.99
Train RMSE	0.18	0.14

Table 8 Quality results temperature model using MLP-Neural Network and Random Forest Tree.

It can be stated that the MLP-Neural Network and the Random Forest Tree perform both significantly better than the baseline model. When comparing the two advanced modeling techniques it can be stated that:

- The fit and score time of the MLP-Neural Network is roughly nine times faster in fitting the model and about seventy times faster scoring compared to the Random Forest Tree.
- Only a tiny deviation is witnessed in the MAE, RMSE and R^2. Meaning that both models are equally good in the prediction of the temperature.
- The scores on the training and test set are exactly identical for the MLP-Neural Network and for the Random Forest Tree only the MEA and RMSE have a tiny offset of 0.01. Hereby it is considered that it is unlikely that the models are overfitting on the data.

7.2. **Energy**

The evaluation of the energy models is processed in the same way as the temperature models. First a Linear Regressor with fivefold cross-validation is used to determine the baseline. The individual results of the five models are displayed in Table 23 of Appendix K, whereas the average results are displayed in Table 9 .

Measure	Score
Fit time	27.93
Score time	1.81
Test MAE	496.79
Test R^2	0.37
Test RMSE	1381.12
Train MAE	497.03
Train R^2	0.37
Train RMSE	1381.45

Table 9 Quality results energy model using Linear Regression.

Based on these results it can be observed that:

- The R^2 score is tremendous low, meaning that the Linear Regressor is not able to find a relation between the predictors and predicted value. It is reasonable to believe that the problem is non-linear and therefore impossible to comprehend by the Linear Regressor.
- The poor quality of prediction is also observed in the MAE, with a score of about 497 WH. To put it into perspective, the average measured energy usage is equal to roughly 510 WH.
- The time to fit and score the model is still quite low, but compared to the temperature model already twice as much.
- The quality results on the training and test set are about equal and therefore it is considered that the model is not overfitting.

Assuming that the poor quality of the linear model is due to the present of a non-linear relation between the predictors and the predicted, the advanced modeling techniques that are able to process non-linear relation may perform much better. The evaluation of MLP-Neural Network (MLP-NN) and Random Forest Tree (RFT) is based on the same principle, where fivefold cross-validation is used to determine the quality of the models. The individual results of the five models are displayed in Table 24 and Table 25 of Appendix K, whereas the average results are displayed in Table 10.

Measure	MLP-NN	RFT
Fit time	2552.70	2947.879
Score time	4.28	83.47333
Test MAE	362.12	314.0391
Test R^2	0.60	0.664768
Test RMSE	1104.12	1004.64
Train MAE	361.97	288.2327
Train R^2	0.60	0.716336
Train RMSE	1103.52	923.8087

Table 10 Quality results energy model using MLP-Neural Network.

Based on these results compared to baseline results of the Linear Regressor it can be stated that:

- The R^2 score compared to the Linear Regressor is better, but still not within an acceptable region.
- The MAE has already decreased to 362 WH, but that is still an offset of 70 percent of the average.
- The quality results on the training and test set are again about equal and therefore it is considered that the model is not overfitting.

Both models, temperature and energy, are constructed using the same dataset, making it useful to also compare the results of the energy model with the temperature models. A couple of differences can be witnessed:

- The time to fit an MLP-NN model is five times larger for the energy model, which might indicate that the energy model is unable to converge properly.
- The suspicion of the converging problem is also strengthened by the R^2 score of the energy model. The R^2 score of the energy model is roughly 0.35 lower.

Based on the results shown of the energy model and temperature model, it is clear that there is a problem with the dataset to predict the energy consumption. There are multiple, possible reasons for this observed behavior. Three of the most common are listed below:

1. The formula and assumptions used to calculate the energy consumption are insufficient and inaccurate, leading to obscure values.
2. Important aspects that have a direct effect on the energy consumption are missing in the dataset, resulting in inaccurate predictions.
3. The dataset is polluted with high amount of noise, which prevents the model from converging.

Research into this problem by looking at the three suggested reasons has shown that there is a problem with the distribution and calculation of the energy consumption of the AHUs.

The first problem is that the formula used to charge the energy consumption of the AHUs to room level is inaccurate during the early hours of the day. The general purpose of the formula is to distribute the total energy consumption over all its connected rooms that are occupied and have a VAV-box that is open. The problem occurs during the early hours of the day, when there are only one or two rooms extracting air from the air ducts of the AHU. This means that this room would be charged with the entire energy consumption of the AHU even with the smallest opening, since it is the only room that is extracting air. A possible solution to this is to set a hard cap on the charged energy consumption, preventing it from charging all energy to a single room.

The second problem is that there are a large number of incidents where the AHU consumes energy only during the early hours of the day and uses no energy in the remainder of the day. In general, this could be explained due to the effect of the thermal wheel in the AHUs. The thermal wheel extracts the heating energy from the outgoing air and uses this to preheat the incoming air. During the early hours when the outgoing air is not sufficient enough to preheat the air, the heat sources are used to alter the temperature, resulting in a spike of energy usage of the AHUs during these hours. Although, field experts stated that the thermal wheel is not sufficient enough on its own and there is always energy required from additional sources. This leads to inaccuracies, which affect the quality of the model.

Based on these finding it was evident to believe that expelling the AHUs from the energy consumption formula would improve the quality of the predictions. The downside of the removal of the AHUs from the energy consumption formula is that it becomes infeasible to measure the effect of adjusting the setpoints. This would mean that the temperature setpoint of the AHUs becomes an input variable instead of a controllable input variable, which makes the convector the only controllable input. Since the convector is the primary source to alter the room temperature it still remains effective to optimize this.

After the removal of the AHUs energy from the formula the models have been reconstructed using the same settings. The Linear Regressor is used to reset the baseline for the advanced modeling techniques. The individual results of the five models are displayed in Table 26 of Appendix L, whereas the average results are displayed in Table 11.

Measure	Score
Fit time	69.56
Score time	2.50
Test MAE	486.43
Test R^2	0.37
Test RMSE	1360.36
Train MAE	486.54
Train R^2	0.37
Train RMSE	1357.82

Table 11 Quality results energy model using Linear Regression.

Based on these results it can be observed that:

- The R^2 score has remained identical even with the modified energy formula. It is still possible that the problem is non-linear and therefore impossible to comprehend by the Linear Regressor.

- The poor quality of prediction is also observed in the MAE, with a score of about 486 WH. With the new formula the average measured energy usage is reduced to 495 WH, meaning that the offset remains roughly identical (about hundred percent).
- The time to fit somehow increased slightly; there is no explanation for it.
- The quality results on the training and test set are about equal again and therefore it is considered that the model is still not overfitting.

Assuming that the poor quality of the linear model is due to the presence of a non-linear relation between the predictors and the predicted, the advanced modeling techniques that are able to process non-linear relation may perform much better. The evaluation of MLP-Neural Network (MLP-NN) and Random Forest Tree (RFT) is based on the same principle, where fivefold cross-validation is used to determine the quality of the models. The individual results of the five models are displayed in Table 27 and Table 28 of Appendix L, whereas the average results are displayed in Table 12.

Measure	MLP-NN	RFT
Fit time	2803.93	5057.47
Score time	4.68	103.29
Test MAE	348.06	301.44
Test $\mathbf{R^2}$	0.61	0.67
Test RMSE	1075.21	977.58
Train MAE	348.15	277.24
Train $\mathbf{R^2}$	0.61	0.72
Train RMSE	1073.81	900.28

Table 12 Quality results energy model using MLP-Neural Network and Random Forest Tree.

The following observation can be made by comparing these results with those of the baseline:

- The R^2 score, compared to the Linear Regressor has improved for both the MLP-Neural Network and the Random Forest Tree, but both are still not within the desired region.
- The MAE has already decreased to 301 WH for the Random Forest Tree and 348 for the MLP-Neural Network, leading to an offset of approximately 65 percent of the average.
- The quality results on the training and test set are again about equal and therefore it is considered that the model is not overfitting.

The modification of the formula has resulted in identical results, and therefore it can be concluded that the scale of the problem is larger than expected. Further analysis on the cause of this problem has been unsuccessful so far. Although, it is reasonable to believe that the assumptions made to calculate the energy usage on room level are not as accurate as expected, resulting in values that are too far off from the reality and the model is unable to find a correlation.

7.3. **Discussion**

The results presented in this chapter have shown that the near future room temperatures can be predicted accurately, whereas the prediction of energy usage on room level is of insufficient quality. Different angles to calculate the energy have been used, but all attempts to model the energy consumption have shown to be unsuccessful. Therefore, it becomes impossible to use these models as reference to select the setpoints for the model predictive control. The quality of the temperature model on the other hand is within the acceptable region. Since the temperature model does not require the energy model as input to predict. An alternate control strategy can be used that functions explicitly on the temperature model. Two of the input variables of the temperature model are the ventilation speed and the position of the convectors valve. The valve position directly regulates the amount of heating or cooling water that flows through the convector, which determines the energy usage of the convector. The energy consumption of the ventilator is based on the speed of the fan. However, the energy consumption of the ventilator is negligible in comparison to the energy consumption from the heating and cooling water. So, primarily reducing the overall flow will result directly in a reduction of energy usage. The opening of the valve and fan speed can be used as objective for the control strategy. The downside is that quantitative analysis on the reduction of energy becomes impossible.

8. CONTROL SYSTEM

This chapter contains the steps: Control method (step 6), Optimization technique (step 7) and Cost function (step 8) of the Methodology. These steps have been presented in a single chapter since they can all be related to the control method. This chapter starts with an in-depth description of the construction of the control method. Followed by a description of the used optimization technique and cost function.

8.1. Control method

The control method that has been used in this thesis is Model Predictive Control. The theory behind MPC is described in chapter 2.6 of the Literature Study. For the implementation, the GEKKO Optimization Suite (Beal, Hill, Martin, & Hedengren, 2018) has been used. An in-depth description of the MPC structure in GEKKO is presented in Appendix M. This chapter describes the process of the selection of the model and the configuration of the MPC structure.

Model

The MPC structure uses the predictions of the model for the selection of the control settings. As described in the Modeling chapter there are three different modeling techniques used (Linear Regressor, Random Forest Tree and MLP-Neural Network). The MPC requires only a single temperature model, meaning that a model has to be selected. There are a couple of aspects that played a vital role in the selection of the model.

- An important aspect, which is also a prerequisite, is that the model is of a good quality. The predictions from the model are indirectly responsible for the selection of the control settings. Inaccurate predictions would result in a poor control selection. Therefore, the quality of the model is a leading factor.
- The speed of prediction is subordinate to the quality of the model, but still an important aspect. The time for the MPC to construct its model is primarily affected by the speed of predictions by the model. By considering that the MPC has to redo its process at every time step for each room in a building, this means that the prediction speed becomes an essential aspect.
- The GEKKO Optimization Suite only supports a limited amount of modeling types. Therefore, only modeling techniques supported by GEKKO become viable.

By comparing the results of the model as described in the Model Evaluation chapter, it can be concluded that the Random Forest Tree and the MLP-Neural network both perform roughly three times better in the prediction quality compared to the Linear Regressor. As previously described, the prediction quality is a crucial aspect, meaning that the Linear Regressor is excluded. The quality of the predictions for the Random Forest Tree and the MLP-Neural Network are relatively identical. The

Random Forest Tree performs only a fraction better. When comparing the prediction time between the Random Forest Tree and the MLP-Neural Network it can be concluded that the MLP-Neural Network predicts significantly faster. The difference in quality of prediction does not outweigh the speed of prediction, meaning that the MLP-Neural Network is the most suitable model. The modeling technique MLP-Neural Network is also supported by the GEKKO Optimization Suite.

Model Predictive Control

As mentioned in chapter 2.6 of the Literature Study, the MPC structure requires a Prediction Horizon, Control Horizon (optional) and a sample time. The timespan of each of these three-time instances are selected based on the following reasons.

- The **sample time** in our case study has been set to 15 minutes. There are a couple of reasons that have led to this decision.
    - The minimum time in which data can be extracted from the GBS is equal to eight minutes, limiting it to a minimum of eight minutes.
    - The MPC structure is room specific and the building contains roughly 250 rooms. This can become computational expensive to do for each room every eight minutes.
    - The situation or the configuration of the desired temperature can change at any time, so increasing the sample time would have a negative effect on the reaction speed of the system to include the changes.

  The 15-minute interval might be reduced to 10 minutes, if during implementation and evaluation of the systems a necessity to do so is shown.
- The **prediction horizon** for our case has been set to two hours. Because, with the possible room changes (desired temperature and situation) of the room, it becomes useless to predict the temperature for the entire day ahead. On the other hand, due to the thermal capacity of the building temperature, adjustments take a substantial amount of time to witness. Therefore, the prediction horizon has been set to twice the model prediction time.
- The **control horizon** has been set equal to the prediction horizon. The short prediction horizon already minimizes the amount of sample time instances and keeps the computational time within acceptable ranges. Therefore, it is not necessary to use a smaller control horizon.

Dynamic Equation

In our case study the temperature is predicted one hour ahead, while the sample time is set to 15 minutes. Therefore, a temperature trajectory is used to estimate how the temperature changed over the hour. At each sample time the expected temperature from the trajectory is used as the new room temperature. The trajectory is constructed with the dynamic equation that is part of the MPC structure. It is logical to assume that the effect of convectors is non-linear, but since the effect is quite small due to the high thermal mass of the building a linear trajectory has been used.

The formula used to calculate the trajectory for our case is set to:

$$60 * \frac{dx}{dt} = -x + \hat{x}(t + d)$$

Target objective

A part of the general objective of the MPC structure is to minimize the distance between the predicted state and the target/setpoint. The target can either be a single setpoint where the objective uses an l2-norm error or a dead-band where a setpoint range is used with an l1-norm error.

In the current situation the system uses a dead-band (or dead-zone) structure when a room is unoccupied in order to keep the room temperature within acceptable temperatures and it uses a single setpoint that is set to the desired temperature when the room is occupied. In a new situation it will be replaced with a dead-band for both an unoccupied and an occupied room. The dead-band size when the room is unoccupied will remain identical and a dead-band with an allowance offset of 0.2 degree Celsius respectively from the desired temperature when the room is occupied. The main reason for using a dead-band is to prevent the system from continually over-correcting, which results in a constant switch between heating and cooling. This would result in an inefficient solution that increases the energy consumption and reduces the thermal comfort.

8.2. **Optimization technique**

In the GEKKO Optimization Suite there are primarily three solvers APOPT (Advanced Process OPTimizer) (Hedengren, Mojica, Cole, & Edgar, 2012), BPOPT and IPOPT (Interior Point OPTimizer) (Wächter & Biegler, 2006). None of these have been mentioned in the Optimization techniques paragraph of the Literature Study, but even so the three listed solvers are viable. The BPOPT and IPOPT work according to the Interior-point method, whereas the APOPT solver works according the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. The BFGS is a Gradient Based algorithm that uses a hill-climbing optimization technique that seeks a stationary point using an approximated Hessian matrix of the problem. The APOPT requires less computational power to converge compared to the BPOPT and IPOPT. The speed of the prediction is an essential part and therefore is the APOPT selected.

8.3. **Cost function**

As stated in our research question the general goal in our case study is to maintain thermal comfort while reducing the energy consumption. To achieve this, a multi-objective cost function has to be constructed that incapsulates the thermal comfort and the energy consumption into a single formula. The cost function can be broken down into three objectives.

1. Maintaining or improving thermal comfort.
2. Reducing the energy consumption by reducing the consumption of cooling and heating water.
3. Reducing the energy consumption by reducing the ventilation speed.

Weights are attached to each of these objectives to prioritize them. In our case study the thermal comfort is superior to the energy consumption, meaning that the thermal comfort has the first priority and highest weight. The energy consumption is based on the consumption of heating or cooling water, and the consumption of energy on the ventilation. Since the energy consumption of the fans is only a small fraction compared to the energy consumed by the heating of cooling water, the penalty of using the fan is substantially lower. Based on the presented objectives the following cost function is constructed.

1) $\min_{c_{1(t+d)},c_{2(t+d)}} (w_1 e_{hi} + w_1 e_{lo} + |c_1| w_2 + c_2 w_3)$

In this formula the e_{hi} and e_{lo} are used to manage the thermal comfort, therefore they also share the same weight w_1. The violation of thermal comfort is measured as follows.

2) $e_{hi} \geq \hat{x}(t+d) - y_{t,hi}$

3) $e_{lo} \geq y_{t,lo} - \hat{x}(t+d)$

In these formulas the $y_{t,hi}$ and $y_{t,lo}$ denote the dead-band region, where as $\hat{x}(t+d)$ represents the predicted temperature. The predicted temperature is obtained from the temperature model that formulated as follows.

4) $\hat{x}(t+d) = f_1(c_1(t+d), c_2(t+d), x(t), x(t-d), x(t-$
$2d), v_{y1}(t), v_{y2}(t), v_{y3}(t), v_{y4}(t), v_{y5}(t), v_{y6}(t),$
$v_{y7}(t), v_{y8}(t), v_{y9}(t), v_{y10}(t), v_{y11}, v_{y12}, v_{y13}, v_{y14}, v_{y15}, v_{y16}, v_{y17})$

The symbols used in this formula correspond to the features displayed in Appendix H.

The controllable variables are subject to boundaries in which they can operate. The valve opening c_1 can be operated percentage wise in both directions (cooling and heating). The negative percentage refers to cooling whereas the positive percentage refers to heating. This results in the following formula.

6) $-100 \leq c_1(t+d) \leq 100$

Since the consumption of cooling or heat is penalized equally, the absolute value is used as displayed in formula 1.

The Fan speed can be controlled percentage wise from 0 up to 100. Although, the maximum fan speed c_2 is lower in most rooms due to noise level regulations. Therefore, the maximum fan speed is set equal to $c_{2,max}$.

7) $0 \leq c_2(t+d) \leq c_{2,max}$

To prevent imbalance between the heating/cooling valve and the fan speed, or situations in which either one of the two is activated the following rules are implemented.

8) $\frac{1}{5}|c_1(t+d)| \leq c_2(t+d)$

9) $|c_1(t+d)| \geq \frac{1}{5}c_2(t+d)$

This ensures, first of all, that either none or both are activated. Besides that, it also ensures that a maximum offset of 80 percent can be witnessed.

9. SIMULATION

The simulation phase is described in this chapter and represents the final step of the Methodology. This chapter starts with the description of the simulation method and afterwards the results from the simulations are discussed in detail.

9.1. Simulation Method

The dataset that has been used for the construction of the models is also used for simulation purposes. The simulation timespan that has been used is equal to an entire day, starting at 0:00 and finishing at 0:00 the next day. Each day primarily consists of three phases:

- The initial phase, when the building is still closed and the objective of the system is to keep the room temperature within acceptable ranges.
- The second phase is the entire time the building is operational and people can access the rooms.
- The third and last phase is equal to the initial phase, when the objective of the system is to keep the temperature within acceptable levels.

The initial state of the simulation is set equal to the state that is captured in the dataset. Afterwards the room temperature and control settings of the convector(s) in the room are entirely based on simulation results. All other aspects, such as the outside weather condition, are extracted from the dataset. In a real-life scenario these values would have been available as well. The main difference with a real-life scenario is that the measured temperature would probably deviate from the predicted temperature, due to inaccuracies in the model. But these inaccuracies would only have a minor impact on the system, since the system recalibrates the room temperature at each sample time instance (15 minutes).

With a simulation time of a workday the MPC will construct a total of 96 separate models with each model representing 15 minutes. From each constructed MPC model only the control settings of the first sample time are actually used. By combining these control settings, the settings over the entire day become visible. For the evaluation of the simulated model only the graph with the combined results are presented and discussed in detail. Only a small selection of the individual models that are used to construct the graph is included in the appendixes as an illustration.

During the analysis of the historic data it was noticed that the capacity of the convectors in certain rooms were below the required minimum to bring the temperature to the desired state. Further investigation showed that the convectors were primarily lacking cooling capacity and that the maximum fan speed of the convectors was restricted in most of the rooms reducing the cooling capacity of the convectors even further. Two examples from different rooms, room V2003 and room V4020, in which this phenomenon occurs are presented in Figure 19. As shown in both of the cases,

the convector is cooling at maximum capacity (positive convector setpoint is heating and a negative value is cooling), but the room temperature is still increasing instead of decreasing as would be expected with a convector that is cooling at max capacity. In this figure the situation line (green) indicates the occupation of the room. At a value of 15 the room is unoccupied, whereas the room is occupied when the value is equal to 16. (Note: both values are increased with 14, such that they are still visible in the reduced y-axis scale).

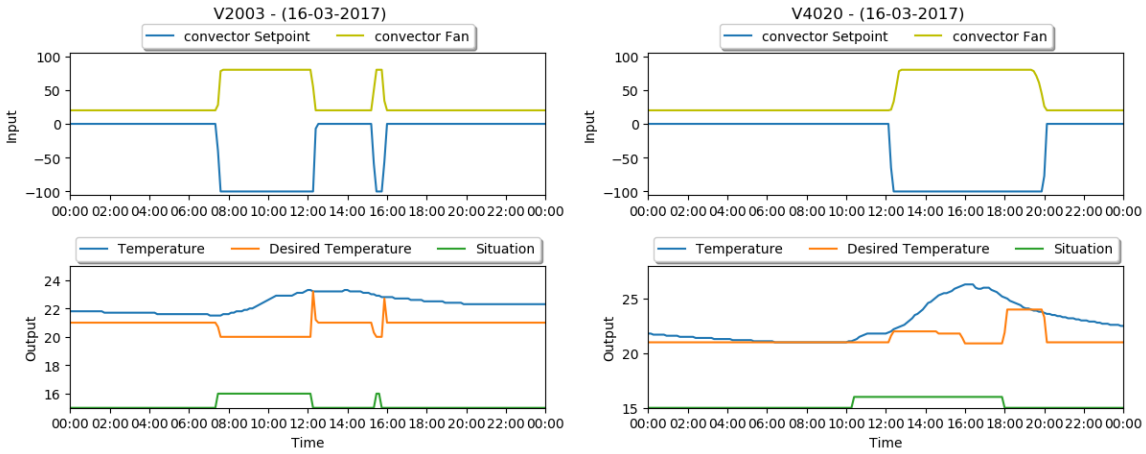


Figure 19 Two examples displaying the effectiveness of the convectors.

The feature importance function from the Random Forest Tree supported these findings and indicated that the level of importance of the features related to the convector were placed below the weather-related features and AHU related features. Therefore, it is evident to assume that there is a lack of capacity for the convectors in general and that the capacity decreases even further for cooling purposes.

9.2. **Simulation Results**

The room that is used for evaluation is selected based on the effectiveness of the convector(s). For the selected room a cold and a warm weather day is used for the simulation. Some general information on the selected room is presented below:

- Room number: V4014
- Floor: 4 (top floor)
- Room type: Workroom
- Workplaces: 2
- Room size: 16.5 M^2
- Number of convectors: 2
- Location: North side
- Maximum fan speed: 60

Scenario 1

The first selected scenario is a cold weather day (29 November 2017) when an average outside temperature was measured of 5.8 °Celsius. The graph on the left side of Figure 20 displays the historic data of that day, whereas the graph on the right shows the results from the simulation. The individual results of the MPC structure between 9 and 12 o'clock are displayed in Appendix N.

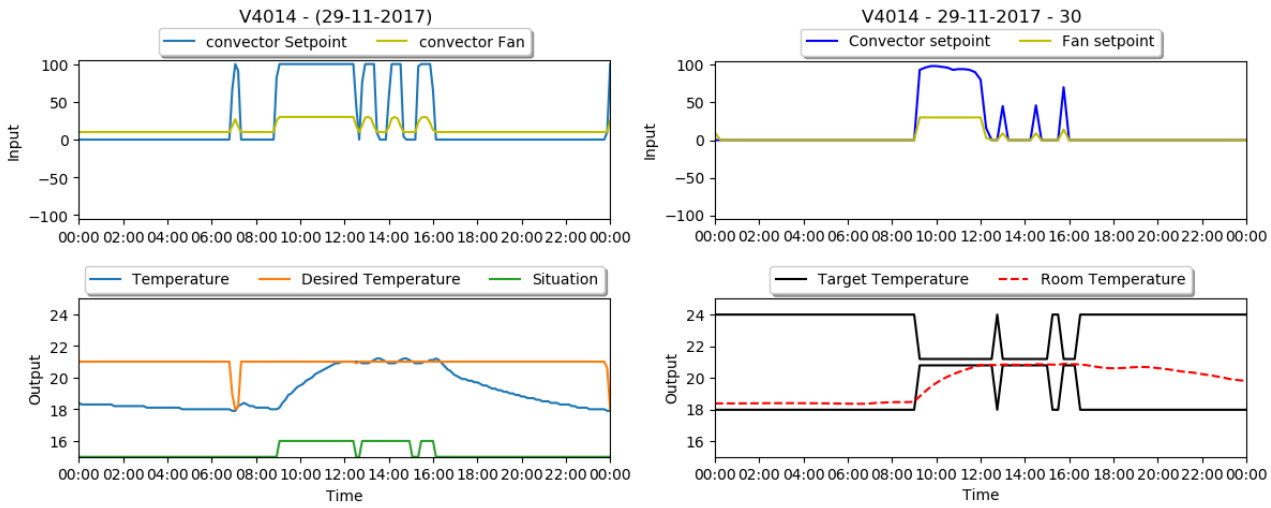


Figure 20 The graph on the left is the historic data and the simulation on the right.

By comparing the room temperature from the figures, it can be observed that the overall trend of the measured temperature over the entire day is roughly identical. The measured room temperature remains around 18°C during the first 9 hours of the day when the building is still closed. Around 9 o'clock the room becomes occupied and the convectors in the room start conditioning at full capacity (fan speed is capped on 30 percent) to get the room temperature up to the desired temperature. About 3 hours later, both the simulated temperature and the measured temperature from the historic data reach the desired temperature. For the remainder of that day when the room is occupied the temperature remains quite stable on 21°C in both graphs. Only in the historic data some minor fluctuations in the measured temperature can be witnessed. After 17 o'clock when the room becomes unoccupied, the temperature decreases slowly in both figures. The measured temperature decreases slightly less in the simulation.

When further analyzing the room temperature and calculating the temperature difference between the historic data and simulation the graph displayed in Figure 21 appears. From this figure it can be observed that the temperature difference remains below 0.6°C up to 17 o'clock. After 17 o'clock when the room becomes unoccupied the temperature difference increases substantially to an offset of about 2°C. A possibility is that the door of room was left open which affected the room temperature. Although, there is no hard evidence to support this hypophysis.

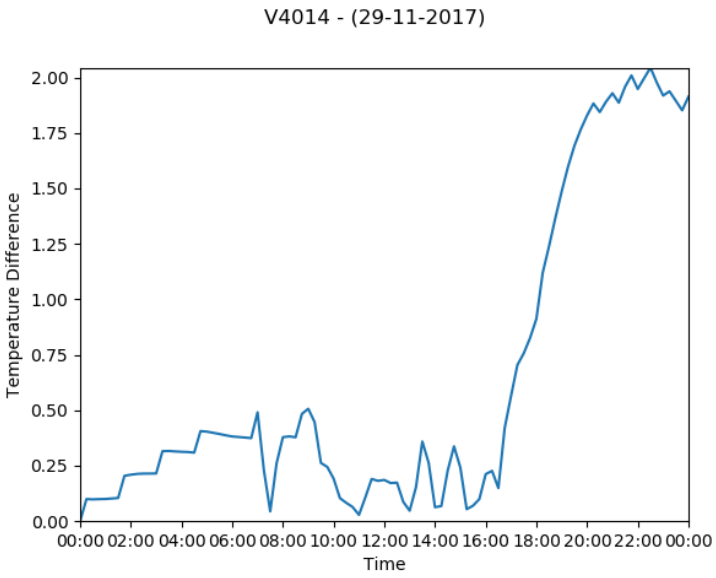


Figure 21 Temperature difference between historic and simulation.

The control settings from Figure 20 show some similarities and differences at the same time. The first noticeable difference is the peak at 7 o'clock that is observed in the historic data. The reason for the spike is that the measured temperature dropped just below the lower threshold of 18 °C. The predicted temperature in the simulation remains on the other hand just above this threshold and therefore remains idle. At 9 o'clock when the room becomes occupied both systems become active by conditioning the room with full capacity. The settings in the simulated situation are only slightly lower than those from the historic data. Around 12 o'clock when the temperature reaches its desired state and the room becomes unoccupied for a short amount of time, the convectors become idle again in both cases. During the remainder of that day when the room is occupied there are three instances in which both convectors become active for a short period of time. The difference that can be witnessed is that the valve position of the convectors in the reconstructed situation switches constantly from 0 to 100 percent, whereas the convectors in the simulated version increase to 50/60 percent. This might explain the fluctuations in the measured temperature that are observed in the historic data, whereas the predicted temperature in the simulation remains stable. For the remainder of that day when the building is closed, the systems remain idle in both situations.

In addition, when calculating the average setpoint of the convectors valve position during the entire timespan that the room is occupied. It can be observed that the historic data indicates an average of 83 percent, whereas the simulation indicates an average of 50 percent. The 33 percent lower average in the simulation indicates a that the settings used in the simulation would reduce the energy consumption. There are no exact numbers to could indicate the exact potential energy reduction, since the energy consumption cannot be measured on room level. Although, it is certain that reducing the average valve position results in less energy consumption.

As a recap of the observations from the first scenario, it can be stated that there are some minor differences between the measured temperature from the historic data and the predicted temperature from the simulation. For instance, the temperature from the historic data indicates some more fluctuations compared to those from the simulation. Besides that, the temperature difference between the simulation and the historic data increases substantially after 17 o'clock when the room becomes unoccupied. The results of the control settings indicate that the historic data consists of primarily two settings, either it is cooling at full capacity or the system is idle, whereas the control settings from the simulation are selected more gradually. The energy usage based on the valve position indicate an average reduction of 33 percent in the simulation. Although, there is no statistics that could illustrate the exact potential energy reduction.

Scenario 2

The second selected scenario is a warm weather day (11 July 2017) when an average outside temperature was measured of 19.2 °Celsius. The graph on the left side of Figure 22 displays the historic data of that day, whereas the graph on the right side shows the results from the simulation. The individual results of the MPC structure between 9 and 12 o'clock are displayed in Appendix O.

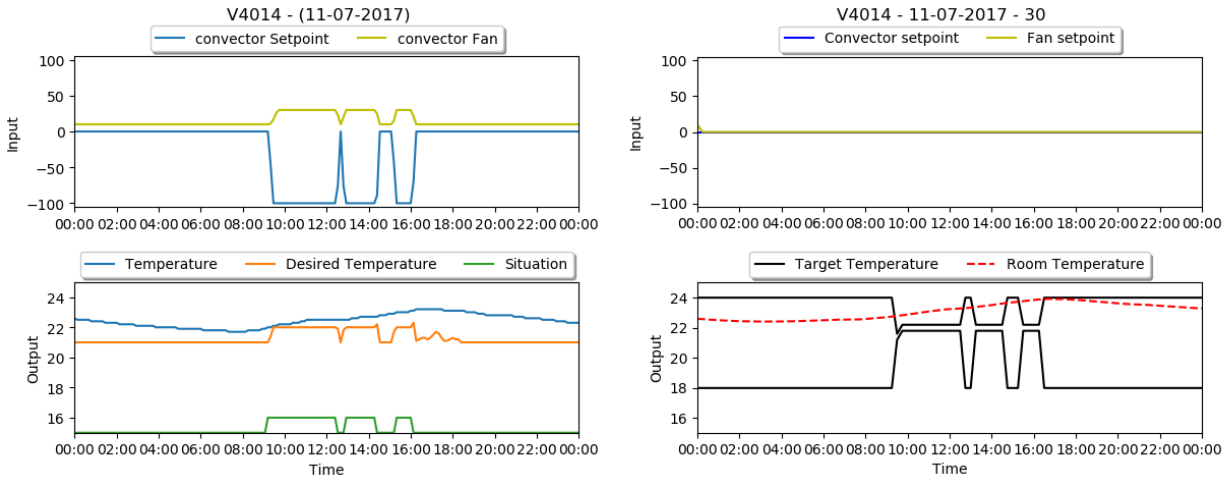


Figure 22 The graph on the left is the historic data and the simulation is on the right.

By comparing the room temperature from the figures, it can be observed that overall trend looks identical in both cases. There are some minor differences, for instance in the first 9 hours of the day when the room is unoccupied the temperature decreases in the historic data whereas the temperature in the simulation remains stable. During the remainder of the day the room temperature behaves in a more identical way. When the room becomes occupied the room temperature increases quite steadily in both cases. This increase continues up till 17 o'clock when the room becomes unoccupied again. After 17 o'clock the room temperature begins its decent in both cases.

When further analyzing the room temperature and calculating the temperature difference between the historic data and simulation the graph displayed in Figure 23 appears. From this figure it can be observed that the temperature difference increases to around 0.8°C in the first ours of the day. During the time the room is occupied the temperature difference fluctuates between 0.6°C and 0.8°C. During the final ours of the day the temperature difference increases slightly further up to 1.0°C.

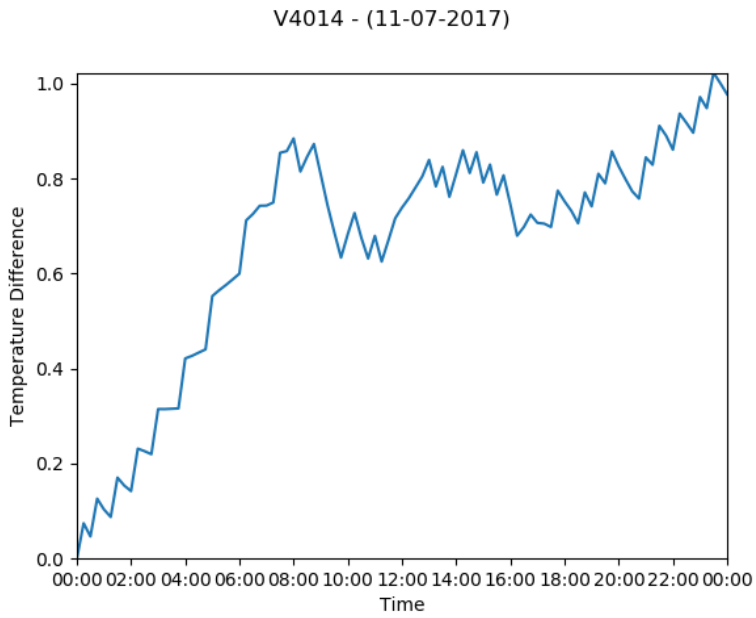


Figure 23 Temperature difference between historic and simulation.

Comparison from the perspective of the convector usage a major difference can be observed in Figure 22. The historic data indicates that the convector is cooling at maximum capacity over nearly the entire period that the room is occupied, whereas the convector in the simulation remains idle over the entire timespan. Presumably, the model predicts an identical room temperature regardless of the settings of the convector. Therefore, the cost function would return a higher value when the convectors are enabled, compared to when the convectors are disabled. It can therefore be convincingly confirmed that the convector configurations from simulation has a lower energy consumption. Although, it is debatable whether the room temperature would behave identically regardless of the convector settings in reality.

By summarizing the observations from the second scenario, it can be stated that the measured temperature from the historic data and the predicted temperature from the simulation show many similarities, whereas the results from the control settings are completely different in both situations. The main question that remains unanswered is whether the measured temperature would have been identical in reality when the convector would have been idle as presented in the simulation, which would explain the decisions made by the MPC. Another aspect that is not captured in the dataset and also cannot be measured is the feeling temperature of the occupants. The occupants may feel

the cool breeze of air from the convector, while the room temperature remains identical. In this situation it becomes a tradeoff between thermal comfort and energy consumption.

9.3. **Conclusion**

Based on the observations from both the scenarios it can be concluded that the predicted temperature from the simulation is quite identical with the temperature that is observed in the historic dataset. By comparing the control strategy, it can be stated that the MPC structure is able to select the optimal control settings during cold weather days when heating is required, but struggles with the selection of optimal control settings during hot weather days. These struggles to select the optimal cooling settings are related to the cooling capacity of the convectors. In the historical data there is plenty of evidence that shows that the cooling capacity is below the minimum to reduce the temperature. Due to this lack of cooling capacity the model does not recognize any changes in the temperature, and so the cost function of the MPC will be lower when the convector is positioned in idle instead of cooling.

10. **CONCLUSIONS**

This chapter reflects on the results from this thesis. It starts with a recap of the presented results and conclusions. Afterwards the research question with the associated sub-questions is elaborated on. Then the limitations of the research are described in detail. Followed by, the recommendation is presented followed by the further research possibilities.

10.1. **Summary**

The building sector has become one of the leading sectors that consumes the most energy, whereof most of the energy is used on thermal comfort. The goal of this thesis is to construct a data-driven approach by improving the HVAC strategy to reduce the energy consumption while maintaining thermal comfort. To achieve this goal the CRISP-DM methodology has been used for the modeling part.

The data used in this thesis is obtained from a utility building that is located in The Hague. The building originates from around 1917, was renovated in 2016 and is equipped with a high-end HVAC system. The data consists of weather-related data, HVAC data, room specific data and data of the general occupancy. The data has been extracted by using data mining techniques and is afterwards prepared according to the specified steps of CRISP-DM.

From the data two models have been constructed to predict the near future room temperature and another to predict the associated energy consumption. The modeling techniques that have been used for the construction of both models are Linear Regressor, MLP-Neural Network and a Random Forest Tree. The models for the room temperature were of a good quality. The Random Forest Tree and the MLP-Neural Network both performed better than the Linear Regressor, but the MLP-Neural Network had a significant advantage in the prediction speed. The models for the prediction of the energy consumption on the other hand were far from acceptable. The poor quality of the energy models was due to the inability to accurately measure the energy consumption on room level, leading to the use of estimation techniques to calculate the energy consumption.

The control strategy MPC has been used in this thesis and is still quite a unique control structure, due to the computational complexity to operate. The recent growth of computational power increased the applicability on this control strategy. The MPC selects the control settings by solving an optimization problem.

The quality of the proposed control strategy is determined based on simulation. Multiple scenarios from the past have been reconstructed and presented to the developed control technique and the results were afterwards compared with the actual situation that is captured in the dataset. The results

from the simulation showed that the constructed control technique was able to select the optimal control settings during cold days when heating was required, but failed to select the optimal cooling setting during warm days. Investigation indicated a problem with the cooling capacity of the convectors, which resulted in situations in which the room temperature would not decrease even when the cooler was enabled on full capacity.

A contribution to the scientific literature is made by showing the applicability of MPC in HVAC control strategies, with the use of a single temperature model for all rooms. Future research objectives for further improvements have been identified and are described in chapter 10.5

10.2. **Research questions**

This chapter elaborates on the research questions from chapter 1.2 by answering the sub-questions and main question based on the discovered results.

- **Which features are required to accurately model the energy consumption and temperature?**

The features used for the construction of the models to predict temperature and energy consumption on room level are selected by using scientific literature, field experts and data mining techniques. The process started by gathering all relevant data sources and listing all features that are present in each dataset. Afterwards, feature extraction techniques were used to construct additional features that were missing in the dataset. The elimination of irrelevant features consists of two stages. First scientific research and expert knowledge was used for the initial selection of relevant features. Second, a wrapper features selection method has been used to qualify the relevancy of the remainder features for both temperature and energy model separately. The final selection of features for the temperature model is displayed in Appendix H and for the energy model is displayed in Appendix I.

- **Which modeling techniques are applicable to forecast energy consumption and temperature?**

In general, all regression modeling techniques are a viable choice, whereby each technique has its advantages and disadvantages. In this research project three different modeling techniques were selected based on scientific literature and expert knowledge. A Linear Regressor was used to set the baseline for the advanced modeling techniques. The advanced modeling techniques that have been used are MLP-Neural Network and Random Forest Tree. Evaluation of the constructed models indicated that both MLP-Neural Network and Random Forest Tree were prediction more accurately compared to the Linear Regressor. The accuracy of prediction between the Random Forest Tree

and MLP-Neural Network were quite identical, only the Random Forest Tree performed slightly better. Another important aspect was the prediction time, since it has a major impact on the MPC process. On this aspect, the MLP-Neural Network scored significantly better compared to the Random Forest Tree and has therefore been selected as the most suitable modeling technique for our research project.

- **Which control technique can be used to optimize HVAC control?**

The control technique used in the research project is Model Predictive Control (MPC). The control technique relies heavily on computational power to operate and with the recent advancements in controller hardware and computational algorithms the control technique has become more viable. A review published by Serale et al., (2018) stated that MPC implementations for climate control are still lacking. Therefore the control technique used in this research project is MPC as a contribution to the this research field.

- **Which optimization technique can be used to find local optima to the cost-function?**

In scientific literature it was stated that GA and PSO were the most common used optimization techniques, but just as the modeling technique each optimization technique has its advantages and disadvantages. The optimization technique used in this research project is the Advanced Process OPTimizer (APOPT). The GEKKO optimization suite contains only a certain amount of optimization techniques, of which the GA and PSO are not part of. From the available optimization techniques, the APOPT required the least amount of time and computational power to converge.

Can a data-driven approach be constructed for the implementation of an advanced HVAC control system that reduces the energy consumption and maintains thermal comfort?

By combining the results from the sub-questions and including the simulation results it can be concluded that a data-driven approach that reduces the energy consumption while maintaining thermal comfort is possible. Although, the successfulness of this approach depends on presence of right circumstances. Three of most crucial aspects that determine the successfulness of the approach are elaborated on.

1. For a successful implementation it is a prerequisite that there is a sufficient amount of data available and the data is of a good quality. Inaccuracies in the data will have a direct effect on the prediction accuracy of the models, which determines the quality of the MPC structure, and therefore has a direct effect on whether goal of reducing energy consumption while maintaining thermal comfort can be achieved. A dataset that contains all weather seasons data is desirable, such that it can accurately predict the temperature in any condition.

2. The HVAC systems in the building should have the capacity to alter the room temperature according to the desired room temperature within a reasonable amount of time. A system that has an insufficient capacity results in a situation in which the model is unable to select the optimal control settings and therefore unable to reach the goal of reducing energy consumption while maintaining thermal comfort.

3. To enable the full potential of an MPC structure, it requires a building in which the desired room temperature is available ahead in time. The strength of MPC is the capability to plan the control setting (far) ahead in time, while keeping inaccuracies to a minimum. The planning strategy minimizes unnecessary preconditioning and ensure that the predefined thermal condition is reached just in time. To minimize the effect of inaccuracies the sample time should be kept to a minimal, such that the MPC can adjust to it accordingly.

10.3. **Limitations**

This chapter discusses the most important limitations to this research project.

- First of all, our conclusions are based on theoretical observations, where simulation is used to validate the applicability of the constructed control technique. The results from the simulation may deviate from reality and evaluation based on real implementation which would have resulted in a different conclusion.
- The conclusions on the energy consumption prediction were based on estimations and calculation methods. The use of a different approach or the presence of room specific energy data may result in different results, in which it would be viable to use the energy models in the control strategy.
- The effectiveness of the MPC structure is case specific meaning that in another building with different specifications the results may deviate (positively or negatively).

10.4. **Recommendations**

During the execution of the research project, several recommendations have been made.

- In the early stages of the project, while analyzing the data, it was observed that the pressure settings of the AHUs consisted of primarily two settings. The Pasific Northwest National Laboratory published a training guide for re-tuning of the AHU static pressure and stated that energy saving is possible by adjusting the static pressure based on the building load. Incorporating an advanced technique that adjusts the static pressure based on the building load or manual adjusting the settings based on the building load could have a possitive effect on the energy consumption.

- Implementation of the constructed system is highly discouraged due to the instability to select the optimal control settings during warm days. It is likely that during warm days the convectors will remain idle even when cooling is required. Even though the capacity of the convector is too low to have a direct effect on the measured room temperature, it may produce a breeze of cold air that reduces the feel temperature of the occupants in the room. The feel temperature cannot be measured and therefore the model is unable to incorporate it.
- Even while the constructed control system is invalid for this specific case, the approach used to construct this control system is still valid. Reproduction using the same approach in a different building, where the convector has the required capacity, is likely to reach the desired state of reducing the energy consumption while maintaining thermal comfort.

10.5. **Further research**

From this research project, there are multiple directions in which further research is advised.

- The approach presented in this research project has been validated by the use of simulation. The results from the simulation indicate that the approach can reach the desired state of reducing energy consumption while maintaining thermal comfort. Although, deployment of the system in a real-life setting has not been validated yet and further research is required.
- In this research project traditional regression modeling techniques have been used for the construction of the temperature model. Evaluation of these models has shown that traditional regression modeling techniques are already capable to accurately predict the near future room temperature. The use of modeling techniques specifically for time series, like recurrent neural networks, have not been tested and further research could identify whether this type of modeling techniques would improve the prediction quality.
- The methods used in this research project to estimate the energy consumption on room level have shown to be inaccurate. Further research could identify whether there are more suitable methods to estimate or calculate the energy consumption on room level.
- There were multiple signs that indicated that the effectiveness of the cooling capacity of the convectors in the building used in this research were below the threshold to effectively reduce the room temperature. This resulted in situation in which the control system was unable to select the optimal control settings. Further research could identify whether this problem also occurs in building with the proper amount of cooling capacity.

REFERENCES

Wang, S., & Ma, Z. (2008). Supervisory and optimal control of building HVAC systems: a review. *HVAC&R Research*, 3-32.

Abdel-Aal, R., & Al-Garni, a. (1997). Forecasting monthly electric energy consumption in eastern Saudi Arabia using univariate time-series analysis. *Energy*, 1059-1069.

Afram, A., & Janabi-Sharifi, F. (2014). Theory and applications of HVAC control systems – A review of model predictive control (MPC). *Building and Environment*, 343-355.

Afram, A., Janabi-Sharifia, F., & FungaKaamr, A. S. (2017). Artificial neural network (ANN) based model predictive control (MPC) and optimization of HVAC systems: A state of the art review and case study of a residential HVAC system. *Energy and Buildings*, 96-113.

American Society of Heating, R. a.-C. (2011). *HVAC Applications Handbook.*

American Society of Heating, R. a.-C. (2013). *ASHRAE Handbook—Fundamentals (SI Edition).* Atlanta, GA, USA: ASHRAE.

Asadi, E., Gameiro da Silva, M., Antunes, C. H., Dias, L., & Glicksman, L. (2014). Multi-objective optimization for building retrofit: A model using genetic algorithm and artificial neural network and an application. *Energy and Buildings*, 444-456.

Bacher, P., & Madsen, H. (2011). Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings*, 1511-1522.

Beal, L., Hill, D., Martin, R., & Hedengren, J. (2018). GEKKO Optimization Suite. *Processes*, 106.

Behrendt, M. (2009, October 2).

Bekkerman, R., El-Yaniv, R., Tishby, N., & Winter, Y. (2003). Distributional Word Clusters vs. Words for Text Categorization. *Machine Learning Research 3*, 1183-1208.

Beltran, A., & Cerpa, A. E. (2014). Optimal HVAC Building Control with Occupancy Prediction. *Embedded Networked Sensor Systems.* Memphis.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 245-271.

Brownlee , J. (2017, januari 30). *How to Decompose Time Series Data into Trend and Seasonality*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com

Bueno, B., Norford, L., Pigeon, G., & Britter, R. (2012). A resistance-capacitance network model for the analysis of the interactions between the energy performance of buildings and the urban climate. *Building and Environment*, 116-125.

Bull, R., Chang, N., & Fleming, P. (2012). The use of building energy certificates to reduce energy consumption in European public buildings. *Energy and Buildings*, 103-110.

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: a new perspective. *Neurocomputing*.

Cavaioni, M. (2017, Februari 2). *Machine Learning: Decision Tree Classifier.* Retrieved from Medium: https://medium.com

Chen, P., Wilbik, A., van Loon, S., Boer, A.-K., & Kaymak, U. (2018). Finding the optimal number of features based on mutual information. In *Advances in Fuzzy Logic and Technology 2017 - Proceedings of* (pp. 477-486). Germany: Springer.

da Costa Sousa, J., & Kaymak, U. (2001). Model predictive control using fuzzy decision functions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 54-65.

Deb, C., Zhang, F., Yang, J., & Lee Eang, S. (2017). A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 902-924.

Donges, N. (2018, Februari 22). *The Random Forest Algorithm.* Retrieved from Towards Data Science: https://towardsdatascience.com

Erickson, V., & Cerpa, A. (2010). Occupancy Based Demand Respone HVAC Control Strategy.

Ferreira, P. M., Ruano, A., Silva, S., & Conceicão, E. (2012). Neural networks based predictive control for thermal comfort and energy savings in public buildings. *Energy and Building*, 238–251.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Machine Learning Research 3*, 1289-1305.

González, P. A., & Zamarreño, J. M. (2005). Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. *Energy and Buildings*, 595–601.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Machine Learning Research*, 1157-1182.

Guyon, I., & Elisseeff, A. (2006). An Introduction to Feature Extraction. *Feature Extraction*, 1-25.

Haberl, J. S., Culp, C., & Claridge, D. E. (2005). ASHRAE's GUIDELINE 14-2002 FOR MEASUREMENT OF ENERGY AND DEMAND SAVINGS: HOW TO DETERMINE WHAT WAS REALLY SAVED BY THE RETROFIT. *The International Conference for Enhanced Building Operations.* Pennsylvania.

Harper, G., & Pickett, S. D. (2006). Methods for mining HTS data. *Drug Discovery Today*, 694-699.

Hedengren, J., Mojica, J., Cole, W., & Edgar, T. (2012). APOPT: MINLP Solver for Differential Algebraic Systems with Benchmark Testing. *INFORMS Annual Meeting*, (pp. 14-17). Phoenix, AZ, USA.

Herkel, S., Knapp, U., & Pfafferott, J. (2005). A preliminary model of user behaviour regarding the manual control of windows in office buildings. *Proceedings buildings simulations*.

Hu, L., Gao, W., Zhao, K., Zhang, P., & Wang, F. (2018). Feature selection considering two types of feature relevancy and feature interdependency. *Expert Systems with Applications*, 423-434.

Hyndman, R. J. (2014, March 31). Measuring forecast accuracy.

Hyndman, R. J., & Athanasopoulos, G. (2013). *Forecasting: principles and practice.* OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 679-688.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. *Proceedings of the Eleventh International Conference* (pp. 121–129). New Jersey: Morgan Kaufmann.

Karn, U. (2016, August 9). *A Quick Introduction to Neural Networks.* Retrieved from the data science blog: https://ujjwalkarn.me

Kaushik, S. (2016, December 1). *Introduction to Feature Selection methods with an example (or how to select the right variables?)*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 273-324.

Kusiak, A., Li, M., & Tang, F. (2010). Modeling and optimization of HVAC energy consumption. *Applied Energy*, 3092–3102.

Kusiak, A., Xu, G., & Tang, F. (2011). Optimization of an HVAC system with a strength multi-objective particle-swarm algorithm. *Energy*, 5935-5943.

Kwak, Y., Seo, D., Jang, C., & Huh, J.-H. (2013). Feasibility study on a novel methodology for short-term real-time energy demand prediction using weather forecasting data. *Energy and Buildings*, 250-260.

Kwok, S., & Lee, E. (2011). A study of the importance of occupancy to building cooling load in prediction by intelligent approach. *Energy Conversion and Management*, 2555-2564.

Lal, T. N., Chapelle, O., Weston, J., & Elisseef, A. (2006). Embedded Methods. In I. Guyon, M. Nikravesh, S. Gunn, & L. A. Zadeh, *Feature Extraction* (pp. 137-165). Berlin, Heidelberg: Springer.

Lan, H. (2017, november 6). *Decision Trees and Random Forests for Classification and Regression pt.2.* Retrieved from Towards Data Science: https://towardsdatascience.com

Larochelle, H., Bengio, Y., Louradour, J., & Lamblin, P. (2009). Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*.

Lazos, D., Sproul, A. B., & Kay, M. (2014). Optimisation of energy management in commercial buildings with weather forecasting inputs: A review. *Renewable and Sustainable Energy Reviews 39*, 587-603.

Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Applying support vector machine to predict hourly cooling load in the building. *Applied Energy*, 2249-2256.

Liu, X., Bai, E., & Fang, J. (2010). Time-variant slide fuzzy time-series method for short-term load forecasting. *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference* (pp. 1:65–8). Xiamen, China: IEEE.

Maaskant, M. (2016). *Data Mining Approaches for Calculating the Energy Consumption of Buildings.*

Magoules, F., & Zhao, H.-X. (2016). *Data Mining and Machine Learning in Building Energy Analysis: Towards High Performance Computing.* Hoboken, NJ, USA: John Wiley & Sons Inc.

Meyers, R., Williams, E., & Matthews, S. (2010). Scoping the potential of monitoring and control technologies to reduce energy use in homes. *Energy and Buildings*, 563-569.

Mirakhorli, A., & Dong, B. (2016). Occupancy behavior based model predictive control for building indoor climate—A critical review. *Energy and Buildings*, 499-513.

Monfeta, D., Corsib, M., Choinièreb, D., & Ark, E. (2014). Development of an energy prediction tool for commercial buildings using case-based reasoning. *Energy and Buildings*, 152-160.

Nadali, A., Naghizadeh kakhky, E., & Nosratabadi, H. (2011). Evaluating the Success Level of Data Mining Projects Based on CRISP-DM Methodology by a Fuzzy Expert System. *Electronics Computer Technology* (pp. 161-165). Kanyakumari, India: IEEE.

Nau, R. (2017, December 14). *Stationarity and differencing*. Retrieved from Statistical forecasting: notes on regression and time series analysis: https://people.duke.edu

Oğcu, G., Demirel, O., & Zaim, S. (2012). Forecasting electricity consumption with neural networks and support vector regression. *Procedia - Social and Behavioral Sciences*, 1576 – 1585.

Oldewurtel, F., Gyalistras, D., Gwerder, M., Jones, C. N., Parisio, A., Stauch, V., . . . Morari, M. (2010). Increasing Energy Efficiency in Building Climate Control using Weather Forecasts and Model Predictive Control. *Clima - RHEVA World Congress*.

Pasific Northwest National Laboratory. (n.d.). *Re-tuning Commercial Buildings Resources.* Retrieved from Pasific Northwest National Laboratory: https://buildingretuning.pnnl.gov/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2825-2830). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2011.

Pérez-Lombard, L., Ortiz, J., & Pout, C. (2008). A review on buildings energy consumption information. *Energy and Buildings, 40*(3), 394-398.

Qin, J., & Badgwellb, T. A. (2003). A survey of industrial model predictive control technology. *Control Engineering Practice*, 733-764.

Rodriguez-Lujan, I., Huerta, R., Elkan, C., & Cruz, C. S. (2010). Quadratic Programming Feature Selection. *Machine Learning Research 11*, 1491-1516.

Serale, G., Fiorentini, M., Capozzoli, A., Bernardini, D., & Bemporad, A. (2018). Model Predictive Control (MPC) for Enhancing Building and HVAC System Energy Efficiency: Problem Formulation, Applications and Opportunities. *Energies, 11*, 631.

Sharma V, A. (2017, March 30). *Understanding Activation Functions in Neural Networks.* Retrieved from Medium: https://medium.com

Srivastava, T. (2015, July 1). *Difference between Machine Learning & Statistical Modeling*. Retrieved from analyticsvidhya: https://www.analyticsvidhya.com

Sun, Y., Wang, S., & Xiao, F. (2013). Development and validation of a simplified online cooling load prediction strategy for a super high-rize building in Hong Kong. *Energy Conversion and Management*, 20-27.

Sutter, J. M., & Kalivas, J. H. (1993). Comparison of Forward Selection, Backward Elimination, and Generalized Simulated Annealing for Variable Selection. *Microchemical Journal*, 60-66.

Swann, W. H. (1969). A survey of non-linear optimization techniques. *FEBS Letters*, S39-S55.

Thomson, D. (1994). Jackknifing multiple-window spectra. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 73-76).

Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 1761-1768.

Wächter, A., & Biegler, L. (2006). On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming. *Mathematical Programming 106(1)*, 25-57.

Wong, I., & Krüger, E. (2017). Comparing energy efficiency labelling systems in the EU and Brazil: Implications, challenges, barriers and opportunities. *Energy Policy*, 310-323.

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-03).* Washington DC.

Zeng, Y., Zhang, Z., & Kusiak, A. (2015). Predictive modeling and optimization of a multi-zone HVAC system with data mining and firefly algorithms. *Energy*, 393-402.

Zhang, Y., Li, S., Wang, T., & Zhang, Z. (2013). Divergence-based feature selection for separate classes. *Neurocomputing*, 32-42.

Zhou, Q., Wang, S., Xu, X., & Xiao, F. (2008). A grey-box model of next-day building thermal load prediction for energy-efficient control. *Energy Research*, 1418–1431.

APPENDICES

Appendix A

#	Name	Description	Unit	Source
1	OUT-TEMP	Outside – Dry-bulb temperature	°C	BAS/WFS*
2	OUT-WSPE	Outside – Wind speed	km/h	BAS/WFS*
3	OUT-WDIR	Outside – Wind direction	Deg	BAS/WFS*
4	OUT-HUMI	Outside – Humidity	% RH	BAS/ WFS*
5	BD-OCCU	Building – Occupancy rate	persons per hour	Entrance gates
6	OUT-ATMP	Outside – Apparent temperature	°C	WFS*
7	OUT-APRE	Outside – Atmospheric pressure	Pa	WFS*
8	OUT-CCOV	Outside – Cloud cover	%	WFS*
9	OUT-LIPR	Outside – Liquid precipitation rate	cm/h	WFS*
10	OUT-OZON	Outside – Ozone	Dobson units	WFS*
11	OUT-PRET	Outside – Precipitation type	{rain, snow, sleet, 0}	WFS*
12	OUT-SNOW	Outside – Snowfall	cm/h	WFS*
13	OUT-UVIN	Outside – UV index	UV index	WFS*
14	OUT-WGUS	Outside – Wind gust	km/h	WFS*
15	OUT-SOLE	Outside – Solar intensity east	B/HFt2	BAS
16	OUT-SOLS	Outside – Solar intensity south	B/HFt2	BAS
17	OUT-SOLW	Outside – Solar intensity west	B/HFt2	BAS
18	OUT-AIRQ	Outside – Air quality	PPM (Parts Per Million)	BAS

Table 13 The available features from the building perspective.

*WFS = Weather Forecasting Service

The list displayed in Table 13 contains all the features that are available from the perspective of the entire building. These features are more general and are equal throughout the entire building. The features are primarily related to weather and occupancy rate. The weather data is either from the on-site weather station that is stored in the BAS or from the external online weather service provider. The occupancy rate is extracted from a separate system that handles the entrance gates.

Appendix B

#	Name	Description	Unit	Source
1	AHU-TSET	AHU – Supply air temperature setpoint	°C	BAS
2	AHU-PSET	AHU – Supply air duct static pressure setpoint	Pa	BAS
3	AHU-STMP	AHU – Supply air temperature	°C	BAS
4	AHU-HVAL	AHU – Heating water valve position	%	BAS
5	AHU-HSTE	AHU – Heating water supply temperature	°C	BAS
6	AHU-HRTE	AHU – Heating water return temperature	°C	BAS
7	AHU-CVAL	AHU – Cooling water valve position	%	BAS
8	AHU-CSTE	AHU – Cooling water supply temperature	°C	BAS
9	AHU-CRTE	AHU – Cooling water return temperature	°C	BAS
10	AHU-PRES	AHU – Supply air duct static pressure	Pa	BAS
11	AHU-RTMP	AHU – Return air temperature	°C	BAS
12	AHU-ITMP	AHU – Intake air temperature	°C	BAS
13	AHU-MTMP	AHU – Mixed air temperature (after thermal wheel)	°C	BAS
14	AHU-LTMP	AHU – Lost air temperature (after thermal wheel)	°C	BAS

Table 14 The available features for each of the AHUs.

The list displayed in Table 14 contains all the features that are available for each of the AHUs that are present in the building. A technical description of the AHU can be found in chapter 4.4 (HVAC-systems).

Appendix C

#	Name	Description	Unit	Source
1	RM-MTMP	Room – Measured temperature	°C	BAS
2	RM-DTMP	Room – Desired temperature	°C	BAS
3	RM-SITU	Room – Situation	{off, stand-by, comfort}	BAS
4	RM-HEAT	Room – Heating	%	BAS
5	RM-COOL	Room – Cooling	%	BAS
6	RM-VENT	Room – Ventilation	%	BAS
7	RM-AIRQ	Room – Air quality	PPM (Parts Per Million)	BAS
8	RM-VAV	Room – Variable air volume	% open	BAS
9	RM-SURF	Room – Surface	M^2	Blueprints
10	RM-TYPE	Room – Type	Set of types listed in Table 4	Blueprints
11	RM-CAPA	Room – Capacity (People)	N	Blueprints
12	RM-WIND	Room – Windows	N	Blueprints
13	RM-NCNV	Room – Number of convectors	N	Blueprints
14	RM-AHUZ	Room – AHU zone	N	Blueprint
15	RM-GKWZ	Room – GKW&CV zone	{NO, ZW}	Blueprint
16	RM-FLOR	Room – Floor	N	Blueprint
17	RM-LOCA	Room – Location	{N, O, Z, W, G}	Blueprint
18	RM-NUMB	Room – Number	{text}	Blueprint

Table 15 The available features for each of the rooms.

The features that are available for each room are displayed in Table 15. The feature originating from the BAS consists of an entire history. The features obtained from the blueprints are a single value for each room. The values from the BAS are used to determine the behavior of the temperature in a room, whereas the features from the blueprints are used to compare rooms. Identical rooms are assumed to behave likewise.

Appendix D

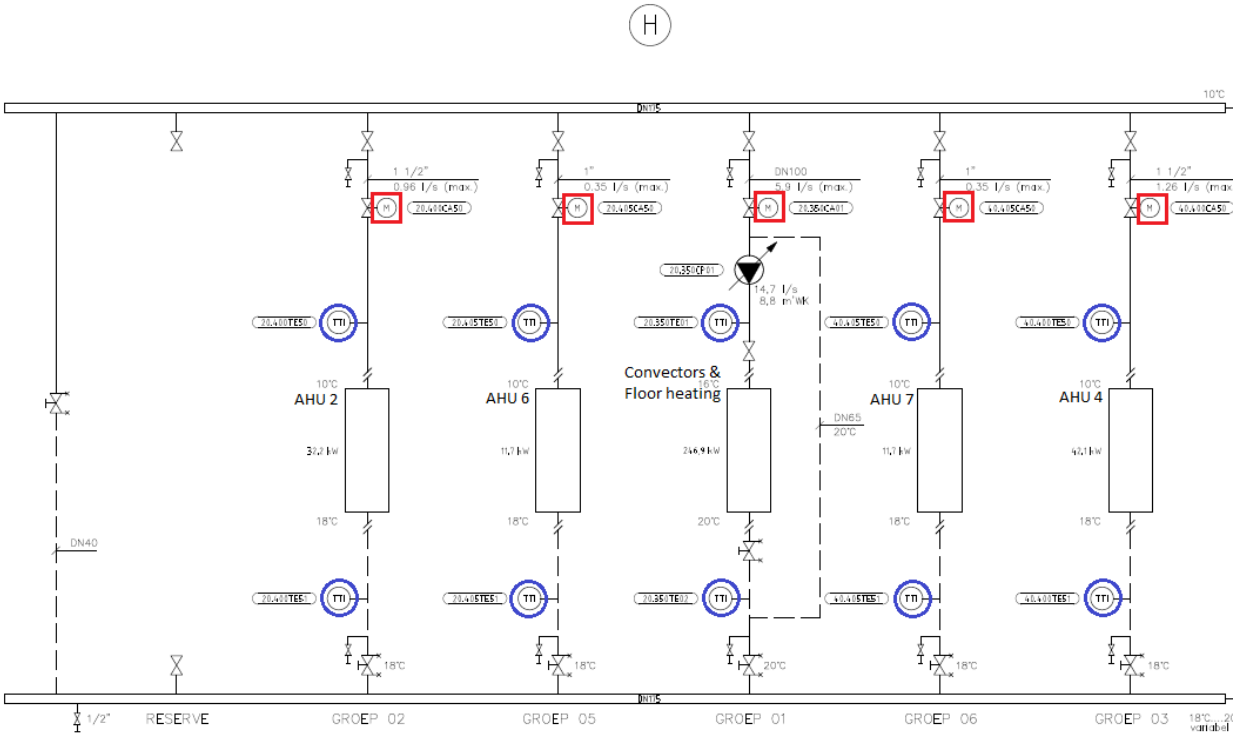


Figure 24 Blueprint of the cooling water distribution for group H.

The distribution of cooling water is divided into the group's "H" and "I". Group "H", displayed in Figure 24, provides the cooling water to the AHUs 2, 6, 7 and to the convectors and floor heating of a single side of the building. Each of these sub-groups is equipped with sensors and a controllable valve. The temperature sensors, indicated with blue circles, measure the temperature of the water. As can be noted, each sub-group is equipped with two temperature sensors. One measures the temperature of the water before entering the system and one sensor is measuring it when exiting the system. The controllable valve, indicated with a red square, regulates the water flow rate. The position of a valve is adjusted by a servomotor that is connected to the BAS. The BAS can instruct the servomotor to open the valve by a certain percentage although this only tells the percentage of the opening of the valve and not the exact flow of water. The exact flow can be determined by using the technical specification of the valve, together with the blueprints of the pipe structure. The technical specification of the valve is used to extract the diameter of the valve, which influences the maximal flow rate. Besides that, it is also used to translate the percentage of the BAS to the exact flow rate that is allowed through the valve. The technical specification contains a figure that shows the characteristics of the valve. The characteristics of one of the valves are displayed in Figure 25.

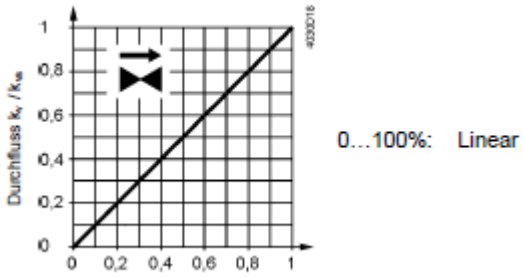


Figure 25 The flow rate based on the opening of the valve for a 2-way valve.

The y-axis of this figure displays the maximum flow rate, where the x-axis displays the percentage in which the valve is opened. In this particular case the relation is linear, so the opening percentage corresponds to the maximal flow rate percentage of the valve although some valves have an exponential curve or a combination of an exponential curve and linear. The pipes towards and after the valves also play a vital role to determine the flow rate. The diameter of pipes can limit the maximal flow. When the diameter of the pipes is smaller than the valves the maximal flow is reduced by the pipes.

By combining all this information, it is possible to determine the amount of cooling energy that is used for each of the sub-groups by using the heat transfer formula.

$Q = M * c * \Delta T$

In this formula Q refers to heat content in Joules, c refers to the specific heat, M refers to the mass flow rate and ΔT refers to the difference of temperature. The ΔT can be obtained by using the two sensors (before and after), c is a given constant which is $4186\,\mathrm{J/(kg * K)}$ for water and M can be obtained from the position of the valve.

As displayed in Figure 24 each AHU consists with its own subgroup and therefore by using the heat transfer formula is it possible to calculate the energy consumption for each of the AHUs, although the goal is to calculate the energy consumption of each room and therefore an additional step is taken. To calculate the amount of energy that is used by each room the following formula is used.

$$E(t,k) = \frac{V_t}{\sum_{i=0}^{n} V_i} * A_k$$

The energy E consumed by room t with its associated AHU k is calculated by dividing the VAV-opining of the room V_t by the sum of all rooms VAV-opening that are connected to AHU k and multiply it with the association AHU energy consumption A_k.

The sub-group for convectors and floor heating is slightly more complex. This sub-group requires a constant flow of water and is equipped with an additional pump that circulates water at a constant speed. Nevertheless, the same formula can be used to calculate the energy that is used over the entire sub-group. The only difference is that M is determined by the speed of the pump and not the

opening of the valve. This would result in calculating the energy usage over the entire sub-group. In the previous case there was only a single AHU and hence specific enough to use. This group on the other hand contains roughly fifty percent of all convectors that are placed in the entire building together with some floor eating. To determine the energy usage of a single convector another approach has to be used. Sensors on each of the convectors would be a sufficient solution, but they are not available. In order to calculate the energy consumption for each convector the temperature difference is taken over the entire group. This gives the best estimation of the temperature difference that occurred in each convector. The high thermal mass of the building ensures that the temperature in the entire building is quite stable and changes very slow. This enables us to assume that the temperature in each room is roughly identical. The position of the valve that controls the water flow to the convectors is available. Each convector is equipped with an identical 6-way valve as displayed in Figure 26. The pipes to the valves are at any time either of the same size or larger and therefore not influencing the maximal flow rate.

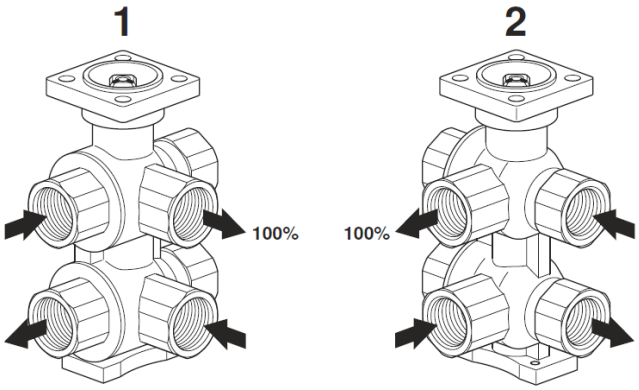


Figure 26 6-way valves for convectors.

The 6-way valve contains two 3-way valves on top of each other that are controlled by the same servomotor. In Figure 26 the top 3-way valve handles the supply heating/cooling water, and the bottom handles the return heating/cooling water. The top 3-way valve obtains heating water from one side (position 1 in the figure) and cooling water from the other side (position 2 in the figure). By adjusting the valve position, it extracts cooling water, heating water or nothing. The servomotor on top of the valve controls this by turning the valve position. The valve position and associating flow rate is displayed in Figure 27.

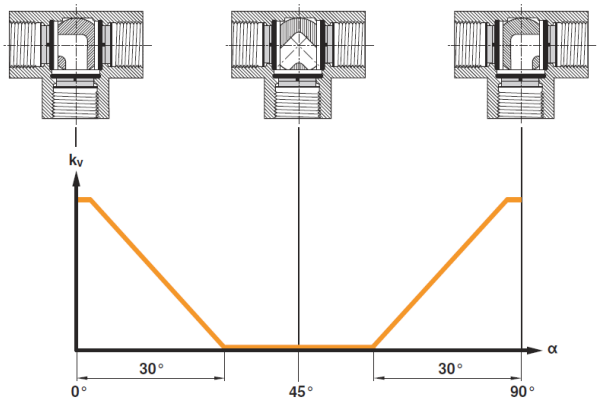


Figure 27 Controls of a 6-way valve.

The BAS displays the cooling and heating power separately in percentages (from 0 to 100%) instead of degrees turned. For the translation from percentages to opening degrees it is assumed that the entire dead zone corresponds to the setting where both cooling and heating display zero percent. There were features in the dataset where the cooling power was equal to ten percent. Which means that if the middle of the dead zone was equal to zero percent, the valve would have still been closed when it was turned by only ten percent. Based on this it becomes possible to determine the flow rate of the heating and cooling water and hence makes it possible to calculate the energy usage of a single convector by using the same heat transfer formula.

Appendix E

The data provided by the BAS has an interval of eight minutes, and therefore the minimal possible prediction horizon. The prediction horizon does not have to be exactly eight or a multiplication of it, but any value in between is not covered by the dataset and therefore cannot be modeled accurately. The final prediction horizon used to select the control setting might be a more trivial value that is close to the prediction horizon used to construct the model. Taking a different value for the control sample time may result in some inaccuracies, but as long as it remains relatively close it should not have a direct effect on the outcome.

The prediction horizon for the model is determined based on the following studies.

- The goal of the first analysis was used to determine the average amount of time it takes to get the measured room temperature to the desired room temperature. The start time was set on the moment that the room situation changed from unoccupied to occupied and the end time was set on the moment that the measured room temperature was within a range of 0.5 degrees of the desired temperature. A range of 0.5 degrees was selected since there might be an acceptance margin in which temperature will not be altered. During inspection of the data, there were cases witnessed in which the measured temperature was already within the 0.5 degrees radius resulting in a time span of zero. Besides that, there were also cases in which the desired temperature was never reached. Both of these cases would have a negative influence on the average time it takes to reach the desired temperature. Therefore, only cases in which the temperature difference was outside the 0.5 acceptance radiance and cases in which the desired temperature was reached within three-hour time span were included in the analysis.
- The second study looks at the effect of the measured temperature between two intervals. In this case a filter is used to only use the data when the room is occupied and the measured temperature is at least 0.5°C off the desired temperature. Hereby it can be assured that the convector is activated to reach the desired temperature. The main goal of this is to determine whether small or large temperature difference can be measured, due to the use of the convector and other systems.
- A third study is used to spot under- and overshoots of the temperature measured within the time interval. This study requires at least an additional measurement within the interval and therefore the eight minutes interval is excluded from this study. To clarify, an example is presented in Figure 28. In this example the temperature changes from 20°C to 23°C during the entire sample time, but measured an even higher temperature in between. Here it can be concluded that the sample time is too long, since the sample time contains a heating and cooling part and therefore contains multiple control settings that cannot be combined.

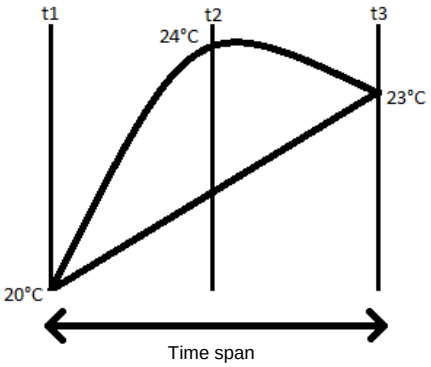


Figure 28 Example of an overshoot.

During the first analysis it was concluded that it takes on average 48 minutes to reach the desired condition.

The results of the second study are presented in Figure 29. Based on this figure it can be concluded that with an interval of 8 minutes the room temperature tends to stay equal, even when the convector in the room was enabled. This is a clear indication that the prediction horizon is to short, since the effect of the convector cannot be measured. It can be noted that the larger the timespan the more likely it is that the temperature has changed.

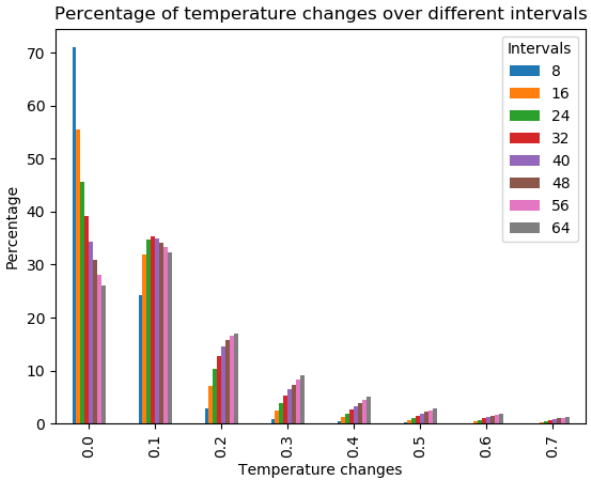


Figure 29 Temperature changes over different intervals, with a minimum offset of 0.5°C.

By increasing the interval, it becomes more likely that the temperature will change in the room. A downside of increasing the interval is the increased possibilities that unmodelled inaccuracies are affecting the temperature. Therefore, the third study looks at the under- and overshoots, presented in Figure 30 and Figure 31. The results are quite conclusive; for each of the intervals over- and undershoots are rarely measured.

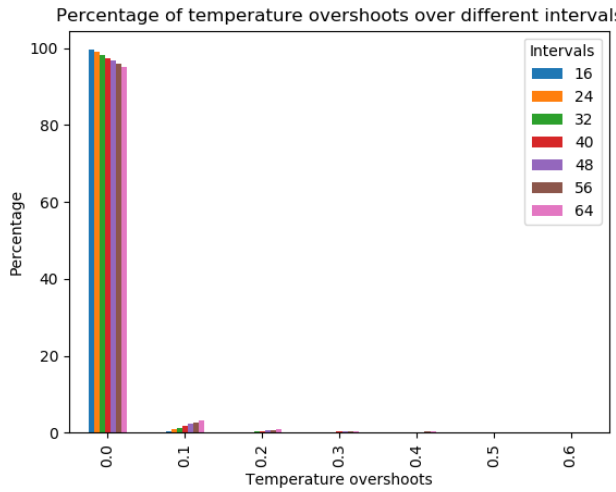


Figure 30 Temperature overshoots over the different intervals.

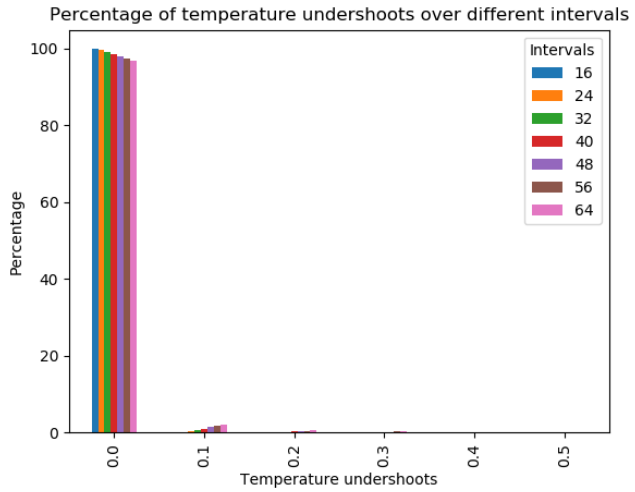


Figure 31 Temperature undershoots over the different intervals.

By combining the results from the three analyses it can be concluded that the minimum timespan should at least exceed the average timespan of 48 minutes to allow for planning ahead. Secondly, it can be confirmed that increasing the timespan would also increase the likelihood in witnessing temperature changes. Thirdly, temperature over-/undershoots are rarely witnessed up to 64. Based on the three analyses it was concluded that the time horizon of 64 minutes would be sufficient to allow for planning ahead with only minor over-/undershoots.

Appendix F

#	Feature	1	2	3	4	5	6	Percentage
1	ahu_temperature_setpoint	True	True	True	True	True	True	100%
2	room_convector_setpoint	True	True	True	True	True	True	100%
3	room_convector_fan	True	True	True	True	True	True	100%
4	ahu_pressure_setpoint	True	True	True	True	True	True	100%
5	room_energy	False	True	True		False	False	33%
6	room_energy_Nd	False	False	False		False	False	0%
7	room_energy_N2d	False	False	True		False	False	17%
8	room_measured_temperature	True	True	True	True	True	True	100%
9	room_measured_temperature_Nd	True	True	True	True	True	True	100%
10	room_measured_temperature_N2d	True	True	True	True	True	False	83%
11	outside_temperature	True	True	True	True	True	True	100%
12	outside_wind_velocity	True	True	True	True	True	False	83%
13	outside_wind_direction	True	True	True	True	True	False	83%
14	outside_humidity	True	True	True	True	True	True	100%
15	building_occupancy	True	True	True	True	True	True	100%
16	ahu_supply_temperature	True	True	True	False	False	False	50%
17	room_desired_temperature	True	True	True	True	True	True	100%
18	room_situation	False	True	True	True	True	True	83%
19	room_air_quality	True	True	True	True	True	True	100%
20	room_surface	True	True	True	True	True	True	100%
21	room_type	True	True	True	True	True	True	100%
22	room_capacity	False	True	True	True	True	False	67%
23	room_windows	False	True	True	True	False	True	67%
24	room_nr_convectors	False	False	True	True	True	False	50%
25	room_nr_vavs	False	False	False	False	False	False	0%
26	room_number	True	True	True	True	True	True	100%
27	room_floor	False	True	True	True	True	True	83%
28	room_location	False	True	True	True	True	True	83%

Table 16 Feature selection Temperature results.

Appendix G

#	Feature	1	2	3	4	Percentage
1	ahu_temperature_setpoint	True	True	True	True	100%
2	room_convector_setpoint	True	True	True	True	100%
3	room_convector_fan	True	True	True	True	100%
4	ahu_pressure_setpoint	True	True	True	True	100%
5	room_energy	True	True	True	True	100%
6	room_energy_Nd	True	True	True	True	100%
7	room_energy_N2d	True	True	True	True	100%
8	room_measured_temperature	True	True	True	True	100%
9	room_measured_temperature_Nd	True	True	False	True	75%
10	room_measured_temperature_N2d	True	True	False	True	75%
11	outside_temperature	True	True	True	True	100%
12	outside_wind_velocity	True	True	True	True	100%
13	outside_wind_direction	True	True	True	True	100%
14	outside_humidity	True	True	True	True	100%
15	building_occupancy	True	True	True	True	100%
16	ahu_supply_temperature	True	True	True	True	100%
17	room_desired_temperature	True	True	False	False	50%
18	room_situation	True	True	True	True	100%
19	room_air_quality	True	True	True	True	100%
20	room_surface	True	True	True	True	100%
21	room_type	True	True	True	True	100%
22	room_capacity	True	True	True	False	75%
23	room_windows	True	True	True	True	100%
24	room_nr_convectors	True	True	False	True	75%
25	room_nr_vavs	True	True	False	False	50%
26	room_number	True	True	True	True	100%
27	room_floor	True	True	False	True	75%
28	room_location	True	True	True	False	75%

Table 17 Feature selection Energy results.

Appendix H

#	Parameter type	Parameter	Name	Description	Unit
1	Optimized input parameter	$c_1(t+d)$	RM-CSTP	ROOM – Convector Setpoint	%
2		$c_2(t+d)$	RM-VENT	ROOM – Convector Ventilation	%
3	Input parameter	$x(t)$	RM-MTMP	Room – Measured temperature	°C
4		$x(t-d)$	RM-MTMP	Room – Measured temperature	°C
5		$x(t-2d)$	RM-MTMP	Room – Measured temperature	°C
6		$v_{y1}(t)$	AHU-TSET	AHU – Supply air temperature setpoint	°C
7		$v_{y2}(t)$	AHU-PSTP	AHU – Supply air duct static pressure setpoint	Pa
8		$v_{y3}(t)$	OUT-TEMP	Outside – Dry-bulb temperature	°C
9		$v_{y4}(t)$	OUT-WVEL	Outside – Wind velocity	km/h
10		$v_{y5}(t)$	OUT-WDIR	Outside – Wind direction	Deg
11		$v_{y6}(t)$	OUT-HUMI	Outside – Humidity	% RH
12		$v_{y7}(t)$	BD-OCCU	Occupancy rate	persons present per hour
13		$v_{y8}(t)$	RM-DTMP	Room – Desired temperature	°C
14		$v_{y9}(t)$	RM-SITU	Room – Situation	{off, stand-by, comfort}
15		$v_{y10}(t)$	RM-AIRQ	Room – Air quality	PPM (Parts Per Million)
16		v_{y11}	RM-SURF	Room – Surface	M^2
17		v_{y12}	RM-TYPE	Room – Type	Set of room types displayed in Table 4
18		v_{y13}	RM-CAPA	Room – capacity (People)	N
19		v_{y14}	RM-WIND	Room – Windows	N
20		v_{y15}	RM-NUMB	Room – Number	{text}
21		v_{y16}	RM-FLOR	Room – Floor	N
22		v_{y17}	RM-LOCA	Room – Location	{N, O, Z, W, G}
23	System output	$\hat{x}(t+d)$	RM-PTMP	Room – Predicted Temperature	°C

Table 18 Final set of selected parameters for the Temperature model.

Appendix I

#	Parameter type	Parameter	Name	Description	Unit
1	Optimized input parameter	$c_1(t+d)$	RM-CSTP	ROOM – Convector Setpoint	%
2		$c_2(t+d)$	RM-VENT	ROOM – Convector Ventilation	%
3	Input parameter	$y(t)$	RM-ENCS	Room – Energy consumption	WH
4		$y(t-d)$	RM-ENCS	Room – Energy consumption	WH
5		$y(t-2d)$	RM-ENCS	Room – Energy consumption	WH
6		$x(t)$	RM-MTMP	Room – Measured temperature	°C
7		$x(t-d)$	RM-MTMP	Room – Measured temperature	°C
8		$x(t-2d)$	RM-MTMP	Room – Measured temperature	°C
9		$v_{y1}(t)$	AHU-TSET	AHU – Supply air temperature setpoint	°C
10		$v_{y2}(t)$	AHU-PSTP	AHU – Supply air duct static pressure setpoint	Pa
11		$v_{y3}(t)$	OUT-TEMP	Outside – Dry-bulb temperature	°C
12		$v_{y4}(t)$	OUT-WVEL	Outside – Wind velocity	km/h
13		$v_{y5}(t)$	OUT-WDIR	Outside – Wind direction	Deg
14		$v_{y6}(t)$	OUT-HUMI	Outside – Humidity	% RH
15		$v_{y7}(t)$	BD-OCCU	Occupancy rate	persons present per hour
16		$v_{y8}(t)$	AHU-STMP	AHU – Supply air temperature	°C
17		$v_{y9}(t)$	RM-SITU	Room – Situation	{off, stand-by, comfort}
18		$v_{y10}(t)$	RM-AIRQ	Room – Air quality	PPM (Parts Per Million)
19		v_{y11}	RM-SURF	Room – Surface	M^2
20		v_{y12}	RM-TYPE	Room – Type	Set of room types displayed in Table 4
21		v_{y13}	RM-CAPA	Room – capacity (People)	N
22		v_{y14}	RM-WIND	Room – Windows	N
23		v_{y15}	RM-NCNV	Room – Number of convectors	N
24		v_{y16}	RM-NUMB	Room – Number	{text}
25		v_{y17}	RM-FLOR	Room – Floor	N
26		v_{y18}	RM-LOCA	Room – Location	{N, O, Z, W, G}
27	System output	$\hat{y}(t+d)$	RM-PECS	Room – Predicted Energy consumption	WH

Table 19 Final set of selected parameters for the Energy model.

Appendix J

The results of the temperature prediction model using five-time shuffle split cross-validation over the three modeling techniques: Linear Regression, Random Forest Tree and MLP-Neural Network are presented in Table 20, Table 21 and Table 22 respectively.

	1	2	3	4	5
Fit time	30.293	34.422	29.196	27.641	33.285
Score time	1.964	0.745	1.327	1.387	0.679
Test MAE	0.285166	0.285737	0.285501	0.285347	0.285292
Test R^2	0.896941	0.896498	0.896386	0.896789	0.896716
Test RMSE	0.403543	0.404541	0.404377	0.403809	0.403463
Train MAE	0.285344	0.285101	0.285197	0.285291	0.285276
Train R^2	0.896592	0.89674	0.896777	0.896643	0.896667
Train RMSE	0.403934	0.403601	0.403655	0.403845	0.40396

Table 20 Results of the temperature prediction model using Linear Regressor.

	1	2	3	4	5
Fit time	3330.559	3331.775	3330.788	3328.383	3326.652
Score time	222.774	222.444	223.509	222.639	224.395
Test MAE	0.077219	0.077524	0.077471	0.077563	0.077426
Test R^2	0.98542	0.985279	0.985145	0.985242	0.985344
Test RMSE	0.151783	0.152566	0.153113	0.152696	0.151983
Train MAE	0.072141	0.07207	0.072055	0.072044	0.072125
Train R^2	0.987116	0.987129	0.987177	0.987155	0.987121
Train RMSE	0.142583	0.142495	0.142274	0.142365	0.142614

Table 21 Results of the temperature prediction model using Random Forest Tree.

	1	2	3	4	5
Fit time	495.039	328.542	428.069	347.46	376.946
Score time	3.234	3.441	2.634	3.301	2.868
Test MAE	0.09874	0.101894	0.100616	0.100588	0.099868
Test R^2	0.978985	0.978237	0.978299	0.97802	0.978268
Test RMSE	0.182228	0.185502	0.185062	0.186347	0.185069
Train MAE	0.099104	0.1017	0.100592	0.100235	0.099936
Train R^2	0.978702	0.978241	0.978399	0.978163	0.978104
Train RMSE	0.183315	0.185271	0.184654	0.185627	0.185951

Table 22 Results of the temperature prediction model using MLP-Neural Network.

Appendix K

The results of the energy prediction in which the energy is based on the combination of energy extracted from the AHU and the energy used by the convector in the room. The results using a five-time shuffle split cross-validation with a Linear Regressor and an MLP-Neural Network are displayed in Table 23, Table 24 and Table 25 respectively.

	1	2	3	4	5
Fit time	31.128	29.781	30.427	22.436	25.858
Score time	1.082	1.899	0.864	1.972	3.214
Test MAE	497.104	497.6552	496.4304	496.5317	496.2143
Test R^2	0.367628	0.365437	0.367928	0.367934	0.366218
Test RMSE	1380.811	1386.037	1375.942	1379.925	1382.9
Train MAE	497.2645	496.5584	497.072	497.1937	497.051
Train R^2	0.366713	0.367446	0.366615	0.366611	0.367181
Train RMSE	1381.558	1379.812	1383.176	1381.852	1380.863

Table 23 Results of the energy prediction model using Linear Regressor.

	1	2	3
Fit time	2949.971	2951.112	2942.555
Score time	83.345	83.358	83.717
Test MAE	313.745	313.9973	314.375
Test R^2	0.660927	0.668308	0.665068
Test RMSE	1007.864	999.696	1006.361
Train MAE	288.1183	288.3544	288.2253
Train R^2	0.717409	0.715265	0.716333
Train RMSE	922.8308	925.4423	923.1531

Table 24 Results of the energy prediction model using Random Forest Tree.

	1	2	3	4	5
Fit time	2498.888	2715.535	2599.259	2141.742	2808.083
Score time	4.306	4.355	4.099	4.793	3.835
Test MAE	357.8244	363.6025	366.2726	365.6078	357.2891
Test R^2	0.598231	0.594035	0.598742	0.590214	0.596088
Test RMSE	1100.618	1108.618	1096.298	1111.1	1103.987
Train MAE	358.4744	362.7435	365.43	365.9145	357.269
Train R^2	0.596403	0.595264	0.601481	0.588969	0.597961
Train RMSE	1102.916	1103.716	1097.156	1113.176	1100.639

Table 25 Results of the energy prediction model using MLP-Neural Network.

Appendix L

The results of the energy prediction model using five time shuffle split cross-validation over the three modeling techniques: Linear Regression, Random Forest Tree and MLP-Neural Network are presented in Table 26, Table 27 and Table 28 respectively. Here the energy usage of the convectors is explicitly used as predictor. The energy from the AHU is excluded, due to obscure witnessed behavior in the data of the AHUs.

	1	2	3	4	5
Fit time	67.301	66.352	67.185	64.589	82.368
Score time	3.284	3.293	2.347	3.248	0.311
Test MAE	485.9285	486.9117	487.2326	485.5018	486.5509
Test $\mathbf{R^2}$	0.366399	0.368565	0.368838	0.369754	0.368163
Test RMSE	1364.501	1358.368	1361.212	1356.683	1361.032
Train MAE	485.985	486.6155	486.2074	486.9861	486.9102
Train $\boldsymbol{R^2}$	0.369156	0.368428	0.368338	0.368034	0.368563
Train RMSE	1356.433	1358.485	1357.533	1359.043	1357.595

Table 26 Results of the energy prediction model using Linear Regressor.

	1	2	3	4	5
Fit time	5061.333	5055.459	5060.497	5057.931	5052.127
Score time	102.492	103.473	104.28	102.649	103.575
Test MAE	300.1661	302.937	300.868	301.7356	301.4786
Test $\mathbf{R^2}$	0.67315	0.669613	0.673667	0.672451	0.672174
Test RMSE	972.89	980.691	974.244	978.981	981.0931
Train MAE	277.6765	276.5851	277.5187	277.2597	277.1469
Train $\mathbf{R^2}$	0.722319	0.722869	0.72182	0.72181	0.72243
Train RMSE	901.1785	899.5101	901.3387	900.4555	898.9301

Table 27 Results of the energy prediction model using Random Forest Tree.

	1	2	3	4	5
Fit time	2557.463	2766.16	3745.435	2202.801	2747.77
Score time	4.214	3.879	5.953	4.965	4.385
Test MAE	345.4625	346.5105	348.9382	352.2639	347.1081
Test $\mathbf{R^2}$	0.603066	0.606278	0.611821	0.599189	0.606552
Test RMSE	1080.003	1072.625	1067.509	1081.914	1074.012
Train MAE	345.1978	346.0112	348.5884	352.9943	347.9576
Train $\mathbf{R^2}$	0.605218	0.604742	0.611162	0.598468	0.605625
Train RMSE	1073.04	1074.69	1065.105	1083.294	1072.902

Table 28 Results of the energy prediction model using MLP-Neural Network.

Appendix M

The GEKKO Optimization Suite is an optimization software for mixed-integer and differential algebraic equations. It contains varies solvers that can solve linear, quadratic, nonlinear, and mixed integer programming (LP, QP, NLP, MILP, MINLP) equations. Different modes are included that allow data reconciliation, real-time optimization, dynamic simulation, and nonlinear predictive control. GEKKO is an object-oriented python library that facilitates local execution of APMonitor.

A part of the GEKKO Optimization Suite is the control technique Model Predictive Control (MPC) that can be used together with predictive modeling. The GEKKO structure that has been used in this research project is displayed in Figure 32. A requirement of GEKKO is that both the model and the MPC are constructed within GEKKO. An active interface between GEKKO and Scikit-learn does yet not exist. Although, Scikit-learn has an option to retrieve the structure and weighing of a trained model. This allows the possibility to reconstruct the structure within GEKKO, by reassembling the structure and setting the corresponding weights.

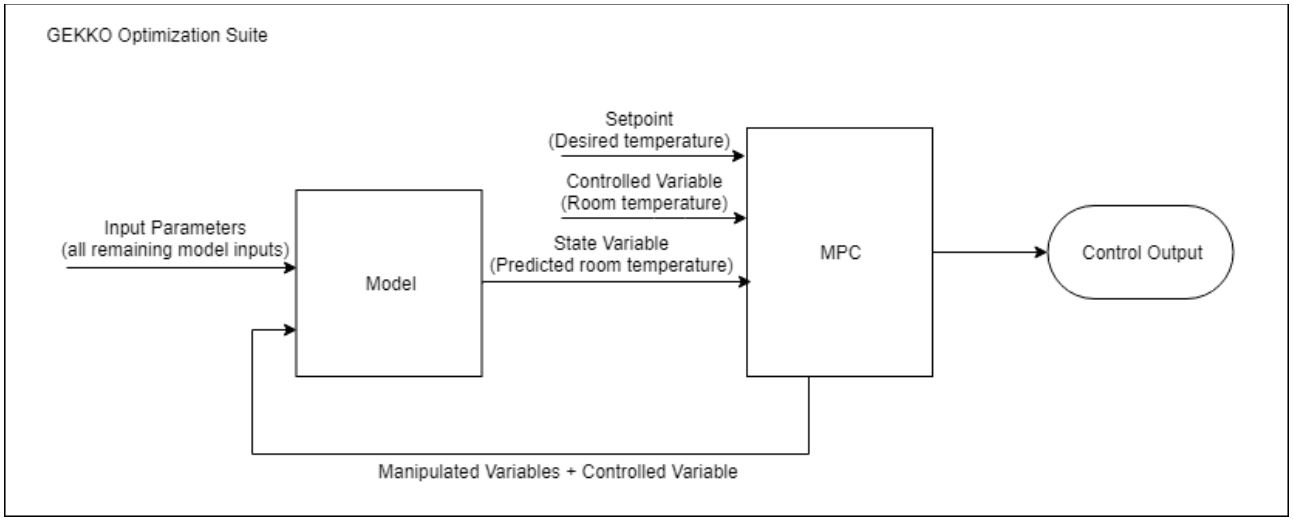


Figure 32 Structure within GEKKO.

As displayed in Figure 32 there are different types of variables visible. Each type of variables has a different functionality and contains a set of options. The function used and options of each of the displayed variables are briefly described.

- The **Manipulated variables** are the direct adjustable variables within the model. Within this research project the setpoints of the convectors within a room are the adjustable variables. The setpoints regulate the amount of heat or cooling the convector is producing in a room and the ventilation speed. The MPC will be able to modify these variables in order to get the controlled variable to its target. As part of the options of the manipulated variable is the ability to set boundaries in which the manipulated variable can be adjusted.

- The **Controlled variable** is the indirect adjustable variable within the model and represents the output state of the system. In our case study that is the temperature of a room. The variable is adjusted based on the output of the value of the model. Therefore, it is considered to be indirect adjustable. One of the critical options of the controlled variable is the setpoint or setpoint region. This/these value(s) describe(s) the goal or target that the controlled variable should reach and follow. When this is a single value, the goal is to keep the controlled value as close as possible to this specific value. Whenever a region of two values is given, the target is to stay within this region.
- The **State Variable** is the variable that contains the predicted value from the model. In our situation this is the predicted near future temperature based on the given parameters, current state of the controlled variable and the selected settings of the manipulated variable by the MPC system.
- The **Parameters** are all the input variables of the model that are not used within the MPC structure itself. In this research project these are all values that are presented in the modeling phase, except those values that are also part of the MPC structure parameters. These values are given by the system and are static.

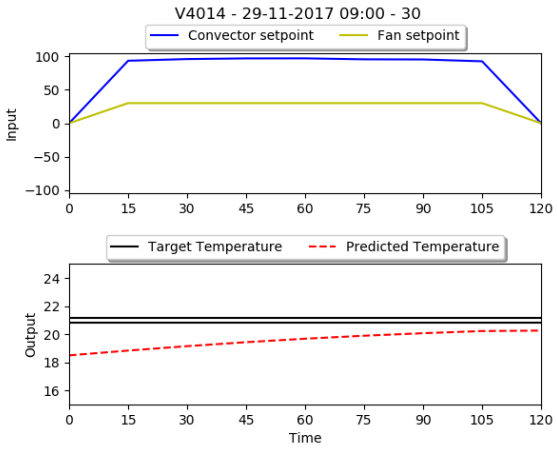
As presented in Figure 32 the coupling between the model and the MPC structure is essential. The output of the model is used as input for the MPC structure, whereas the output of the MPC is input for the model. The MPC structure is the dominant system of the two and uses the model to operate.

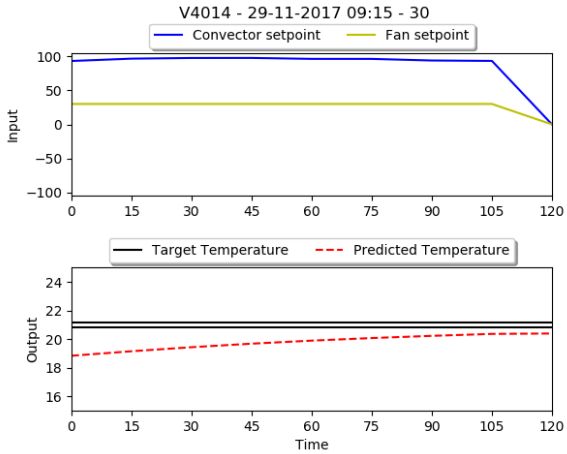
The MPC structure is displayed as a single box in Figure 32, but requires some additional settings.

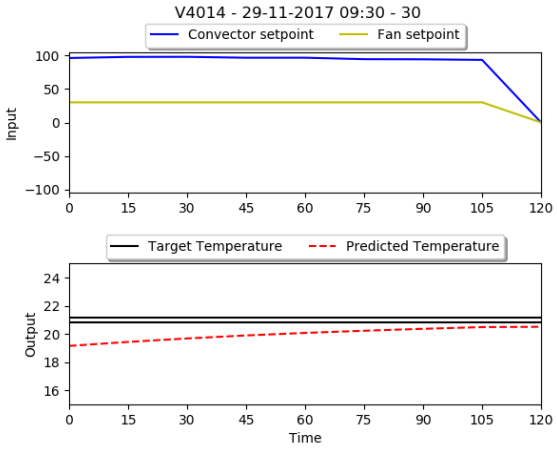
- The prediction horizon, control horizon (optional) and sample time have to be configured. Whenever the control horizon is not selected, the control horizon is set equal to the prediction horizon.
- An objective function should be included in the form of a minimization or maximization problem. This formula is often described as the cost function of the model. The cost function can be based on multiple objectives, where a weight determines the importance of each objective. In general, both the manipulated variables and the controllable variables are present in the cost function.
- The value of the controllable variables at each time unit is adjusted based on a differential equation.

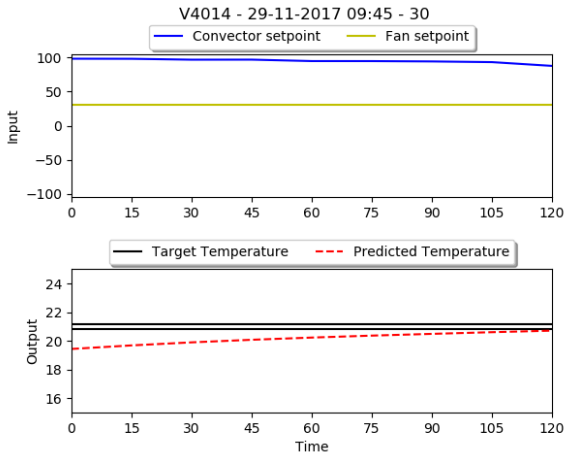
Appendix N

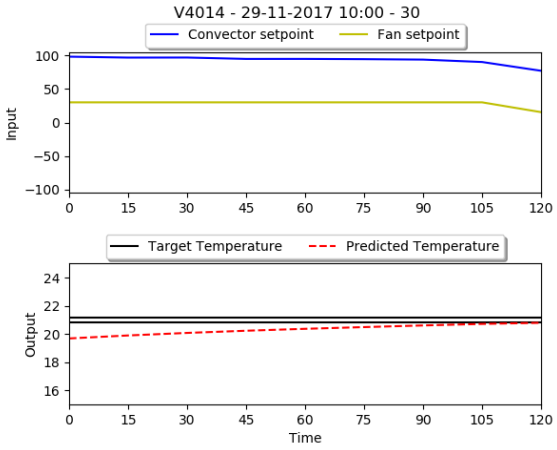
The individual constructed MPC graphs of room V4014 on 29 November 2017 between 9:00 and 12:00.

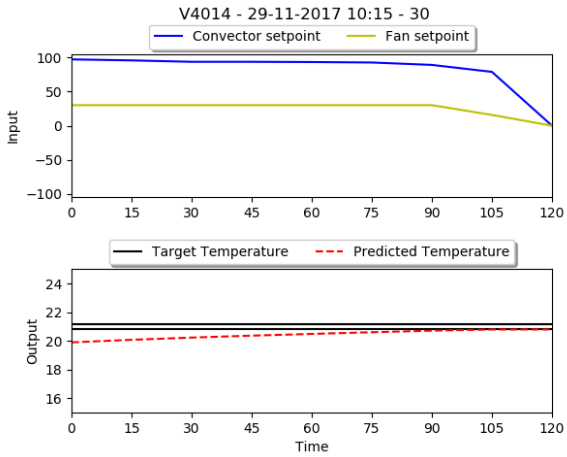


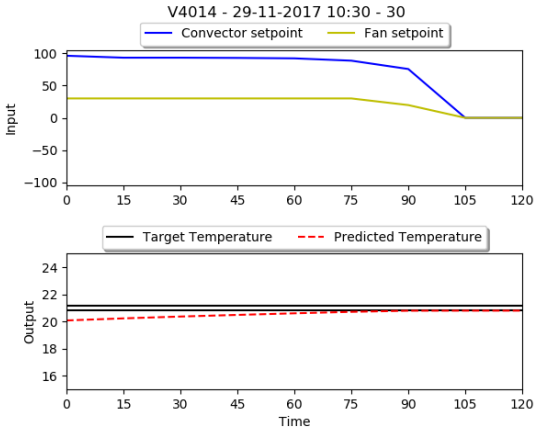


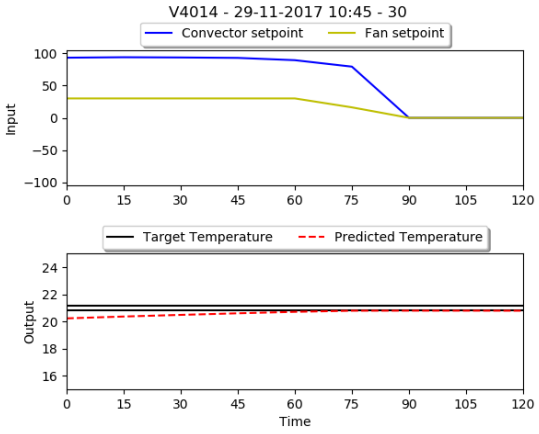


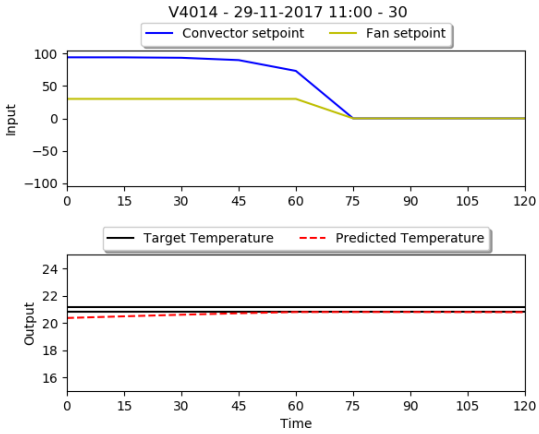


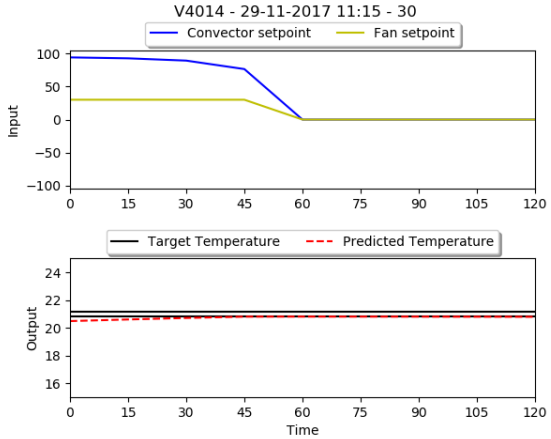


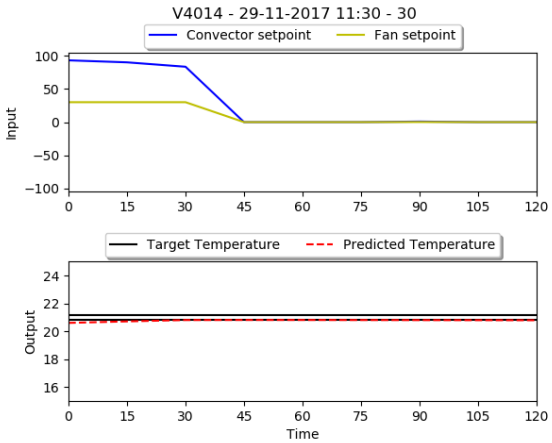


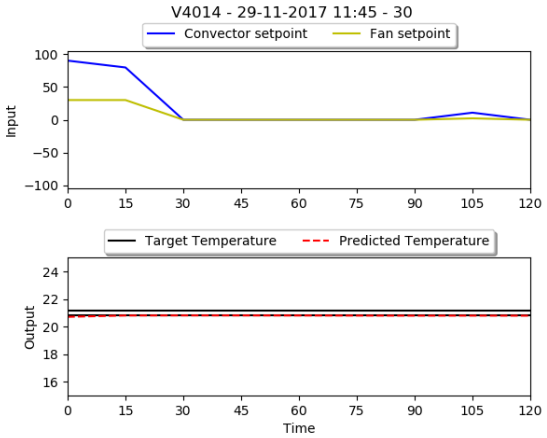

V4014 - 29-11-2017 10:30 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time


V4014 - 29-11-2017 10:45 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time


V4014 - 29-11-2017 11:00 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time

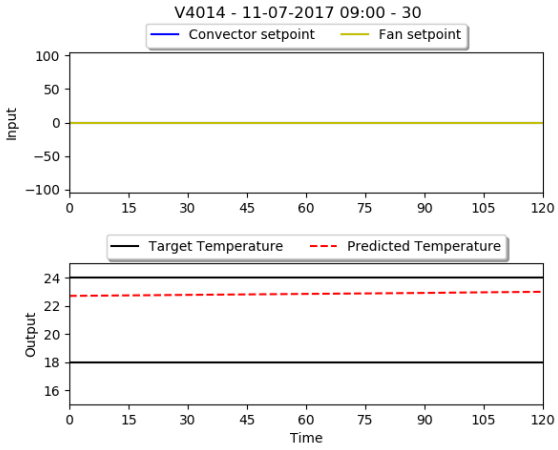

V4014 - 29-11-2017 11:15 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time

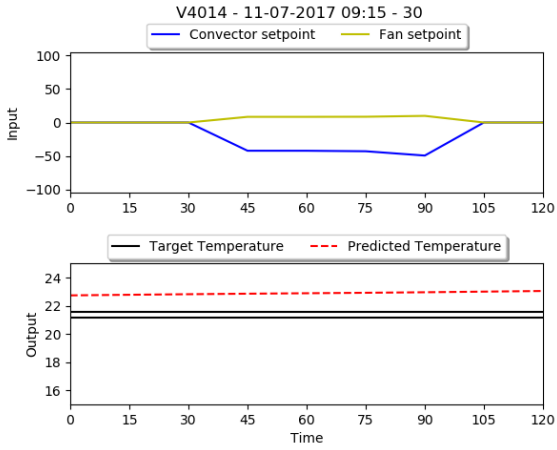

V4014 - 29-11-2017 11:30 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time

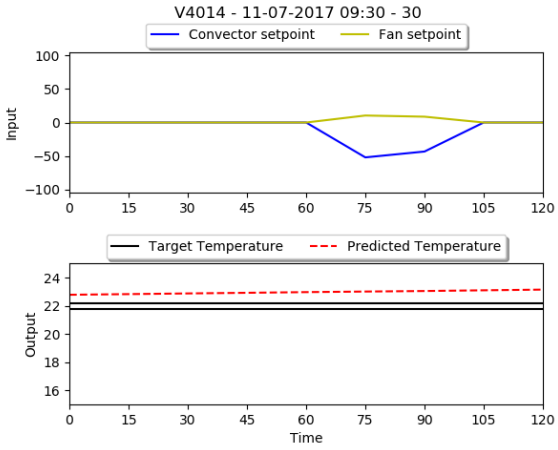

V4014 - 29-11-2017 11:45 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time

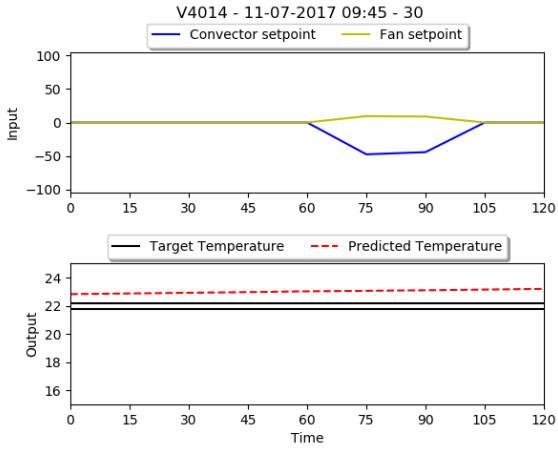
Appendix O

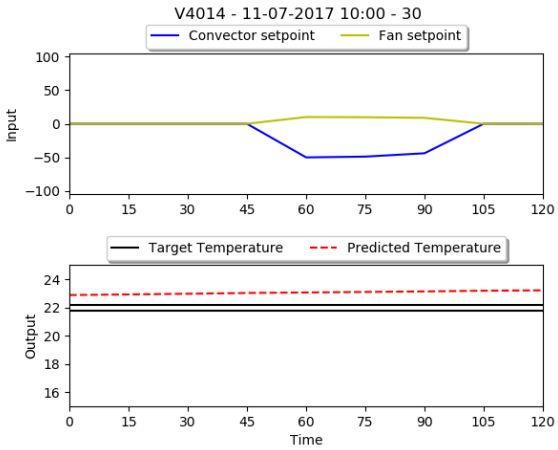
The individual constructed MPC graphs of room V4014 on 11 July 2017 between 9:00 and 12:00.

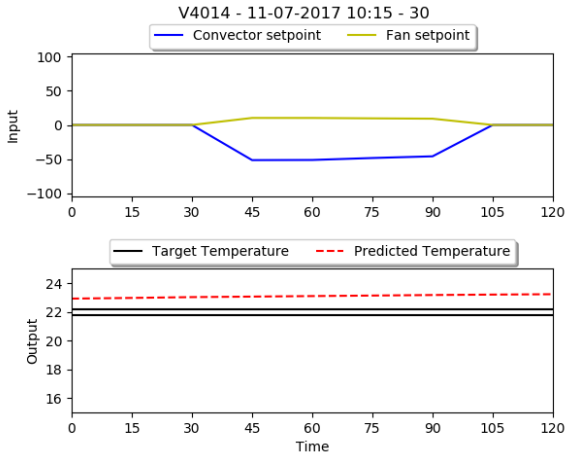


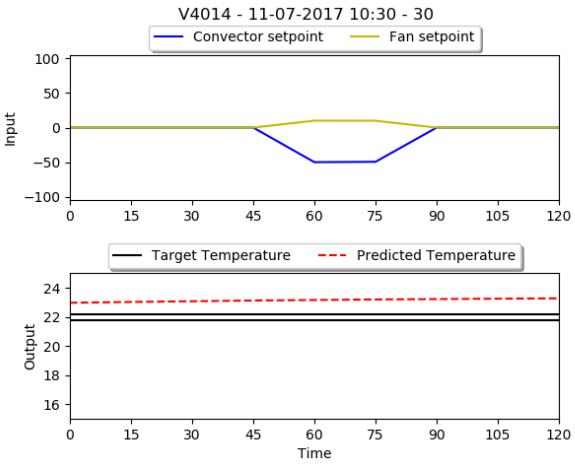


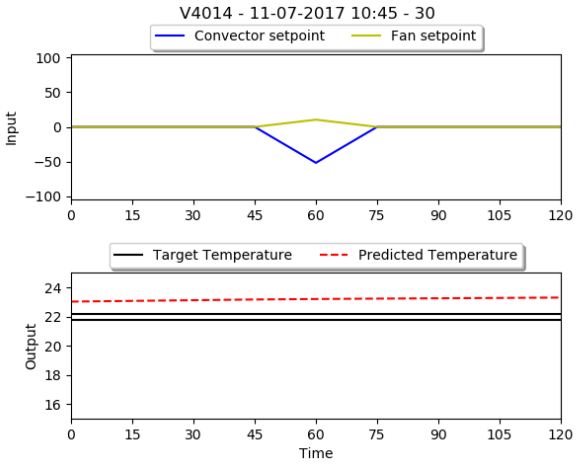


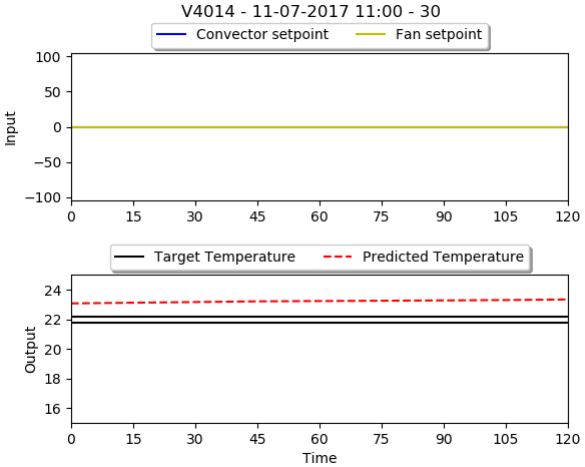


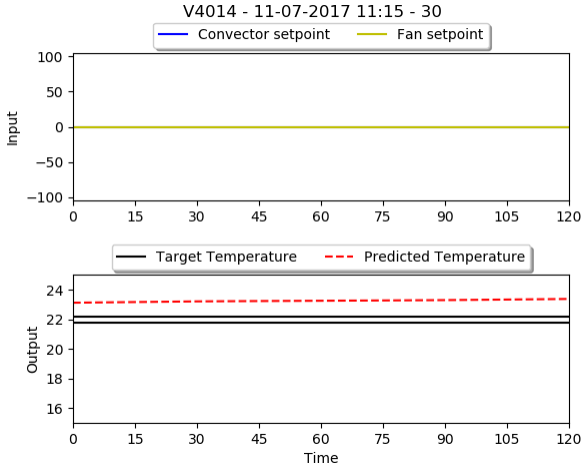


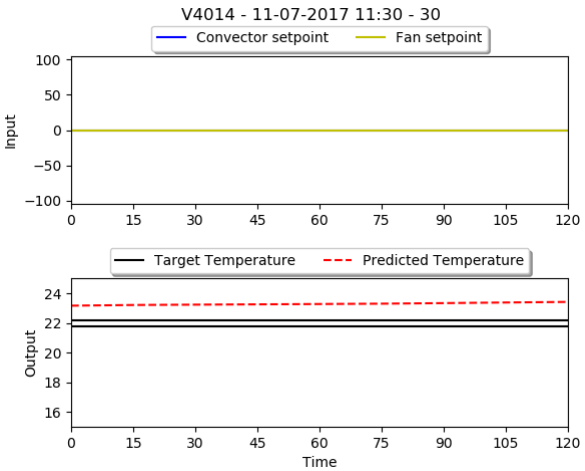


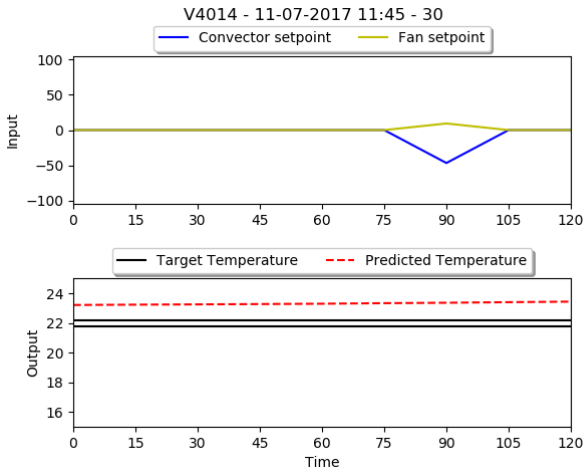

V4014 - 11-07-2017 10:30 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time


V4014 - 11-07-2017 10:45 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time


V4014 - 11-07-2017 11:00 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time


V4014 - 11-07-2017 11:15 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time


V4014 - 11-07-2017 11:30 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time


V4014 - 11-07-2017 11:45 - 30
Convector setpoint
Fan setpoint
Input
Target Temperature
Predicted Temperature
Output
Time