Short-term forecasting of the electric demand of HVAC systems

Mathieu Le Cam

A Thesis

In the Department

of

Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Building Engineering) at

Concordia University

Montréal, Québec, Canada

July 2016

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Mathieu Le Cam

Entitled: Short-term forecasting of the electric demand of HVAC systems

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Building Engineering

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Ming Yuan Chen — Chair

Dr. Louis Lamarche — External Examiner

Dr. Pragasen Pillay — External to Program

Dr. Andreas Athienitis — Examiner

Dr. Hua Ge — Examiner

Dr. Radu Zmeureanu and Dr. Ahmed Daoud — Thesis Supervisor

Approved by

Chair of Department or Graduate Program Director

Dean of Faculty

ABSTRACT

Short-term forecasting of the electric demand of HVAC systems

Mathieu Le Cam, Ph.D.

Concordia University, 2016

Heating, Ventilation and Air Conditioning (HVAC) systems of large buildings have a high contribution to the electric grid peak demand. During those periods, electric utilities face important mismatch issues in power supply and demand. In the context of Demand Response (DR) programs, there is a need from building energy managers for tools to forecast the electric demand of HVAC systems to plan for fast-DR control strategies. This thesis contributes to the DR research field by proposing a method for multi-step forecasting of the electric demand of existing HVAC cooling systems on the short-term in large commercial and institutional buildings.

Two forecasting methods are proposed: a cascade-based (global) method and a component-based method. The cascade-based method includes a sequence of forecasts of target variables. First, the air flow rate supplied by the AHUs is forecasted, followed by the cooling coils load, the cooling load of the whole building, and finally the electric demand of the primary cooling system is forecasted. The component-based method forecasts the electric demand of one component of the HVAC system such as a fan. Data-driven models are developed based on Building Automation System (BAS) trend data for electric demand forecasting of HVAC cooling system over the next six hours with a time-step of 15 minutes.

The large amount of data collected through the BAS presents a gold mine of information which could be used for better understanding of the actual building operation and performance. Data mining techniques are used as pre-processing steps to help in the development of the forecasting models, for the selection of regressors, to identify typical daily profiles of the target variable and to better understand the operation of HVAC systems. Different sequences of pre-processing steps are tested and their impact on the forecasting performance is compared.

The proposed forecasting methods are validated on two case studies: the Genomic research center on Loyola Campus of Concordia University and an office building located in Shawinigan-

Sud (Québec). The thesis compares the forecasts with measurements, and discusses the quality of forecasting results.

ACKNOWLEDGMENTS

I first want to express my gratitude to my supervisors, Drs. Radu Zmeureanu and Ahmed Daoud, for their guidance, and support through my doctoral degree. I am grateful to have had the opportunity to pursue my doctoral studies at Concordia University among the members of the Smart Net-Zero Energy Buildings Strategic Research Network. I greatly appreciated to have had the chance to stay for a few months among the researchers of the Hydro-Québec research institute and learn from their expertise.

The support of the Concordia Facilities Management, Luc Lagacé and Mathieu Laflamme was strongly appreciated. I thank them for sharing their knowledge of the building and volunteering their time to help us with data collection.

I would like to thank the researchers of the Hydro-Québec research institute in Shawinigan for their warm welcome at each one of my stays at the lab.

I thank my colleagues of the 16th floor for making this journey through the doctoral studies good memories.

I would like to acknowledge the financial support of Hydro-Québec, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Smart Net-Zero Energy Buildings Strategic Research Network (SNEBRN), and the Faculty of Engineering and Computer Science at Concordia University.

À mes parents,

pour m'avoir donné la chance de pouvoir choisir.

Table of contents

List of figures

List of tables

Chapter I. Introduction

I.1 Problem statement

The growth of energy consumption in the world is a significant issue. The International Energy Agency recorded an increase of 20% in the total energy consumption and 23% of the CO2 emissions during the last ten years (2003-2013) [1]. The buildings sector was responsible for about 41% of the primary energy use in the United States, in 2010 [2]. In Canada, in 2010, the total consumption of Commercial and Institutional (CI) buildings accounted for 12% of the secondary energy use [3]. Especially in office buildings, the Heating, Ventilation and Air-Conditioning (HVAC) systems, lighting and electric appliances are the main energy end-users.

During peak demand periods, utilities face important mismatch issues in power supply and demand, and for this reason the electric suppliers propose incentive programs to customers to change their electricity use. The building energy managers of large CI buildings involved in Demand Response (DR) programs modify their control strategies to reduce or shift the peak electric demand during the period of DR events. The HVAC system of large CI buildings presents a significant potential in demand response programs because: (1) HVAC systems are the main electric demand users of CI buildings; (2) the inertia involved in the thermal mass of the building allows to temporarily shut off or reduce the setpoints of HVAC systems; and (3) usually HVAC systems are almost completely automatically controlled through the Building Automation System (BAS) making easier to implement a different control strategy during DR events.

So far control strategies have been investigated with a one-day ahead notification from the supplier [4-6]. The interest focuses now on fast-DR with a notice a few hours before the DR event. In such a context, the building managers need tools to forecast the electric demand of the HVAC systems on a time horizon of a few hours.

I.2 Scope of the thesis

This thesis focuses on the short-term forecasting of the electric demand of HVAC systems in the context of DR programs. Data Mining (DM) techniques can bring useful information in the development of forecasting models for the electric demand of HVAC systems under real operating conditions.

The two main purposes of this thesis are as follows:

1) In most existing CI buildings, a very large amount of data is collected through the BAS. These measurements present a gold mine of information which could be used for better understanding of the actual building operation and performance. DM techniques present a high potential to extract knowledge about the as-operated conditions of the HVAC system from available measurements. The interest of this thesis focuses on this derived knowledge that is useful in the development of data-driven forecasting models, such as for the selection of regressors or for the identification of typical daily profiles of the electric demand**.**
2) The main purpose of this thesis is to provide the building energy managers with a forecasting tool that allows to assess the impact of some demand response strategies. The tool aims at forecasting the electric demand profile of existing HVAC systems over a horizon of prediction up to six hours. The forecasting performance of data-driven models is investigated for the electric demand of HVAC systems based on available measurements. A forecasting method is developed for the electric demand of the primary and secondary cooling system of existing CI buildings.

I.3 Overview of the thesis

The literature review, in Chapter II, is organized into four parts. In the first section, the process Knowledge Discovery in Databases (KDD) and Data Mining techniques are introduced. The section briefly summarizes and explains the existing DM techniques in each step of the KDD process. Examples of application of DM techniques in HVAC-related studies are presented. The second section of Chapter II presents a review of the literature on the published models for energy demand forecasting. The third section of Chapter II presents a discussion and comparison of

existing forecasting models. The advantages and disadvantages of each model are listed. The forecasting models are first compared based on published results. A preliminary analysis of the forecasting performance of a few models is then conducted: three inverse models are compared based on measurements from a case study. The last section of Chapter II presents the conclusions from the literature review and the objectives of this thesis.

Chapter III describes the proposed methodology for short-term forecasting of the electric demand of HVAC systems, in the context of DR strategies. Two forecasting method are briefly presented for the forecasting of electric demand of cooling systems: a cascade-based (global) method and a component-based method.

Chapter IV and Chapter V present the application of the cascade-cased forecasting method to a case study, the Genomic research center on Loyola Campus of Concordia University. Chapter IV describes the pre-processing steps performed in the cascade-based method. Chapter V focuses on the results of the forecast of the electric demand of the cooling system.

Chapter VI presents the application of component-based method to a second case study, an office building located in Shawinigan-Sud (Quebec).

Finally, Chapter VII presents the main contributions of this thesis and the future work.

Chapter II. Literature review

This chapter reviews first the studies conducted on the application of DM techniques in HVAC-related studies (section II.1.1) and their use for the help in the development of forecasting model (section II.1.2). A review of the literature is then performed on the existing models for the forecast of energy demand (section II.2). Different types of inverse models are compared in section II.3 based on published results and a case-study. And finally, the objectives of the thesis are presented in section II.4.

II.1 The process of Knowledge Discovery in Databases and Data Mining techniques

A very large amount of data is collected through the BAS in existing CI buildings. These measurements are a gold mine of information which could be used for better understanding of the actual building operation and performance, to optimize the control strategy of equipment or for fault detection. This section presents how data mining techniques allow to extract useful information from BAS trend data, for the development of forecasting models.

The terminological confusion between the terms of Knowledge Discovery in Databases and Data Mining is discussed in [7, 8]. DM corresponds to only one step of the process of KDD, in which information is extracted from available measurements. It comes along with four other steps: data selection, preprocessing, transformation and interpretation. When the DM techniques are used within the process of KDD, their performance is significantly improved [8].

The survey conducted by Kurgan et al. [7] presents the historical evolution of the knowledge discovery systems. The concept of process of KDD was first introduced in 1989 [9]. The first set of processing steps of the KDD process was proposed in 1996 by Fayyad et al. [10]. DM techniques can be classified into six groups (the first two for prediction and the rest for description): classification, regression, clustering, summarization, dependency modelling as well as change and deviation detection [10]. Several other processes of KDD emerged in the mid-1990s, such as the CRoss-Industry Standard Process for DM (CRISP-DM) [7]. Most models compared in [7] present

a similar sequence of steps. The most cited is the nine-step model presented by [10], which was implemented into a software called MineSet™.

II.1.1 Applications of DM techniques to HVAC-related studies

The application of DM techniques in such HVAC-related studies is still in the early stages. The first reported application of a KDD process to the building engineering field was presented by Buchheit et al. in 2000 [11] for the purposes of extracting trends in the energy performance of an HVAC system, and proving the value of the KDD process and DM techniques for application in building engineering.

DM algorithms have been employed to develop virtual sensors, which derive information from physical sensors, on variables that are either difficult or expensive to measure. Kusiak et al. [12] used four different DM algorithms to develop virtual models of indoor air quality from twelve measured variables. The best predictions were obtained by a neural network (NN) model with the mean absolute error (MAE) of 0.03°C for air temperature, 6.42 ppm for carbon dioxide and 0.11% for relative humidity.

DM techniques are useful to analyze occupants' behavior and provide recommendations for energy-saving opportunities. Yu et al. tested three types of DM techniques in [13]. A clustering analysis based on the K-means algorithm was performed on a database gathering information about residential buildings to group the buildings depending on their level of energy use. The occupants' behavior in each cluster can then be studied. A classification analysis based on a decision tree is developed to estimate the cluster attribution of new buildings. The structure of the decision tree also brings useful information to help understand building occupants' behavior. An association rule mining examines association and correlations among the user activities helping in providing recommendations for energy-saving opportunities.

DM techniques were investigated for the purpose of fault detection and diagnosis in HVAC systems. Du et al. [14] used two combined neural networks to detect anomalies in the supply air temperature control loop of the air handling unit. The two neural networks model the controlled variable (supply air temperature) and its most relevant variable (return water temperature). A fault is detected when the combined relative error between the observations and the predictions of the two networks exceeds a threshold value. A historical dataset of measurements including normal

and faulty operation is partitioned by the clustering algorithm into groups corresponding to normal operation and faulty variables. The clustering algorithm diagnoses the source of a new detected fault through an adaptive classification pointing at the faulty variable (e.g., cooling coil valve, temperature sensor …).

DM techniques can bring information about the consumption patterns of customers, characterize their demand profiles and classify new consumers. Ramos et al. [15] used four different clustering algorithms to identify typical electric load profiles of existing customers from a database of a utility company. They assessed the quality of typical daily profiles by using eight cluster validity indices. The best partition of daily profiles was obtained by the K-Means algorithm with four clusters, supported by five out of eight validity indices. The energy consumption corresponding to each cluster was obtained by using a rule-based modelling technique. The paper also presented rules for the automatic classification of new consumers according to those typical load profiles.

II.1.2 Application of DM techniques for the forecasting of building energy use

The performance of DM techniques is significantly improved when applied with the other steps of the process of KDD [8]. Only a few studies have applied the complete process of KDD for the purpose of forecasting building energy use; some are presented in this section.

DM techniques could help for input selection and data reduction, in the development of predictive models. Kusiak et al. [16] developed four inverse predictive models of the energy use by an AHU, by using DM techniques for the parameter selection and dimension reduction. This approach for the selection of independent variables with the highest impact had the potential of resulting in more accurate models. The forward neural network provided the best results compared with the other three models, and was coupled with a single-objective optimization model that was solved by the particle swarm optimization algorithm. Fan et al. [17] used DM techniques to select inputs, from a large set of measurements, for the forecasting models of the whole building next-day energy consumption and peak demand. The clustering analysis was used to identify and remove abnormal building energy consumption data. Eight different predictive models and one ensemble model (composed of the eight predictive models) were developed. The models were optimized using a genetic algorithm. The mean absolute percentage error (MAPE) for the next-day peak power

demand by the eight models was between 3.34% and 8.74%, while for the ensemble model was 2.85%. Kusiak et al. [18] used data mining algorithms to select the most appropriate independent variables, from a data set of measurements over three years, for the development of inverse models for the forecasting of the building steam load of the following year. They compared 10 different data mining algorithms, and concluded that a neural network model (MLP Ensemble) gave the best forecasts of daily steam load with the MAPE of 14.44% over the testing data set.

II.1.3 Conclusions on DM techniques for HVAC-related studies

Data mining techniques are used for two main purposes which are prediction and description. The extraction of information from patterns cannot be performed directly. The previous steps of the KDD process have to be conducted. They reduce the size of the dataset, clean the dataset and reduce the number of dimensions by selecting relevant regressors. The important issue in data mining is how the problem is defined and presented: how to reshape and pre-process the data so that they would have the appropriate frame to provide interesting results. For instance, when analysing the energy demand of a building, it could be more interesting to perform the analysis on the daily profiles of selected variables rather than on one-value measurement.

Interesting information about the operation of HVAC system and the building behaviour can be extracted by applying DM techniques on the measurements gathered through the BAS. For instance, by looking at the decrease of room temperature in perimeter zones during unoccupied periods: hints about insulation, infiltration and thermal inertia can be derived. The minimum amount of fresh air brought in by the system can be deduced by observing the coldest days when the system starts.

Over the past few decades, only a few studies have been conducted on the use of DM for building energy performance analysis and prediction. With the increasing amount of measurements available in buildings, there is an interesting potential in using DM techniques for the analysis of existing building operation. They can be applied for fault detection, electric demand forecasting, or also optimization of HVAC system operation to minimize energy consumption or peak demand.

II.2 Energy demand forecasting

The problem of modeling and forecasting the energy use in a building can be very complex due to the number and diversity of factors influencing it, such as HVAC system operation, thermal properties of the building envelope, weather conditions, or occupants and their behavior.

Electric load forecasting methods can be divided into three groups depending on the time horizon of the prediction [19-21]: for long, medium and short term load forecasting. The three groups of forecasting models defer from each other by their final utility. Long term forecasts present a prediction horizon of a few decades. They are used by electric utilities to manage the reserves of energy when planning to increase their generation or transmission capacities for instance. The medium term forecasts cover a period from a few weeks to one year. They allow the electric utilities to schedule the maintenance of their equipment and negotiate new contracts. The third group includes short term load forecasts giving predictions of the electric load over a horizon of a few minutes to a few days. Short-term forecasts are important in day-to-day operation of the equipment. They are useful for building energy managers to schedule the operation of HVAC systems during DR events. Our interest focuses on short-term forecasting for demand response application.

Three main groups of important factors were identified in the literature [19, 20, 22] for their impact on the short-term forecasting of electric load. These factors are classified as follows: time factors, environmental factors and type of customer. The time factors include seasonal trend, weekly and daily patterns as well as holidays. The behaviour of the occupants and the schedule of operation of the system is described by the time factors. These factors can characterize changes in the load between weekends, holidays and working days. The time factors are encoded to represent the hour of the day, the day of the week, and the season of the year for instance. The environmental conditions are identified as a factor of impact of the electric load. The following weather parameters can be relevant: outdoor air temperature, humidity, wind speed and solar radiation. They characterize the seasons of the year and fluctuations of climate conditions; they may affect the building load. Weather parameters are identified as the most important factors for short-term forecasting of the electric load in [19]. But, in some CI buildings, HVAC systems are still manually controlled and less affected by the environmental conditions. This scenario is handled in [23] by using a classification algorithm based on a working day schedule. Finally, the class of customers

is highlighted in [19] as an important factor impacting the electric load. Residential, commercial and industrial customers present different type of energy demand profiles and magnitudes. In addition to the other factors, an historical data set of previous values of the electric demand from one hour to several years should be considered. The previous electric demand values are representative of the building structure and characteristics (e.g. thermal envelope, HVAC system capacity…) and also of the behaviour of its occupants. The trend data of the electric demand over the time is of prime importance for modeling and forecasting the electric demand.

Diverse forecasting models were developed and tested in the past few decades. The most employed forecasting algorithms can be classified into (1) forward, (2) data-driven and (3) hybrid approaches; they are presented in sections II.2.1, II.2.2 and II.2.3. Literature surveys of popular predictive models covering the three groups are presented in [21, 24-26]. The surveys concluded on the need for refining the models from the BAS trend data to perform forecasts at the system level. Another issue is the use of artificial intelligence techniques for forecasting models development.

II.2.1 Forward models

The forward, or white-box, models are based on physical principles. They are mainly used at the design phase, for HVAC equipment sizing, for instance. Forward models use a comprehensive set of detailed physics-based equations to model building components and HVAC equipment. The white-box models can predict the behaviour of a whole building including energy consumption and thermal comfort. Their strong detailed physics-based equations allow the forward models to precisely describe the dynamics of the building.

The forward models have been widely used and improved for several decades. They were implemented into several building performance simulation software; some are presented in [27, 28]. The most used simulation software are DOE-2, EnergyPlus, ESP-R, TRNSYS, and eQuest among others. The simulation software can estimate the energy consumption of a building over a year. The simulation of a building can be performed in different location with dedicated weather files. Detailed reports about hourly building energy consumption can be generated. However, these forward models are extensively time-consuming to develop and solve. A significant amount of time is required for the users to obtain and set all the parameters and to compute the results.

Detailed information are required about the building structure, HVAC equipment installed, weather conditions, location and orientation. This knowledge is usually obtained through design plans, manufacturer data or on-site measurements if applicable.

This study aims at forecasting the energy demand of HVAC systems in existing buildings under as-operated conditions. The interest is not on the design and sizing of equipment but on the analysis of the building energy performance. The building performance simulation software are effective and accurate; they can be calibrated to fit well with measurements if available. Calibrated forward models can forecast the electric demand of HVAC systems and of the whole building. They do not adapt easily to a change in the HVAC system operation.

In the context of demand response, calibrated forward models are discarded for this work due to the time and detailed knowledge requirements about the building for their development and calibration, and limited flexibility if adapting to changes in HVAC systems. Statistical learning theory and artificial intelligence are in constant development and they present an alternative for future work in energy performance analysis and forecast of electric demand of existing buildings. The review of the literature focuses on inverse models in the following section.

II.2.2 Inverse models

The data-driven, also called black box, models are based solely on measurements. Their accuracy depends on the amount and quality of data available and data preprocessing. They are used to estimate the energy performance of a HVAC sub-system up to the whole building. They can model nonlinear phenomena and adapt to changes by training and learning based on new available observations. However, the black box models do not present a physical interpretation of results.

The inverse models can be divided between statistical and artificial intelligence methods. Statistical models include different types of regression and time series techniques. The artificial or computational intelligence methods gather fuzzy logic, evolutionary programing and Genetic Algorithms (GAs) as well as Neural Networks (NN) and Support Vector Machines (SVM).

II.2.2.1 Statistical models

Statistical models are traditional inverse approaches that have been widely used for the forecast of the electric load. The models correlate the electric load of the building with one or more

independent factors relevant to the building load. These empirical models are developed based on a history of previous values of the building load.

II.2.2.1.1 Parametric regression models

Regression analysis is commonly used for the forecasting of the electric load of a building [19, 20]. The regression models identify a relationship between the dependent variable of interest, the building energy consumption, and some independent variables called regressors in this study. The regressors are the important factors presented at the beginning of section II.2; they can represent the time, the climate conditions and historical electric load of the customer. The regression function correlating the electric load with its regressors can be of various kinds: linear or non-linear, univariate (using one regressor) or multivariate (using several regressors) …

The first statistical regression model used in the building engineering field for energy consumption was introduced by Fels in 1986 [29]. The Princeton Scorekeeping Method (PRISM) is a univariate linear regression model linking the building energy consumption to the daily average outdoor air temperature. In the PRISM method, the building energy consumption is correlated to one weather regressor, the outdoor dry bulb temperature. These specific models are also called energy signatures. Other regressors characterizing the climate conditions are added to the temperature in statistical regression models, such as the relative humidity or the solar radiations when available. They have showed a better performance, at estimating the value of the heat loss coefficient of the building for instance [30]. The inclusion of time factors, such as day-type index, as a regressor of statistical regression models improved their performance [31]. A nonlinear relationship between the building energy consumption and the regressors can be described by polynomial regression models. The polynomial models are still considered as a linear regression in terms of the unknown parameters (coefficient of regression) which are estimated based on the measurements.

The linear regression methods are easy and quick to implement. They do not require detailed information about the building structure and HVAC equipment. However they are based solely on data and a large dataset of measurements is required. It is quite difficult to develop a very accurate statistical regression model. These models cannot forecast a deviation in the electric load due to a sudden change in the climate conditions for instance. The forecasting performance of statistical regression models can suffer from autocorrelation or cross-correlation between the

regressors. These models are suitable for predicting an average energy consumption over quite long periods such as days, weeks or even months.

II.2.2.1.2 Nonparametric regression models

As opposed to parametric regression models, the nonparametric models are not restricted to a family of parametric functions, such as polynomial or trigonometric functions. When no parametric model obviously fit the data, a nonparametric regression is an option. It lets "the data speak for themselves". The nonparametric regression models are often called regression smoothers. There are many methods of obtaining a nonparametric regression models. A significant subgroup of regression smoothers are the kernel-based regression smoothers, called kernel smoothers [32]. The kernel smoothers are intuitive and relatively simple to understand mathematically. They were introduced by Rosenblatt in 1956 [33] and Parzen in 1962 [34].

The kernel smoothers have shown a good performance in predicting the energy demand of a building. A kernel smoother predicts the energy demand of a building in real time in [35] based on factors describing the time and climate conditions. The kernel smoother was tested on four building case studies and benchmarked against the ASHRAE Predictor Shootout dataset. It showed a better prediction performance in three cases out of four. The kernel smoothers require a large amount of observations to present a good performance. They accumulate the data and search through them to estimate the output. They can be time-consuming in case of very large dataset. An evolutionary local kernel regression algorithm is proposed in [36]. It can be parallelized to reduce the computing-time requirement. The evolutionary model presents a better performance in power prediction than usual kernel smoothers.

Kernel regression models are intuitive and simple to implement. They require to identify a few parameters: one for each selected regressor. The parameters are physically interpretable; they give some insight on the relevancy of each regressor to the building energy consumption. The kernel regression models require a significant amount measurements available and they can be time-consuming.

II.2.2.1.3 Time series models

Time series forecasting has been studied for decades in econometrics and finance market analysis. A comprehensive review of the time series forecasting models since the 1980s is presented in [37].

Time series data present a natural temporal ordering. The time series forecasting models assume that there is an internal structure in the historical dataset of the electric load of the building (e.g. autocorrelation). The time series of the electric load can be forecasted in the time domain or in the frequency-domain. In the time-domain, stationary and non-stationary techniques are distinguished. The stationary time series models assume that the mean, the variance and autocorrelation of the time series do not change over time. The future value in the time series of the electric load can be modelled by a linear combination of the previous loads. AutoRegressive (AR) and Moving Average (MA) processes are involved in these models. They can be merged into an autoregressive moving average (ARMA) process also known as Box-Jenkins model. ARMA models are developed in [38] to forecast the price of electricity over one week ahead based only on their past hourly values.

The non-stationary time series are transformed into stationarity series to apply ARMA models. A parameter is introduced corresponding to the number of time the series is differenced to achieve stationarity. This parameter leads to the AutoRegressive Integrated Moving Average (ARIMA) model. The monthly electric energy consumption of the residential sector in the eastern province of Saudi Arabia is forecasted with ARIMA model in [39]. Five years of monthly measurements are used by the ARIMA model to forecast the following year with an average percentage error of 3.8 %.

The ARMA and ARIMA models can be extended to take into account other independent variables, describing the climate conditions for instance. These independent variables are called exogenous variables. The stationary (ARMA) and non-stationary (ARIMA) models with eXogenous variables are respectively named ARMAX and ARIMAX. An ARMAX model is developed in [40] to forecast the daily electric load of Czech up to the next two days. The model takes into account the daily outdoor air temperature.

ARMA and ARIMA models show some limitations if there are nonlinear dependencies in the time series; if it presents a variation in its variance and covariance over time. Non-linear time series forecasting models are introduced, such as the AutoRegressive Conditional Heteroscedasticity (ARCH) model. ARCH models assume that the variance of the error at time t is a function of the previous errors. A generalized version of the ARCH model is developed in [41]

to forecast the wind power density of five wind farms in the United Kingdom. The ARCH model shows a good forecasting performance over long-term horizon (five years).

Time series can also be forecasted in the frequency-domain. The Fourier transform is frequently used to represent the time series in the frequency-domain. The Fourier series is a decomposition of the time series in a linear combination of sine and cosines. A Fourier series model estimates accurately the hourly heating and cooling energy use of commercial buildings in [42]. The model is able to represent the nonlinear dependence of the energy consumption against outdoor temperature and time. High frequency components have to be included in the model, in case of sudden changes in the energy consumption. This significantly increases the computational cost.

Time series models, in general, show a satisfactory forecasting performance over long-term prediction horizon. Other independent (exogenous) variables can be involved in the time series model and improve its performance. Nonlinear time series can model the non-linear dependencies of the building energy consumption with the time or the outdoor temperature for instance. Nonlinear models require an important number of parameters to identify. It requires a significant computational time when dealing with large data sets.

II.2.2.1.4 Kalman filters

The Kalman filter is named after Rudolph E. Kalman who introduced the algorithm in 1960 [43]. The Kalman filter is a recursive algorithm computing one-step-ahead forecasts. It is a linear model based on a discrete state-space representation. The Kalman filter is an explicit mathematical model that estimates the state of a system and minimizes the covariance of the error at each iteration. The Kalman filter recursively repeats the two steps: (1) predict the future state of the system and (2) update the model with new measurements and give feedback. Kalman filters were used extensively for guidance and navigation. They were applied in time series analysis as well for signal processing and econometrics. A few studies have applied Kalman filters for the forecast of electric load. In [44], a Kalman filter performs iteratively one-step-ahead forecasts of the hourly electric load of Crete island over the next day. The parameters are identified over a training period of ninety days. The model is tested over different conditions: starting at different times of the day and over two years.

II.2.2.2 Artificial Intelligence (A.I.) models

A review of the literature on AI models is presented in [45-47] and their application in energy-related studies for buildings. The algorithms are briefly presented and explained. The AI models present a large diversity due to the number of scientific fields involved: computer science, mathematics, neuroscience, psychology… The AI models applied in building-related studies can be separated in two classes [45]: Computational Intelligence (e.g. fuzzy systems, neural networks and evolutionary computation) and Distributed Artificial Intelligence (e.g. intelligent agent, expert systems). Krarti focuses on the use of neural networks, fuzzy systems and genetic algorithms for building energy systems [46]. A selection of applications is presented: weather forecast, short-term forecast of the electric load, fault detection and diagnosis, control of thermal energy storage systems or modelling of heat transfer through the building foundations. The review performed in [47] focuses on neural networks and their application, such as solar radiation and wind speed prediction as well as energy consumption in buildings.

II.2.2.2.1 Fuzzy systems

The fuzzy sets theory was introduced in 1965 by Zadeh [48]. The fuzzy sets are presented as classes with membership functions (fuzzy rules) which determine the affiliation of an element to a class (fuzzy set). The fuzzy systems try to mimic the human ability to face with uncertainty in judgement and decision fields. Traditional fuzzy systems are composed of three modules: “fuzzifier”, “interference engine” and “defuzzifier”. The three modules map the inputs to fuzzy sets and determine the output through IF-THEN rules that match fuzzy sets together or not.

Fuzzy systems were mainly used for fault detection and diagnostic applications in the building field. They have been combined with neural networks into fuzzy-neural models [49]. This hybrid model shows advantages of fuzzy systems and neural networks. The fuzzy system can deal with uncertainties and manage sudden changes in the inputs. The neural network adapts to the system to model and perform the forecasts. In the literature, fuzzy systems are usually as a complementary module in a hybrid model. An hybrid model using fuzzy systems and neural networks forecasts of the hourly electric load of Crete island over the next day in [44]. An entire year of hourly measurements of the load is used to train the model.

II.2.2.2.2 Evolutionary computation

Evolutionary computation originated in the 1950s, based on the theory of Darwin, for solving problems. They are heuristic techniques traditionally used for optimization and search problems. The evolutionary computation include nature-inspired algorithms, such as Genetic Algorithms (GAs), ant colony optimization and particle swarm optimization.

The GAs were introduced in the engineering field by Goldberg in 1981 to identify parameters of a simple mechanical system composed of mass, spring and damper [50]. These algorithms mimic the biological processes of natural evolution including mutation, inheritance, selection or crossover. These optimization algorithms present some advantages. No derivatives are required unlike in the gradient descent algorithm for instance. They perform global search and present less chances to find local minima. Evolutionary computation involves probabilistic rules rather than deterministic.

GAs have been mostly applied, in building–related studies, for model predictive control of HVAC systems or fault detection and diagnostic. GAs are usually combined with other forecasting models such as regressions or neural networks. They are usually used for an optimization process such as parameters identification in regression problems. A GA is used in [51] to minimize the energy consumption of a chiller during its operation. The GA tunes the temperature setpoints of the refrigerant at the output of the chiller. The energy consumption of the chiller for a given set point is estimated by a neural network trained on actual measurements.

II.2.2.2.3 Artificial Neural Networks (ANNs)

ANNs are originally based on the perceptron and the theory of the brain mechanisms, introduced by Rosenblatt in 1961 [52]. An overview of the different applications of ANNs in building-related studies is presented in [47]. ANNs have been extensively studied in the last twenty years and especially their application for the forecast of energy consumption in buildings. They have shown good skills in dealing with non-linear problems. Only a few examples of published studies are presented in the following paragraphs with some insight on the specific issues when developing an ANN for the forecast of building energy consumption.

An adaptive ANN is developed in [53] to predict the energy demand of a building. Previous studies were mainly using static models for building energy prediction. This work aims at

developing dynamic models that adapt to new measurements when available. Two training approaches are tested: an accumulative and a sliding window. The adaptive ANNs are retrained each time that new measurements are available, as opposed to the static models which are trained only once. In this study, the sliding window showed better results in terms of forecasting performance on the testing set, in the case of real measurements. The size of the window might be specific to one building, more investigations are required.

Another type of ANN is tested in [54] to predict the electric load of a building. The ANN presented in this study includes a module that forecasts the outdoor temperature at the following time step and use as input to the ANN. Most studies in the literature use variables describing the climate conditions for electric load forecasting. Usually only current and previous values of these variables are involved. The ANN is tested against the winners of the ASHRAE Predictor Shootout and showed a better performance with a coefficient of variation (CV) of about 1.5%.

A new approach is presented in [55] for the short-term forecast of the total electric consumption of a building. The total consumption of the building is divided into independent processes for each energy end-user (e.g. HVAC components). For each end-user, 96 ANNs are developed to forecast its electric consumption at each time-step of the day (15-min interval). The forecasts of each end-user are summed to get the profile of the total electric consumption of the building over the day. This study also focuses on the selection of the training period which significantly impact the performance of the ANN. A few interesting days are selected for training from the whole dataset; they include diverse conditions in terms of the day-type and climate conditions.

Two issues usually faced when using an ANN to predict the hourly energy consumption in buildings are studied in [56]. The problems discussed are the selection of the relevant inputs and the identification of the optimal architecture of the ANN (number of hidden layers and neurons). This paper performs statistical tests in two steps. First, all the potentially significant inputs are considered; hidden neurons are added one by one until the performance of the model stops improving. In the second step, redundant hidden neurons are identified and discarded as well as irrelevant inputs.

ANNs have also been applied for modeling the nonlinear behaviour of some HVAC components. For instance in [51], an ANN is trained to estimate the electric consumption of a four-compressor

chiller in different conditions of operation. Once the model is trained, a genetic algorithm is used to optimize the control strategy. The temperature setpoints of the refrigerant is tuned to minimize the energy consumption of the chiller estimated by the ANN.

ANNs have been widely exploited in the past few decades in energy-related studies of buildings. Their showed an ability to accurately capture nonlinear relationship between inputs and outputs from synthetic or real datasets. They can be trained based on examples from available measurements and generalize to forecast the future state of the system in new conditions. They can adapt in real-time to take into account new measurements. They can model complex systems with multiple inputs and multiple outputs. Neural networks models have shown accurate forecasting performance of energy consumption in buildings.

However, the accuracy of ANNs depends on the amount of measurements available and their quality. The forecasting performance of ANNs is also impacted by the data pre-processing steps including the selection of regressors. The training of the ANN is of prime importance; it covers among others, selecting the dataset period and avoiding over-fitting. When over-fitted, the ANN memorizes the noise in the measurements; it reduces its generalization skill. Another key-point is the difficulty in the design of the architecture of the ANN (number of hidden layers and neurons and the type of activation function). And the designed architecture is dedicated to a specific building or a regression problem; it must be investigated again each time.

II.2.2.2.4 Support Vector Machines (SVMs)

Support Vector Machines were introduced by Vapnik in the late 1960's based on statistical learning theory [57]. SVMs were initially developed for classification problems. They presented a very good forecasting performance and were extended to regression problems and time series forecasting, among other research fields. SVMs are increasingly used for prediction; around 3,500 papers have been published during the last decade. It is still less compared to the number of studies on ANN-based models. Some studies were conducted on the use of SVMs for building energy analysis. Examples are presented in the following paragraphs of published work using SVMs for predicting the thermal load or energy consumption in buildings.

SVMs are used in [58-60] to predict the hourly cooling load of a building. In the three studies, the prediction performance of SVMs is compared to ANNs based on synthetic data generated for an office building. The SVM shows better performances than ANN-based models

over the three studies with synthetic data. The researchers showed that the architecture of ANNs is difficult to design; there are no general rule. SVMs only require a few parameters.

More recently, SVMs have been utilized for the forecast of electric load. For example in [61], an SVM predicts the yearly electric consumption of Taiwan from 1985 to 2003 based on the consumption from 1945 to 1984. A simulated annealing optimization algorithm is used to identify the three parameters of the SVM. At the scale of a building, SVMs are presented in [62] for the forecast of the energy consumption of four buildings located in the tropical region of Singapore. The model is trained over three years of monthly data of the energy consumption for four commercial buildings. The performance of the SVM over the following year is less than 3% in terms of coefficient of variation of the root mean square error.

Studies have been conducted lately on forecasting electric demand using measurements with a shorter time interval. Hourly electric load are predicted in [63] and [64] for a city in the province of Hunan, China, and respectively for a province in Mongolia. Two SVMs models were compared in [65] for the forecast of the daily peak of the electric load in eastern Slovakia over the next month. The models are trained based on half-hour load data of the two previous years and averaged daily temperature. The identification of the parameters of SVMs showed a significant impact on the accuracy of the models. Two approaches for identifying the parameters are tested in the two models: a grid-search approach and a less time-expensive Bayesian framework.

Few studies were performed on the application of SVMs for regression in HVAC-related studies. A very few examples were found in the literature. For instance, SVMs are involved in [12] to model a virtual sensor representative of the indoor air quality. They can model the complex nonlinear relationships between the involved variables (temperature, relative humidity and CO2 level) for air quality assessment. A SVM is utilized in [66] to predict the cooling load of a chiller and estimate its electric demand.

SVMs showed promising results in solving nonlinear problems and for time series forecasting. SVMs presented a forecasting performance at least as good as ANNs. They can be performed with a small amount of data available and require only three parameters to be identified. The parameters show a significant impact on the accuracy of the predictions. The approach for identifying the parameters requires more investigations. The application of SVMs to HVAC-related studies is a promising research topic.

II.2.3 Hybrid modeling

The hybrid models combine two or more forward and/or data-driven methods to benefit from their respective advantages. Grey-box models are hybrid models using forward and inverse models. For instance, a grey-box model can be based on physical laws using parameters that are calibrated with available measurements. Hybrid models can also combine two or more data-driven approaches. Examples of each type of hybrid models are given in sub-sections II.2.3.1 and II.2.3.2.

II.2.3.1 "Grey-box" modeling

Grey-box or semi-physical models are based on physical principles. They also are data-driven because the parameters are calibrated using measurements. The grey-box models stand between the white-box models, completely based on physical equations, and black-box models, learning from the data. The grey-box models are physically understandable and they fit to real measurements.

Grey-box models are popular for modelling thermal systems with lumped capacitance models. They have been widely applied for modelling the thermal exchanges in a building and with its environment [67-69]. In this case, the grey-box model is called a thermal network involving parameters, such as thermal capacitance and resistance. The parameters are identified from available measurements.

A thermal network is developed in [67] a for the prediction of the building cooling load and indoor temperature. The thermal behaviour of the building is modeled by a thermal network with thirteen thermal resistance and eight capacitances. Different combinations of thermal resistance and capacitance are considered for internal and external walls, ceilings and floors and windows. A non-linear regression algorithm tunes the parameters of the thermal network to fit with the measurements of the indoor temperature. The study concludes that one to two weeks of data are required to train the model.

Another thermal network approach is implemented in [68, 69] to model the thermal load of a building. In both studies, a lumped capacitance model with three thermal resistances and two capacitance (3R2C) is used for the building envelope (roof and external wall) and a 2R2C-structure for the internal mass. The parameters of the thermal network are tuned by a genetic algorithm in [69]. Two modules based on empirical inverse models are added in [68] for the prediction of

weather conditions (solar radiation and outdoor temperature and relative humidity). The grey-box model finally predicts the hourly cooling load of the building on the next day.

A simplified grey-box model was proposed in [70] to model reciprocating chillers with a minimum number of parameters. A grey-box steady-state model of an air-cooled chiller is presented in [71]. Four models are developed based on physical equations; one for each sub-component: compressor, condenser, valve and evaporator. The grey-box model of the chiller was able to predict the total cooling and electrical energies consumption over about a month with a good accuracy.

Grey-box models present good skills for modelling thermal loads and energy consumption of a building. They combine the advantages of forward and inverse models. They are physically understandable and they are calibrated to fit measurements of an existing system.

II.2.3.2 Hybrid models combining statistical and/or Artificial Intelligence approaches

Hybrid models can also be defined as models joining at least two statistical and/or artificial intelligence algorithms. Many studies performed these AI hybrid models for the forecast of electric load. The AI hybrid models mostly involve ANNs or SVMs as the basis forecasting model because of their good performance. In a consequent number of studies the ANN-based hybrid models are associated with fuzzy rules. It gives the possibility of including human knowledge in the model. The SVM-based models are generally combined with nature-based heuristics, such as genetic algorithms, to identify the learning parameters of the model. Some published studies of ANN-based and SVM-based models for forecasting electric load are presented in the followings.

II.2.3.2.1 ANN-based hybrid models

Neuro-fuzzy approaches combining an ANN-based model with fuzzy if-then rules were introduced Jang in 1993 [49]. The neuro-fuzzy models make possible to include human knowledge in the model through if-then rules along with empirical knowledge on the system learned from the data. The fuzzy rules developed by the hybrid model make it more understandable to humans. Neuro-fuzzy models are presented in [72, 73] for the forecasting of the energy consumption of an hotel and respectively a library building in China. The neuro-fuzzy models are improved using a genetic

algorithm to optimize the number of fuzzy rules and the structure of the network. They showed a better forecasting performance compared to ANN [73].

A different ANN-based hybrid model is used in [74] to estimate the energy consumption of the residential building sector in Canada. The hybrid model, the Canadian Hybrid Residential End-use Energy Model (CHREM), uses an ANN to predict the energy consumption of electric appliances, lighting and domestic hot water. These components of the energy consumption are strongly correlated to the behaviour of occupants. The ANN is coupled with a forward model developed in ESP-r that estimates the energy consumption associated to space heating and cooling in the residential houses.

II.2.3.2.2 SVM-based hybrid models

Two main types of SVM-based hybrid models were identified from the literature, coupling SVMs with times series models or heuristic algorithms. A time series model (ARIMA) is used in [75] to forecast the hourly electric load of a city in China over one day. A SVM model is coupled to the ARIMA model and corrects the deviation of the forecast by the time series model. The ARIMA model takes into account only the past values of the load; the SVM includes the nonlinear part of the load sensitive to the environmental conditions. An ARIMA model is combined with an SVM in [76] to forecast the daily peak load of an electric utility in Iran over one month. The ARIMA models the linear component of the load and the SVM is used for the non-linear component. This hybrid model allows to capture the linear and nonlinear components of the load and guarantees a good forecasting performance in any condition of the load.

SVMs can also be combined with evolutionary algorithms to optimize the parameters of the SVM model giving the best forecasting performance. A genetic algorithm tunes the parameters of a SVM in [77]. The SVM is tested on the forecast of the electric load of the wide area of Taiwan. Another heuristic algorithm is utilized in [78] to identify the best set of parameters for the SVM: a particle swarm optimization algorithm. The SVM model forecasts the monthly electric load of a province in China over a year.

II.3 Discussion and comparison of inverse models

This section compares different types of inverse models based on published results in sub-sections II.3.1 and II.3.2. A preliminary comparison of the forecasting performance of three inverse models is performed in sub-section II.3.3 based on a case-study. This comparison analysis was published in a conference paper [79].

II.3.1 Advantages and disadvantages of data-driven models for energy forecasting

Most of the inverse models present the following advantages as opposed to forward models. They do not require detailed knowledge about the building characteristics and HVAC equipment. They are developed to fit to measurements and require less time to generate: the calibration step is embedded in the development. However most inverse models are black-boxes which do not present any physical interpretation.

From the review of the literature, the advantages and limitations of the different classes of data-driven forecasting models are listed in Table II.1 for the traditional statistical models and in Table II.2 for the AI models.

Table II.1 Advantages and limitations of statistical models

	Advantages	Disadvantages
Parametric Regression	▪ Simple model requiring few parameters: one for each regressor. ▪ Light computational time. ▪ Easy and quick to implement. ▪ Easily understandable. ▪ Performs multivariate regression. ▪ Handles nonlinear relationships by transformation of the regressors. ▪ Can be based on physical equations: becomes a grey-box model. ▪ Suitable for predicting an average energy consumption over quite long periods (days, months, years).	▪ Limited in prediction accuracy. ▪ Not generalizable: do not adapt to new conditions. ▪ Cannot handle sudden and significant changes in the regressors or the electric demand. ▪ Requires large datasets of measurements covering different types of condition of operation.
Nonparametric Regression	▪ Accurately performs non-linear multivariate regression without prior knowledge about the inputs. ▪ Few parameters: one for each regressor. ▪ Simple to implement. ▪ Good generalization skill. ▪ Parameters can be physically interpreted.	▪ Requires a significant amount of measurements for a good accuracy ▪ Time-consuming in computation for very large datasets
Time Series	▪ Satisfactory forecasting performance over long-term prediction horizons. ▪ Forecasts the electric demand of the building multiple step ahead. ▪ Time series of regressors can be included. ▪ Handles nonlinear variation of the time series	▪ Many parameters to identify. ▪ Time-consuming in computation to identify all the parameters. ▪ Handles with difficulty sudden and significant changes in the regressors of the electric demand.
Kalman filter	▪ Optimal one-step ahead predictor: minimizes the error at each iteration. ▪ Adaptive: new measurements processed in real time when available. ▪ Filters the data to remove noise. ▪ Extended Kalman filter to model nonlinear relationships. ▪ Based on physical equations.	▪ Complex modelling: cannot simply fit to measurements. ▪ Difficult to calibrate. ▪ Less community knowledge for time series forecasting.

Table II.2 Advantages and limitations of Artificial Intelligence models

	Advantages	Limitations and Drawbacks
Artificial Neural Networks	▪ Accurately performs non-linear multivariate regression without prior knowledge about the inputs. ▪ Learn complex nonlinear relationships from examples. ▪ Present a generalization skill: adapt to new conditions. ▪ Can extract empirical knowledge from very large dataset with noise.	▪ Black-box model: no-physical explanation ▪ Requires a significant amount of data available. ▪ Present the tendency to over-fit the data: memorize noise. ▪ Architecture difficult to design: no general rule. ▪ Architecture dedicated to one building and one application: degrade the generalization skill. ▪ Time-consuming to perform an efficient training.
Support Vector Machines	▪ Accurately performs non-linear multivariate regression without prior knowledge about the inputs. ▪ At least as accurate as ANNs. ▪ Require only three parameters to be identified. ▪ Small amount of data required for training. ▪ Present a good generalization skill to adapt to new conditions.	▪ Black-box model: no-physical explanation

II.3.2 Comparison of the prediction accuracy of inverse models based on published results

A comparison in terms of forecasting performance between the different classes of inverse models is performed based on published case studies from the literature. The accuracy of the models presented in the literature on case studies is presented in Table II.3 for the statistical models and in Table II.4 for the AI models. The forecasting performance of the models is described in terms of Mean Absolute Percentage Error (MAPE), Normalized Mean Squared Error (NMSE) and Coefficient of Variation (CV) over the testing set (equations [II.1] to [II.3]).

$$MAPE = \frac{1}{n} * \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad [II.1]$$

$$NMSE = \frac{n * \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} \hat{y}_i * \sum_{i=1}^{n} y_i} \quad [II.2]$$

$$CV = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} / \bar{y} \quad [II.3]$$

where $\hat{y}_i$ is the forecasted value, y_i is the measured value and n is the number of observations.

Table II.3 Comparison of prediction accuracy of statistical models on published case studies

Class	Model specification	Case study	Ref.	Data available for training	Prediction horizon	Accuracy		
						MAPE	NMSE	CV
REGRESSION	Kernel regression	Four buildings (library, offices and labs)	[35]	Hourly electric load over one year and a half	~			10.86%
	Kernel regression	Electric load provided by the electric utility ComEd	[80]	Hourly electric load over one year	From 1 to 48 hours with a window of 24h	2.6% (1h) 4.1% (2d)		
	Kernel regression	Electric load provided by the electric utility Hydro Quebec	[81]	Hourly electric load over the previous 300h	Up to 120 hours	0.87% (1d)		
TIME SERIES	Smooth Transition Periodic Auto-Regressive	Electric load of an Australian province	[82]	Half-hourly electric load over four years	Up to seven days	3.42%		
	Periodic Auto-Regressive	245 time series of electric load of utility substations in Belgium	[83]	Hourly electric load over four years	One hour	<5%	<5%	
	ARMA with eXogenous input	Czech electric utility	[40]	Hourly electric load over two years	One day	1.4%	1.2%	
	ARIMA					2.5%	3.5%	
	Modified Generalized Auto-Regressive Moving Average	Brazilian electric utility	[84]	Hourly electric load over two years	From one to seven days	4.4%		
KALMAN FILTER	~	Electric load of Crete island	[44]	Hourly electric load over two years	From 1 to 24h	3.77% (1h) 13.12% (1d)		
	~	Canada electric utility	[85]	Hourly electric load over two years	Prediction over the next 24h	<1%		

Table II.4 Comparison of prediction accuracy of AI models on published case studies

Class	Model specification	Case study	Ref.	Data available for training	Prediction horizon	Accuracy		
						MAPE [%]	MSE	CV [%]
ARTIFICIAL NEURAL NETWORKS	Static training	Electric demand of a chiller	[53]	Hourly electric demand over 130 hours	24h			0.23 - 0.26
	Dynamic training with sliding window							0.26
	Dynamic training with accumulative window							2.53
	1-step predictor	Dataset 1: The Great Building Energy Predictor Shootout I	[56]	Hourly electric demand of a building over six months	24h	1.50		2.39
	24-step predictor					3.67		5.47
	1-step predictor	Dataset 2: office building located in Athens, Greece.		Hourly measurements of electric power consumption over 1 year		2.57		2.97
	24-step predictor					13.39		12.58
	Recurrent neural network	Dataset 1: The Great Energy Predictor Shootout I	[54]	Hourly electric demand of a building over six months	Two months	1.95		2.55
		Dataset 2: The Great Building Energy Predictor Shootout II		Hourly electric demand of a building over 21 days	One week			1.44
	96 models: at each time-step in a day	University in Italy	[55]	Building electric demand at 15-min time-step over a year	24h	6.45		
SUPPORT VECTOR MACHINES	~	Energy consumption of four office buildings in Singapore	[62]	Monthly electric consumption over two years	From one month to one year	<4		<3
	~	Electric load of Taiwan	[61]	Yearly electric consumption over 40 years	From one to nine years	1.76		
	~	Electric load of a city in China	[63]	Hourly electric load over one month	48h step-by-step		1.26	
	~	Electric load of a province in China	[64]	Hourly electricity measurements over 22 months	24h	<3		

II.3.3 Comparison of three inverse models based on a case study

In this section, three data-driven models are selected from the previous comparison based on published studies from the literature: model 1 Kernel regression, model 2 dynamic ANN and model 3 SVM for Regression called SVR. The forecasting performance of the three models is compared on a case study: the electric demand of a cooling plant of a university campus. Measurements a t15-min time-step are gathered through the BAS and include historical electrical demand, operating conditions of two chillers and local environment variables over three cooling seasons (from April to September 2009 and 2010, and from April to June 2011). The chillers have a design cooling capacity of 3165 kW and electric demand of 550 kW. The monitored data have already been used in a study carried out by other researchers [86]. Their results are used for comparison in this study.

Four regressors are selected to forecast the electric demand of one chiller. They have been chosen based on their statistical correlation to the electric demand of the chillers: the hour of the day, the outdoor temperature and the supply chilled water and condenser water temperatures.

II.3.3.1 Presentation of the inverse forecasting models

The first two models could be qualified as dynamic in the sense that, at each time step, they use previous and current values of the regressors (e.g., at time t-1, t) as well as previous values of the dependent variable to forecast the dependent variable at time t; they are called Nonlinear AutoRegressive models with eXogenous inputs (NARX). In the first approach, models 1 to 3 are compared to the Multi-Polynomial (MP) inverse model presented in [86], a static model, where the regression coefficients are calculated once based on previous data of training set; the electric demand at time t is estimated based on the current values (at time t) of regressors. Model 3 can also be considered as a static model. The comparison of predicted performance of those four inverse models is carried out on a 24-hour testing set with a training set size varying from 7 to 21 days.

II.3.3.1.1 Kernel regression model

The estimated future electric demand is a weighted-average of the past values of the demand in similar conditions. The similarity of the conditions is defined through parameters called the bandwidth. The accuracy of the model depends on the value of the parameters and the amount and

diversity of the training set. A cross-validation step is performed to define bandwidths while minimizing the prediction error over a given training set.

II.3.3.1.2 Dynamic ANN model

A dynamic ANN is implemented in Matlab Neural Network Toolbox 7. The ANN is composed of four input neurons, one hidden layer with ten neurons using sigmoid transfer functions; and one neuron in the output layer using a linear transfer function. The model also includes a 24-step delay of the regressors and electric demand. The ANN is trained over a time period from 7 to 21 days using the Levenberg-Marquardt algorithm which tunes the weight of each neuron. The closed-loop ANN is set to use the forecasts as delayed input.

II.3.3.1.3 SVR model

A SVR model is developed in Matlab libSVM toolbox [87] for the forecast of the electric demand of the chillers. A radial basis kernel function is used to map the regressors to a higher-dimensional space. A linear regression is performed in this space corresponding to a nonlinear regression in the initial space. The regression coefficients are tuned by minimizing a loss function composed of two terms: one characterizing the level of complexity of the model and the other one describing its accuracy. A trade-off between the two terms is determined: this is the principle of Structural Risk Minimization. Three parameters ($C, \varepsilon \; and \; \gamma$) related to the loss function and kernel function are tuned using a cross-validation procedure over a training set varying from 7 to 21 days.

II.3.3.2 Discussion of results

The parameters of inverse models 1 and 3 are tuned by a genetic algorithm to minimize the prediction error over the training set. The optimization of the weights in model 2 are performed by the Levenberg-Marquardt algorithm. Once trained, the inverse models can be tested to forecast the electric demand of the chillers.

The prediction accuracy of the inverse models is presented in Table II.5 for the prediction of the electric demand of the chillers over the next 24 hours, at a 15-minute time-step. The training set size is varied from 7 to 21 days to analyze its impact on the prediction accuracy. Figure II.2 shows the predicted and measured electric demand profiles of the chiller over the testing set. The Coefficient of Variation (CV) and Root Mean Square Error (RMSE) (equation [II.4]) over the training and testing set are displayed in Table II.5.

$$RMSE = \sqrt{\Sigma_{i=1}^{n}(\hat{y}_i - y_i)^2 / n} \qquad [II.4]$$

where $\hat{y}_i$, y_i and $\bar{y}$ are the forecasted, measured and average of measured electric demand, respectively.

These models can capture the general shape of the electric demand profile but some phenomena remain difficult to predict. SVR and Kernel Regression show a CV(RMSE) of less than 5% over the 24-hour testing set, regardless of the training set size. The CV(RMSE) of dynamic ANN varies from 3.8% (for 21 days training data) to 6.6% (for 7 days data). The lower CV(RMSE) value over the testing set compared to the training set could be explained by their respective size: the testing set is only over 24 hours and the training set size is from 7 to 21 days.

Models 1 and 2 use current and previous values of regressors and electric demand to estimate the future value. The more historical values they have, the better they perform. The range of variation of the electric demand would be better described and so the predictions more accurate. Model 1 shows an improvement of its performance with the increase of the training set size. The issue is in defining the number of previous values that are relevant. The averaged value of the regressors or their peak values over previous hours or days could be relevant to describe the future electric demand.

Figure II.1 displays the evolution of CV(RMSE) of the three inverse models. The training set size is fixed to 14 days and a larger prediction horizon of 7 days is considered. The three models present forecasts with a CV(RMSE) of less than 10% over the following week. A noticeable increase of CV(RMSE) appears after the first day. In terms of RMSE, the SVR and Kernel regression models present values of 15.1-16.4 kW, over the 24-hour testing set, regardless of the training set size. The RMSE of the dynamic ANN varies from 12.5 kW (with 21 days training data) to 21.6 kW (with 7 days). The MP regression model has smaller RMSE values, of about 50% less.

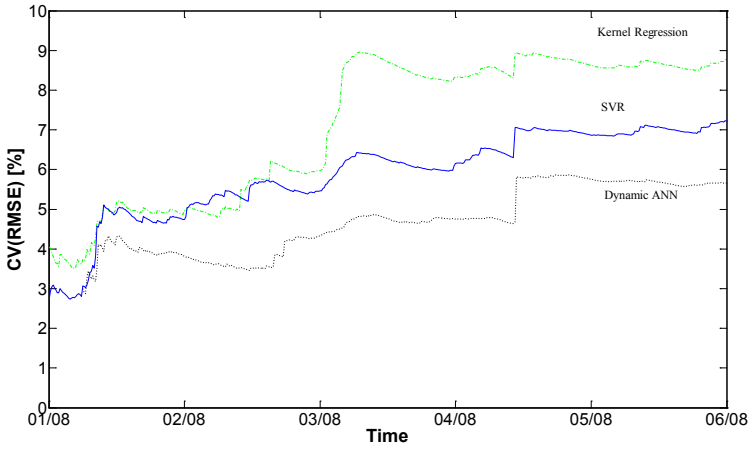

Figure II.1 Evolution of CV over the prediction horizon for the forecasting of the electric demand of the chillers

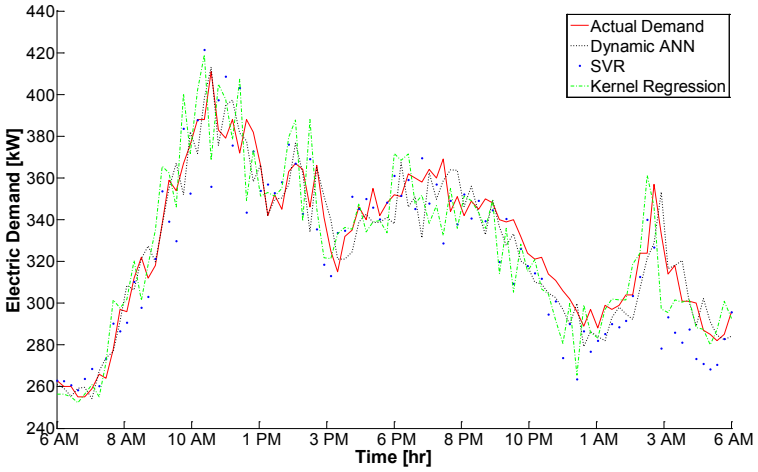

Figure II.2 forecasting of the electric demand of the chillers over the next 24 hours with a 2-week training set

Table II.5 Comparison of inverse models on case study

Model	Training set Size	Training set		Testing set (next 24h)	
		CV (%)	RMSE (kW)	CV (%)	RMSE (kW)
Model 1.	7 days	8.9	26.5	4.9	16.2
Kernel	10 days	8.7	26.6	4.9	16.4
Regression	14 days	10.9	31.3	4.9	16.2
	21 days	14.1	33.9	4.9	16.4
Model 2.	7 days	7.4	22.1	6.6	21.6
Dynamic	10 days	6.4	19.8	4.1	13.5
ANN	14 days	8.9	26.0	5.4	17.9
	21 days	11.6	28.1	3.8	12.5
Model 3.	7 days	6.2	18.4	4.8	15.6
SVR	10 days	6.2	19.0	4.7	15.4
	14 days	7.2	20.5	4.7	15.3
	21 days	9.2	22.1	4.6	15.1

MP	7 days	3.9	12.4	3.7	8.8
Regression[1]	10 days	3.8	11.1	4.0	9.7
	14 days	3.9	10.9	3.8	11.8
	21 days	3.7	10.8	2.9	9.7

In a second approach, the idea is to use the inverse models to perform "forecasts" instead of "predictions": the future values of the regressors are unknown. The regressors are reduced to the time of the day, the day of the week and the forecast of outdoor temperature. The inputs specific to the operation of the chiller (supply chilled and condenser water temperatures) are removed. The three inverse models are trained using a 14-day training set. The optimized parameters are presented in Table II.6 and Table II.7. The predicted and measured profiles of the electric demand over the week following the training period are presented in Figure II.3 for each model. SVR and Kernel Regression models present a CV(RMSE) of about 15% in the first day and converges at about 20% over the complete testing set of seven days. The dynamic ANN presents a bigger CV(RMSE) of 24.9% over the next 24 hours and reaches 22.4% for a one week prediction horizon. The three inverse models tested on this case study showed a good ability in electric demand forecasting. The SVR model presents interesting generalization aptitude. More reflection is required on the selection of regressors describing the electric demand.

Table II.6 Optimized parameters for Model 1 - Kernel Regression

Parameter	Kernel bandwidth	Optimized value
h	**Hour of the day**	**3.33**
d	**Day of the week**	**3.33**
t	**Outdoor air temperature**	**3.33**

Table II.7 Optimized parameters for Model 3 - SVR

Parameter	Description	Optimized value
γ	**Kernel bandwidth**	**6.75**
C	**Penalty parameter of Kernel function**	**50**
ε	**Penalty parameter of loss function**	**20.06**

[1] [86] Monfet D, Zmeureanu R. Ongoing commissioning of water-cooled electric chillers using benchmarking models. Applied Energy. 2012;92:99-108.

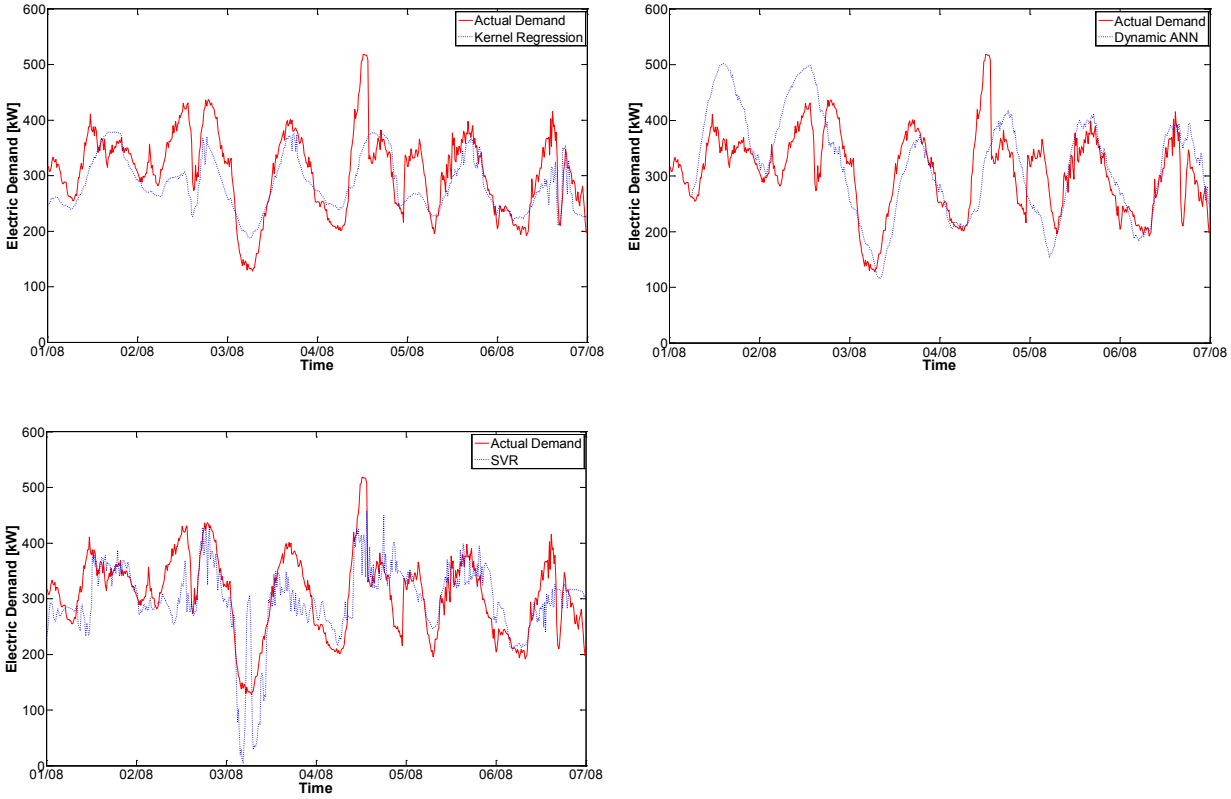

Figure II.3 Forecast of electric demand over the first week of August 2009

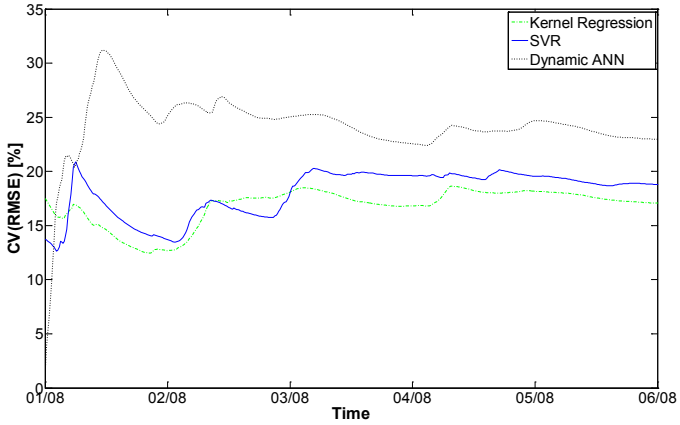

Figure II.4 Evolution of CV over the prediction horizon for the forecast of electric demand of the chillers using the 2nd approach

From the literature review of HVAC-related studies, it has been noticed that in most cases the forecasts were performed of the whole building electric demand [17, 88] in which all energy end-uses are integrated in one single signal, or of a large campus [89], or of a whole utility company [90], or of a province [91, 92]. The forecast of the total electric demand integrates several usages with their own schedules. It is relatively easier to achieve, compared with the forecast of the electric demand of a single HVAC equipment, which is more sensitive to changes in operation. Most published studies used measurements of one-hour or higher time steps, while the BAS contains data recorded at 15-min time step that can be used for the forecasting. Very few studies focused on the forecast of the electric demand at the system level [93]. The forecast of electric demand of existing HVAC sub-systems at sub-hourly time steps, which is of high interest for the DR programs [94], requires additional research effort.

In terms of algorithms used in the previous publications, the neural networks and support vector machines presented the best forecasting performance. The review of the literature showed that the selection of regressors is a challenging issue; the result of selection has an important impact on the forecasting performance of those forecasting methods. The development of a forecasting model following the process of KDD is still in the early stages and therefore requires more investigation in the case of using BAS trend data recorded at 15-min time step for the forecasting applications in building energy systems and equipment.

In the research field of smart grid, a large number of studies was carried out in the residential sector, on the forecast of the aggregation of many small individual loads such as space heating and cooling or water heating [95, 96]. However more investigations are still necessary to develop demand response strategies in commercial buildings [97, 98], and to forecast the electric demand of end-users such as HVAC systems [99].

In this thesis, the term “prediction” refers to the estimation of the output variable at time t+1 knowing the value of the inputs at the same time step t+1. The term “forecasting” corresponds to the estimation of the output value at future time steps (t+1, t+2, …) knowing the input values at the current and past time steps (t, t-1, …). Some studies do not indicate the forecasting time horizon and some others do not specify if they performed a prediction or a forecasting. The development of a forecasting model giving an estimation of the future trend of the electrical demand of a HVAC component is of interest for electric utilities.

II.4 Objectives of the thesis

In the context of Demand Response programs, there is a need from building energy managers for tools to forecast the electric demand of HVAC systems to plan for fast-DR control strategies. This thesis contributes to the DR research field by presenting a method for multi-step forecasting of the electric demand of HVAC sub-system on the short-term**.** The main objectives of this thesis are the following:

1) Development of a method for short-term forecasting of the electric demand of HVAC cooling system at a 15-min time-step with a prediction horizon up to six hours; for this purpose ANN and SVM inverse models are used.
2) Use of data mining techniques to help in the development of forecasting models, for the selection of regressors, to identify typical daily profiles of the target variable and to better understand the operation of HVAC systems.
3) Validation of the proposed methodology on two case studies.

Chapter III. Proposed research method

The scope of this research is the development of a method for the forecasting over the next six hours of the electric demand of the cooling system of the building under normal conditions of operation. It is called the "baseline" electric demand of the building and includes the secondary air system and primary cooling system electric demands. Two methods are proposed for the forecasting of electric demand of cooling systems: a cascade-based (global) method and a component-based method. The cascade-based method includes a sequence of forecasts of target variables from the AHU to the chillers. The component-based method forecasts the electric demand of one component of the HVAC system such as a fan.

The forecasted baseline electric demand is compared to the upper and lower benchmark values developed using previous observations of the electric demand. By comparing the forecast of electric demand with upper and lower benchmarks, the building operator will notice in advance (of up to six hours) if the current conditions will lead to a high electric demand.

The upper (U_{b,t_i}) and lower (L_{b,t_i}) benchmark values are calculated at each 15-min time step of the day (t_i) using the average value at that time step ($\mu_{\dot{E},t_i}$) plus or minus the standard deviation ($\sigma_{\dot{E},t_i}$) of the electric demand ($\dot{E}$), which are calculated at that time step from past measurements (Equation [III.1]). Another option is to use a constant value for the upper and lower benchmarks using the daily average and standard deviation of the electric demand instead of each time step.

$$U_{b,t_i} = \mu_{\dot{E},t_i} + \sigma_{\dot{E},t_i}$$
$$L_{b,t_i} = \mu_{\dot{E},t_i} - \sigma_{\dot{E},t_i} \qquad [III.1]$$

An example of the forecasted "baseline" electric demand of chillers is displayed in blue on Figure III.1 from 12 PM to 6 PM and the upper and lower benchmarks in green. The measured electric demand is shown in red; its values after 12 PM are presented for purpose of comparison with the forecasted profile.

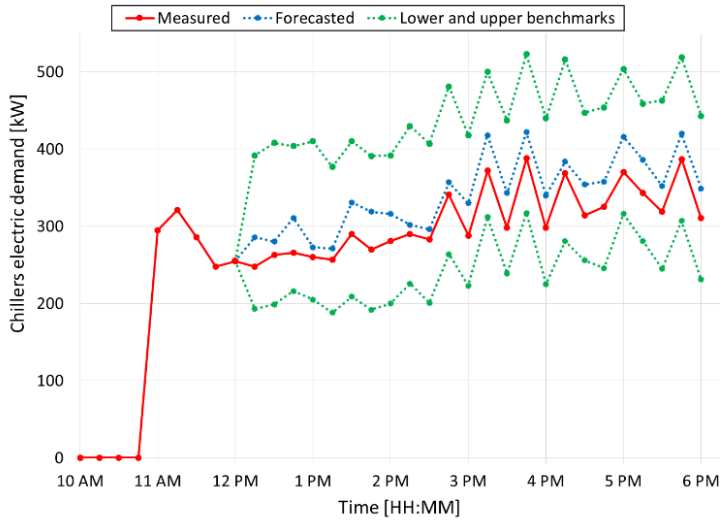

Figure III.1 Example of forecast of electric demand as a baseline vs benchmarks and measured electric demand

The forecasted electric demand could also be compared with the limit for the electric demand imposed by the utility company for the case of demand response strategy. This limit allows the building operator to develop strategies of operation with the scope of limiting or shifting the peak electric demand during specific time periods. For instance, the operator can assess the impact of a change in the thermostat setpoints (e.g. indoor air temperature or supply air temperature of the AHU) on the future electric demand. For this purpose, a calibrated model of the building can be used to generate synthetic data representative of the actual operation of the building. These synthetic data can be used to develop correlations between the cooling coil load or supply air flow

rate and the room air temperature. Another way is the use of correlations extracted from case study building, based on real measurements, between the cooling coil load or the supply air flow rate and the room air temperature [100].

The two forecasting methods are presented in sections III.1 and III.2 for the forecasting of electric demand of cooling systems: the cascade-based method and the component-based method. The cascade-based forecasting method is applied on the case-study building of the Genomic Research Center located on Loyola campus of Concordia University. The component-based method is applied on the case-study of an office building in Shawinigan-Sud, QC.

III.1 Cascade-based forecasting method

The cascade-based method forecasts the target variables of interest of different HVAC equipment and the electric demand of primary and secondary cooling systems over the next six hours. It starts with the forecasting of the air flow rate supplied by the AHUs to the thermal zones, and ends with the forecasting of the electric demand of the cooling system (Figure III.2). The following target variables are successively forecasted: the supply air flow rate ($\dot{V}_{SA}$), the cooling coil load ($\dot{Q}_{CCL}$), whole building cooling load ($\dot{Q}$), and electric demand of the primary and secondary cooling systems ($\dot{E}_{cooling}$). These successive forecasts of the target variables enable the investigation of the impact of each target variable on the primary electric demand. The future value of each target variable, for instance at time t + 15 min, is forecasted based on its previous values and of other regressors. The 15-min time-interval corresponds to the sampling time step of electric meters used in this case study, however, the proposed method could use any sampling time of the BAS trend data. Examples of regressors used for each target variable are given on the flowchart (Figure III.2).

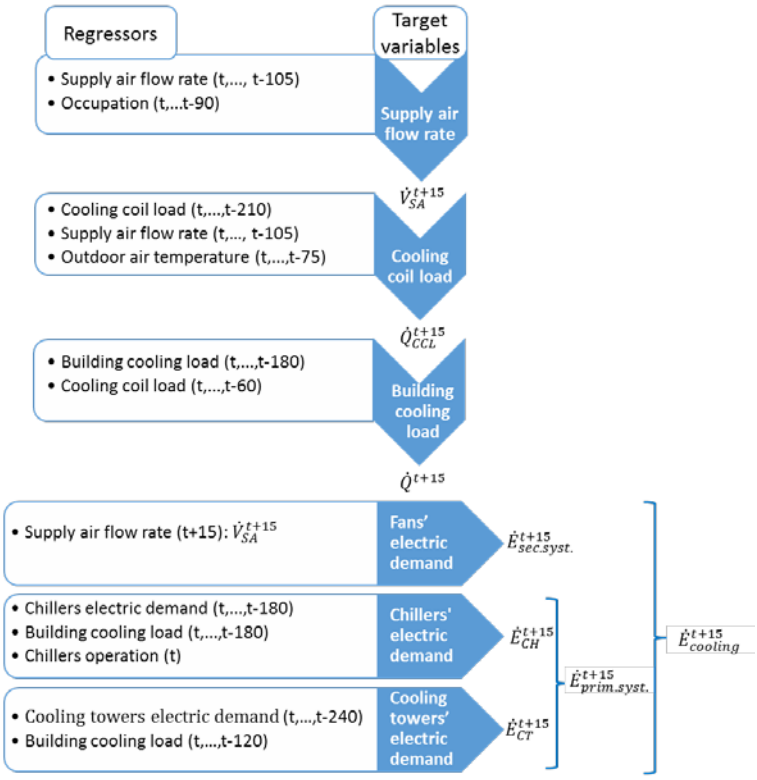

Figure III.2 Cascade-based forecasting of the electric demand of cooling system

For each one of the six target variables in Figure III.2, some investigations are performed using data mining techniques as a preprocessing step. The preprocessing steps intend to get a first knowledge of the target variable to forecast, help in the selection of the variables to be used as regressors and identify typical daily profiles, if any. The regressors are selected from a database, extracted from the BAS trend data. The database includes measured and derived variables that characterize the operation of HVAC system in key points (e.g., air temperature and relative humidity, chilled water temperature), the operation modes of equipment, and environmental conditions. The measured and derived variables of database are listed in the nomenclature. For each target variable, the following preprocessing steps are undertaken:

1. Exploratory analysis of the target variable.
2. Selection of the regressors and the relevant past values by using :
   - Cross-correlation analysis of the target variables to the available variables, and
   - Filtering method that ranks these variables according to their cross-correlation coefficient to the target variable; and selection of those variables with a coefficient above a specified threshold (greater than 0.7 in this study).
3. Identification of the typical daily profiles of each target variable.

A forecasting model is then developed for each target variable by using the selected regressors. More details about the pre-processing steps are given in the section VI.1 of Chapter VI.

III.1.1 Electric demand of the cooling system of the Genome building

The proposed cascade-based forecasting method can be applied to the electric demand of the HVAC cooling system of any CI building with available BAS trend data, from which the target variables and regressors can be extracted. The following details of the method are explained for the case study of research center, called Genome (GE) building, on the Loyola Campus of Concordia University in Montreal, Quebec, Canada.

The forecast of electric demand of the cooling system of the GE building at next time step (t+15 min) ($\dot{E}^{t+15}_{cooling,GE}$) includes two terms (Equation [III.2]). The first term corresponds to the electric demand of four fans of 29.8 kW each of the AHUs ($\dot{E}^{t+15}_{sec.syst.}$). The second term refers to the electric demand of the primary cooling system in the central plant ($\dot{E}^{t+15}_{prim.syst.}$) that corresponds to the cooling requirements of the GE building.

$$\dot{E}^{t+15}_{cooling,GE} = \dot{E}^{t+15}_{sec.syst.} + \dot{E}^{t+15}_{prim.syst.} \quad [III.2]$$

The electric demand of the Variable Air Volume (VAV) fans in the AHUs ($\dot{E}^{t+15}_{sec.syst.}$) is estimated using Equation [III.3], where Y is the ratio of the electric demand at part-load to the maximum measured electric demand $\dot{E}_{max}$ over the period considered (equation [III.4]):

$$\dot{E}^{t+15}_{sec.syst.} = \dot{E}_{max} * Y \quad [III.3]$$

$$Y = a_0 + a_1.X + a_2.X^2 + a_3.X^3 \quad [III.4]$$

Where X is the part-load ratio of the supply air flow rate: $X = \dot{V}_{measured}/\dot{V}_{max}$.

The electric demand ($\dot{E}^{t+15}_{prim.syst.}$) is calculated by multiplying the electric demand of the cooling system of the central plant $\dot{E}^{t+15}_{cooling,RF}$ (equation [III.5]) with the ratio (α^{t+15}) of the cooling load of the GE building ($\dot{Q}_{GE}$) to the total cooling load of the central plant ($\dot{Q}_{RF}$) (equation [III.6]).

$$\dot{E}^{t+15}_{prim.syst.} = \alpha^{t+15} * \dot{E}^{t+15}_{cooling,RF} \quad [III.5]$$

$$\alpha^{t+15} = \dot{Q}^{t+15}_{GE}/\dot{Q}^{t+15}_{RF} \quad [III.6]$$

The electric demand ($\dot{E}_{cooling,RF}^{t+15}$) includes two chillers ($\dot{E}_{CH}^{t+15}$) with 3165 kW cooling capacity and electric demand of 549 kW each, along with two fans in the cooling towers ($\dot{E}_{CT}^{t+15}$) with a motor of 29.8 kW each (equation [III.7]).

$$\dot{E}_{cooling,RF}^{t+15} = \dot{E}_{CH}^{t+15} + \dot{E}_{CT}^{t+15} \qquad [III.7]$$

III.1.2 Multi-step-ahead forecasting: iterated and direct strategies

The problem of multi-step-ahead forecasting is commonly solved using two types of strategies: an iterated strategy or a direct strategy [101]. The iterated strategy uses a repetition of one-step-ahead forecast with the same forecasting model. For instance, when the current time is t, the model forecasts the target variable at time t+15 min using the current and previous observations of the target variable at times t, t-15, t-30 etc. The 2nd forecast, at time t+30 min, is performed using the previous observations at times t, t-15, t-30 etc. and the first forecast of the target variable at time t+15 min, and so on. In the example given in Equations [III.8], the model f forecasts the value of the target variable y by iteration over a horizon of prediction of one hour with a 15 min time interval. The window width of regressors of the test set is kept constant with $n + 1$ observations, from t to t-n for $\hat{y}^{t+15}$, and t+15 to t-n-15 for $\hat{y}^{t+30}$ etc.

$$\begin{aligned}
\hat{y}^{t+15} &= f(y^t, y^{t-15}, \dots, y^{t-n}) \\
\hat{y}^{t+30} &= f(\hat{y}^{t+15}, y^t, y^{t-15}, \dots, y^{t-n-15}) \\
\hat{y}^{t+45} &= f(\hat{y}^{t+30}, \hat{y}^{t+15}, y^t, y^{t-15}, \dots, y^{t-n-30}) \\
\hat{y}^{t+60} &= f(\hat{y}^{t+45}, \hat{y}^{t+30}, \hat{y}^{t+15}, y^t, \dots, y^{t-n-45})
\end{aligned} \qquad [III.8]$$

Where, y is the measured value of the target variable and $\hat{y}$ is the forecasted value.

In the direct strategy for multi-step-ahead forecasting, one forecasting model is developed and trained for each time step of the prediction horizon. For instance, four different forecasting models ($f_1, \dots, f_4$) must be developed for the forecast of the variable y at times t+15 to t+60 min. For this purpose we use the same set of regressors and training dataset, from previous ($t - n$) to current (t) observations (Equations [III.9]). Other regressors (x) can also be added. The direct strategy is time consuming since a different forecasting model is developed and trained for each time step of the prediction horizon.

$$\begin{aligned}
\hat{y}^{t+15} &= f_1(y^t, y^{t-15}, \dots, y^{t-n}, x^t, \dots, x^{t-n}) \\
\hat{y}^{t+30} &= f_2(y^t, y^{t-15}, \dots, y^{t-n}, x^t, \dots, x^{t-n}) \\
\hat{y}^{t+45} &= f_3(y^t, y^{t-15}, \dots, y^{t-n}, x^t, \dots, x^{t-n}) \\
\hat{y}^{t+60} &= f_4(y^t, y^{t-15}, \dots, y^{t-n}, x^t, \dots, x^{t-n})
\end{aligned} \qquad [III.9]$$

In this study, the iterated approach is applied. Other regressors are added to the current and past values of the target variable, and for such a regressor a different forecasting model is developed. For instance, in Equations [III.10], the target variable (y) is forecasted by the model (f_1) based on its current and previous values from time t to $t-n$, plus two other regressors (x_1, x_2) from time t to $t-m$, and respectively t to $t-l$. Forecasting the value of the target variable at the following time step ($\hat{y}^{t+30}$) requires the estimates of the target variable and regressors at the previous time step ($\hat{y}^{t+15}, \hat{x}_1{}^{t+15}, \hat{x}_2{}^{t+15}$). In the example of Equations [III.10], two autoregressive forecasting models (f_2, f_3) are developed to estimate the future value of each regressor ($\hat{x}_1^{t+15}$ and $\hat{x}_2^{t+15}$).

$$\begin{aligned}
\hat{y}^{t+15} &= f_1\begin{pmatrix} y^t, \dots, y^{t-n}, \\ x_1^t, \dots, x_1^{t-m}, x_2^t, \dots, x_2^{t-l} \end{pmatrix} \\
\hat{x}_1{}^{t+15} &= f_2(x_1^t, x_1^{t-15}, \dots, x_1^{t-m}) \\
\hat{x}_2{}^{t+15} &= f_3(x_2^t, x_2^{t-15}, \dots, x_2^{t-l}) \\
\hat{y}^{t+30} &= f_1\begin{pmatrix} \hat{y}^{t+15}, y^t, \dots, y^{t-n-15}, \\ \hat{x}_1{}^{t+15}, x_1^t, \dots, x_1^{t-m-15}, \\ \hat{x}_2{}^{t+15}, x_2^t, \dots, x_2^{t-l-15} \end{pmatrix} \\
\hat{x}_1{}^{t+30} &= f_2(\hat{x}_1{}^{t+15}, x_1^t, x_1^{t-15}, \dots, x_1^{t-m-15}) \\
\hat{x}_2{}^{t+30} &= f_3(\hat{x}_2{}^{t+15}, x_2^t, x_2^{t-15}, \dots, x_2^{t-l-15})
\end{aligned} \qquad [III.10]$$

The inputs of the autoregressive model are selected by autocorrelation analysis. The past values of the regressor that are moderately to strongly correlated to the value at time t are retained, with an autocorrelation coefficient greater or equal to 0.7 [102].

III.2 Component-based forecasting method

The component-based method uses the same sequence of pre-processing steps as in the cascade-based method including: (1) exploratory analysis, (2) identification of the typical daily profiles and (3) selection of the regressors in each cluster (for each subset of typical daily profiles).

The main difference with the cascade-based method is that the component-based method focuses on one component of the HVAC system such as a fan, a chilled water pump or a chiller. The pre-processing steps are applied to the variable modulating the HVAC component. This control variable is extracted from the BAS and recorded. The electric demand of the HVAC component is forecasted based on the current and past values of the control variable and the maximum electric demand of the equipment. The selection of the regressors is restrained to the selection of the relevant past values of the control variable to its future value; no other regressor is involved. For the component-based approach, the iterated strategy is employed for multi-step-ahead forecasting of the control variable (Equation [III.8]).

For instance, if the HVAC component is the supply fan, the preprocessing steps are performed on the modulation of the supply fan, which is the control variable of the fan.

In the component-based method, a hybrid physical/non-physical model is developed for the forecasting of the electric demand of the HVAC component. The pre-processing steps are performed: the original dataset is partitioned into subsets of similar daily profiles of the control variable and the relevant previous values of the control variable are selected for each subset. An autoregressive neural network is used for time series forecasting of the control variable; i.e. only based on its current and previous values. The artificial neural network is trained over two weeks of measurements. The weights associated with each neuron of the network are tuned such that the forecasted value of the control variable fits well the recorded values over the training set. Once the model is trained, the value of the control variable is forecasted over the test set. The iterated strategy enables to forecast the future value of the control variable over the next six hours using the same model.

The forecasted value of the control variable is then passed as an input to a physical model that estimates the future value of the electric demand by using also the maximum electric demand of the fan.

In chapter VI, the component-based approach is applied to forecast the electric demand of the supply fan of an office building located in Shawinigan-Sud.

Chapter IV. Cascade-based forecasting method Case study: pre-processing steps

This chapter presents the pre-processing steps in the development of forecasting models including: exploratory analysis, selection of the regressors and identification of typical daily profiles by clustering analysis. Three different sequences of pre-processing steps are used in the study, and the forecasting results are compared. The exploratory data analysis is the first step that is common to all sequences. The same filtering technique is applied in all three sequences for the selection of regressors.

1) In the sequence A, the regressors of each target variable are selected by using the whole data set available, followed by identification of different clusters (i.e., the typical daily profiles). For each target variable, a forecasting model is developed for each cluster by using the regressors from the entire data set.

2) In the sequence B, the clusters are first identified, followed by the selection of regressors of each target variable and for each cluster. The target variable is forecasted for each cluster by using regressors that are specific to that cluster.

3) In the sequence C, the regressors are selected for each target variable by using the whole data set available (the same regressors as in sequence A). For each target variable, only one forecasting model is developed using the regressors from the entire data set, since there are no clusters.

The pre-processing steps are used for the development of forecasting models for the target variables: from the supply air flow rate of AHUs up to the electric demand of the secondary and primary cooling system. Sections IV.2 to IV.5 present the pre-processing steps of the target variables by following the sequence A. The selected regressors for each target variable and for each sequence of pre-processing steps are presented in section IV.6. And the comparison of forecasting results from the three sequences A, B and C is presented in Chapter V.

IV.1 Building case study: the Genomic Research Center

The proposed forecasting method is applied to the GE building, which has a total floor area of 5,400 m^2, over five levels, including a basement (Figure IV.1). The building is oriented 60° NW and has a window-to-wall ratio of 33%. The building contains 48 offices, three conference rooms and corridors, which account for about 53% of the total floor area, and laboratories with fume hoods that account for about 30%. The remaining space is occupied by kitchen/lounge and restrooms on each floor.

More details about the envelope of the building and the HVAC system are given in [103, 104].

Figure IV.1 Genomic Research Center.

The cooling load required by the Genome building is provided by the central plant (Figure IV.2). The primary cooling equipment includes two large chillers with a cooling capacity of 3,165 kW (900 tons) each and a coefficient of performance of 5.76 at design conditions. The chillers are cooled down by two cooling towers with a designed capacity of 4,750 kW (1350 tons) each. The operation of the two chillers is alternated over summer. They are both operated in the same time only when one cannot fulfill the demand. The two centrifugal chillers supply chilled water at a set-point of 6.7 °C via two constant speed pumps. The two pumps are connected in parallel and circulate the chilled water to the AHUs and fan coils of several buildings on the campus. Each pump is controlled in association to the chillers; if one chiller is working, only one pump is operated and both are operated when both chillers are working. Two other constant speed pumps are used similarly in the condenser water loop to reject heat to the environment. These two pumps

are operated in association with the cooling towers. Additional information about the primary cooling equipment and its operation are presented in [105].

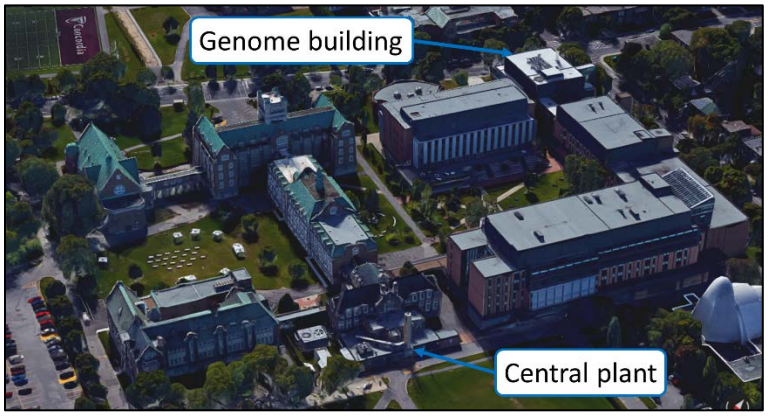

Figure IV.2 Loyola campus, Concordia University, Montreal, Quebec. Image from Google Earth®

IV.1.1 Database of measured and derived variables

A database is created including the measured variables available through the Building Automation System (BAS) as well as derived variables. The database includes about 170 variables, at a 15-min time interval, over nine years. 36 variables are selected from the database based on the domain knowledge (Table IV.1). The measured variables describe the temperature of the chilled water entering and leaving the Genome building as well as its flow rate; the air conditions in the Air Handling Units (AHUs) 1 and 2 (in Table IV.1 only AHU#1 is presented), as well as the environmental conditions on site.

The derived variables include for instance the enthalpies calculated from the measured temperature and relative humidity and the total supply air flow rate in both AHUs. They also include the cooling coil load on the air side and the cooling load from the building on the water side. Some indices describing the time of the day are also developed. The hour of the day (HH:MM) and day of the week (Mon, …, Sun) are encoded using periodical functions to reflect the cyclic character of these variables (Equations [IV.1] and [IV.2]).

$$HOUR = \cos(2\pi . t/T_{HOUR}) \quad [IV.1]$$

$$DAY = \cos(2\pi . t/T_{DAY}) \quad [IV.2]$$

where, t is the incremented numerical value recorded by the BAS and reflecting the time. $T_{HOUR} = 24\ hours$ and $T_{DAY} = 7\ days$.

A Boolean variable is encoded to differentiate the working days from the weekends and holidays. Another Boolean variable is extracted from the operation of the fans in the AHU (Chapter 6, Section 1), differentiating the occupied periods from unoccupied. The last derived variable in Table IV.1 indicates how many chillers are operated.

Table IV.1 Available measured and derived variables

Points	Symbol	Description	Measured or Derived (M/D)	Units
Variables of interest				
GE CHW_FLOW_SUP	$\dot{V}_{CHW,GE}$	Chilled water flow rate supplied to the building	M	L/s
GE CHW_T_SUP	T_{CHWS}	Supply chilled water temperature	M	°C
GE CHW_T_RET	T_{CHWR}	Return chilled water temperature	M	°C
AHU1&2 T_SUP	T_{SA}	Average supply air temperature of both AHUs	M	°C
AHU1&2 FLOW_SUP	$\dot{V}_{SA}$	Supply air flow rate of both AHUs	D	L/s
AHU1&2 CC_LOAD	$\dot{Q}_{CCL}$	Cooling coil load on the air side of both AHUs	D	kW
GE CG_LOAD	$\dot{Q}_{GE}$	Building cooling load	D	kW
AHU #1				
AHU1 FLOW_SUP	$\dot{V}_{SA,1}$	Supply air flow rate	M	L/s
AHU1 FLOW_RET	$\dot{V}_{RA,1}$	Return air flow rate	M	L/s
AHU1 MOD_MIX_DAMP	$mod_{D,1}$	Mixed air damper modulation	M	%
AHU1 H_SUP	$RH_{SA,1}$	Relative humidity of supply air	M	%
AHU1 H_RET	$RH_{RA,1}$	Relative humidity of return air	M	%
AHU1 T_MIX	$T_{MA,1}$	Mixed air temperature	M	°C
AHU1 T_RET	$T_{RA,1}$	Return air temperature	M	°C
AHU1 MOD_CG_V	$mod_{CCV,1}$	Cooling coil valve modulation	M	%
AHU1 E_SUP	$h_{SA,1}$	Enthalpy of supply air	D	kJ/kg
AHU1 E_RET	$h_{RA,1}$	Enthalpy of return air	D	kJ/kg
Outdoor air				
OUT_H	RH_{OA}	Outside air relative humidity	M	%
OUT_T	T_{OA}	Outside air temperature	M	°C
OUT_E	h_{OA}	Outside air enthalpy	D	kJ/kg
Time				
HOUR	H	Hour of the day	D	-1,…,1
DAY	d	Day of the week	D	-1,…,1
DAYTYPE	DT	Working day or weekend	D	0,1
OCCUPATION	Occ	Occupied and unoccupied periods	D	0,1
Other indicators				
CG_NET S_S_CH	CH_{op}	Operation of the two chillers	D	0, 1, 2

IV.1.2 Dataset cleaning: Missing and faulty values

The variables measured by the sensors of the BAS present some missing or faulty values due to uncalibrated sensors or unexpected events in the system operation. The dataset cleaning is an important step in the process of Knowledge Discovery in Databases (KDD) [8, 106]. Having a

reliable dataset is a requirement considering that the forecasting models are developed based on the measurements.

The missing values of continuous variables, such as the air flow rate or damper position, over a time period shorter than two hours are interpolated. The missing values are ignored when the time period is longer than two hours and all the other observations at the same time step are removed.

Inconsistencies were detected in the measurements of the outdoor air temperature and humidity by comparing them to other sensors located in other buildings on the same campus. The outdoor air temperature and humidity were corrected using on-site measurements.

IV.2 Pre-processing of the supply air flow rate

IV.2.1 Exploratory analysis of the supply air flow rate

The daily operation of the supply fans in both AHUs of the Genome building is recorded through the BAS, at a 15 min time interval, over summer 2014, from the 1st of June to the 31st of August. The daily operation is presented on the carpet plot (Figure IV.3). The x-axis refers to the hour of the day and on the y-axis is the date over the selected time period. The colour of each cell corresponds to the supply air flow rate as percentage of the fans capacity in both AHUs. It varies from 0% in dark blue to 100% in dark red (42,472 L/s for both AHUs). The white cells from 10th to 12th of June correspond to missing data.

The fans are operated at a higher level (50 to 60% of the capacity) on working days from 8:00 to about 18:00 (vertical red dashed line). As a first estimation, the starting and stopping time are approximately determined based on the colour of cells. The starting time clearly appears, while the stopping time is more variable. The fans work at a lower level at night and on weekends (around 40% of the capacity).

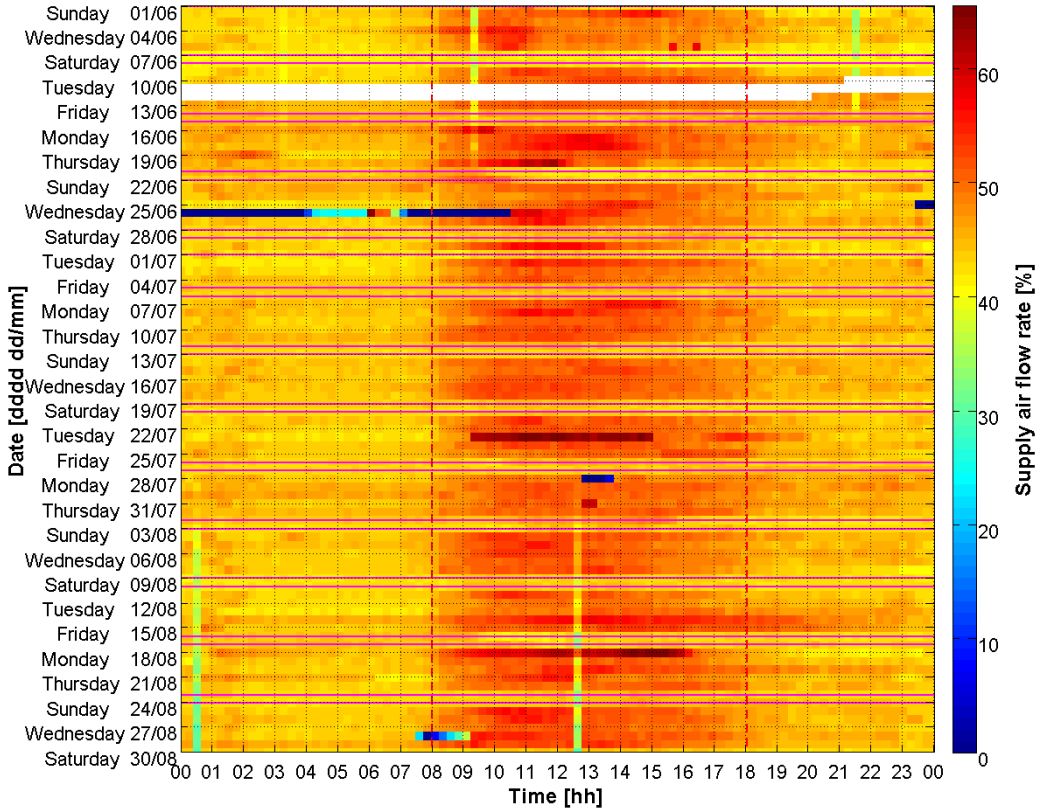

Figure IV.3 Schedule of operation of the supply fans in both AHUs in summer 2014

The probability distribution function of supplied air flow rate from both AHUs over the summer season is presented in Figure IV.4 as normal distributions. Two modes of operation can be noticed, in black for occupied period and white for unoccupied. The median supply air flow rate is about 18,500 L/s (44% of the design capacity) during unoccupied periods. The median increases to reach 21,300 L/s (50% of the design capacity) during the occupied periods.

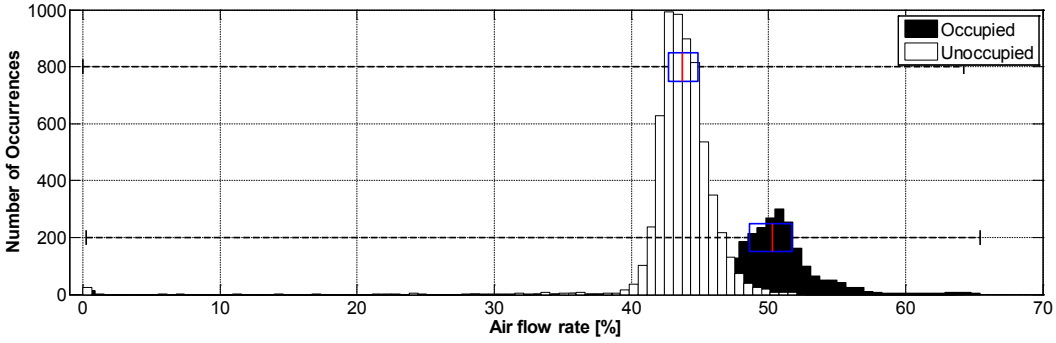

Figure IV.4 Probability distribution of supply air flow rate corresponding to occupied and unoccupied periods

IV.2.2 Selection of the regressors for the forecasting of the supply air flow rate

The selection of regressors is an important step in the development of the model that impacts significantly on its forecasting performance. The issue of selecting a subset of variables is presented and discussed in details in [107].

In this study, a cross-correlation analysis is performed to identify the potential regressors and the past values of the regressors, to forecast the target variable.

IV.2.2.1 Identification of the relevant past values by cross-correlation analysis

The value of the target variable at the next time step, t+15 min, is forecasted by using a number of relevant past values (known values at previous times steps t, t-15, t-30 min etc.), which are selected as regressors. A filtering technique is then used based on the calculated cross-correlation coefficients of the target variable with the independent variables, to rank and select the variables whose coefficients are above a defined threshold.

The cross-correlation coefficient is calculated using equation [IV.3], where $\boldsymbol{y^{t+15}}$ is the vector of the target variable including j observations over summer 2014. $\boldsymbol{y^{t+15}}$ contains the previous values of $\boldsymbol{y}$ until and including the observation at time t and t+15 min. $\boldsymbol{x^{t+15}}$ is a vector containing the previous observations of one of the potential regressors until and including the observation at time t and t+15 min. $\overline{y^{t+15}}$ and $\overline{x^{t+15}}$ are respectively the average values of the vectors $\boldsymbol{y^{t+15}}$ and $\boldsymbol{x^{t+15}}$ over the *j* measurements.

$$R(i) = \frac{\sum_j (y_j^{t+15} - \overline{y^{t+15}})(x_j^{t+15} - \overline{x^{t+15}})}{\sqrt{\sum_j \left(y_j^{t+15} - \overline{y^{t+15}}\right)^2 \sum_j \left(x_j^{t+15} - \overline{x^{t+15}}\right)^2}} \qquad [IV.3]$$

The cross-correlation coefficient is calculated between the target vector $\boldsymbol{y^{t+15}}$ and n vectors of potential regressors $[\boldsymbol{x_1^{t+15-i}}, \dots, \boldsymbol{x_n^{t+15-i}}]$ with i varying from 15 to 300 min (equivalent to five hours). A time delay in the potential regressors greater than 5 hours is not considered to limit the number of regressors.

Figure IV.5 presents those variables that might have an impact on the forecasting of the supply air flow rate at time t+15, out of all available variables (see the nomenclature); those variables are displayed on the x-axis. A value of the coefficient greater than or equal to 0.7 indicates a moderate

to strong correlation [102], presented by colors from orange to red with black diamonds, such as the occupation from time t to t-90 min. The figure gives an indication on the possible time lag between the target variable and the relevant regressors. The y-axis shows the value of the time lag in minutes. For instance, the supply air flow rate from both AHUs (AHU1&2 FLOW_SUP) at time t+15 min is correlated to its previous values from time t to t-105 min (black diamonds on the left of Figure IV.5).

The variable labelled "occupation" is a variable derived from the operation of the fans. It takes a value of 1 during occupied periods (08:00 to 18:00 on weekdays) and 0 during unoccupied periods.

The correlation between the total supply air flow rate and the return air flow rate in both AHUs is due to the fact that the supply and return fans have a similar schedule of operation. The return air flow rate in both AHUs is not being kept as a regressor. The supply air flow rate in each AHU (AHU1 FLOW_SUP and AHU2 FLOW_SUP) are already included in the derived variable which is the summation of both (AHU1&2 FLOW_SUP).

IV.2.2.2 Filtering of the relevant variables to the supply air flow rate

The techniques of selecting the regressors can be classified into three groups: the filter, wrappers and embedded techniques [107]. The filter methods rank the variables according to a unique metric, such as the Pearson linear correlation coefficient, which grades their relation to the target variable. The wrapper methods rate the variables depending on their usefulness to the forecasting algorithm. A variable is considered useful to the prediction algorithm when its use as a regressor reduces the prediction error. In this case, all the combinations of variables as regressors are tested. The embedded methods are specific to a forecasting model; the algorithm selecting the regressors is embedded in the prediction algorithm. The regressors are selected by minimizing a prediction error; it is embedded in the learning process.

In this study, a filtering approach is used to select the potential regressors among all the available variables gathered through the BAS, and over the entire data set. All the variables presenting a moderate to strong cross-correlation coefficient with the target variable, greater than 0.7, are retained. These variables are highlighted in Figure IV.5 by a black diamond.

The regressors retained to forecast the supply air flow rate are ranked in descending order of the cross-correlation coefficient and listed in Table IV.2.

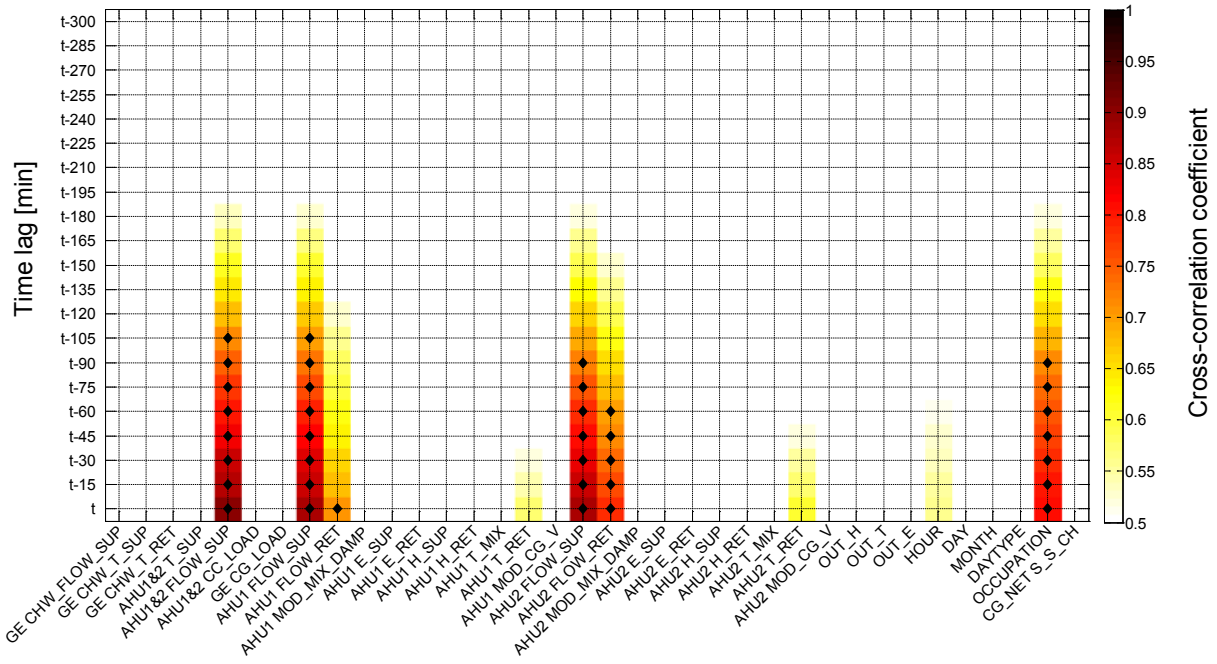

Figure IV.5 Relevant variables for the forecasting at time t+15 of the supply air flow rate and their time lag

Table IV.2 Ranked regressors retained for the forecasting at time t+15 of the supply air flow rate and the corresponding cross-correlation coefficient

Rank	Regressors	Description	Time lag (min)	Correlation coefficient
1	AHU1&2 FLOW_SUP	Supply air flow rate	t	0.899
2	AHU1&2 FLOW_SUP	Supply air flow rate	t-15	0.874
3	AHU1&2 FLOW_SUP	Supply air flow rate	t-30	0.854
4	AHU1&2 FLOW_SUP	Supply air flow rate	t-45	0.828
5	OCCUPATION	Occupied and unoccupied periods	t	0.807
6	AHU1&2 FLOW_SUP	Supply air flow rate	t-60	0.804
7	OCCUPATION	Occupied and unoccupied periods	t-15	0.799
8	OCCUPATION	Occupied and unoccupied periods	t-30	0.788
9	AHU1&2 FLOW_SUP	Supply air flow rate	t-75	0.774
10	OCCUPATION	Occupied and unoccupied periods	t-45	0.772
11	OCCUPATION	Occupied and unoccupied periods	t-60	0.753
12	AHU1&2 FLOW_SUP	Supply air flow rate	t-90	0.743
13	OCCUPATION	Occupied and unoccupied periods	t-75	0.737
14	AHU1&2 FLOW_SUP	Supply air flow rate	t-105	0.712
15	OCCUPATION	Occupied and unoccupied periods	t-90	0.711

As a result, the supply air flow rate of both AHUs can be forecasted at time t+15 using its current and past values as well as the occupation (Equation [IV.4]):

$$\text{AHU1\&2 } FLOW\_SUP(t+15) = f\left(\text{AHU1\&2 } FLOW\_SUP(t,\dots,t-105), \text{OCCUPATION}(t,\dots,t-90)\right) \quad [IV.4]$$

IV.2.3 Identification by clustering analysis of typical daily profiles of supply air flow rate

The process of data mining can be used for "prediction", i.e., to estimate the value of a variable, and for "description", i.e., to extract patterns that can be presented and understood by a user [8]. In this study, a fuzzy C-means clustering algorithm is used, for the "description" purpose, to analyse the daily profiles of the target variable and identify similar profiles based on a distance metric. The daily profiles of the supply air flow rate are the individuals of interest for which the clustering analysis looks for similarity. There are 92 potential daily profiles (from the 1st of June to the 31st of August), each profile being composed of 96 measurements at 15-min time-step.

The mean and standard deviation of the supply air flow rate over portions of the day are used as features to characterize those profiles (Table IV.3). The specific portions of the day have been determined based on a detailed analysis of the fans operation (Figure IV.3). Three days with missing data, from 10th to 12th of June, are removed from the dataset.

Table IV.3 Features characterizing the supply air flow rate daily profiles

Features	Description
$\boldsymbol{\mu_1, \sigma_1}$	Mean and standard deviation over the time period from 00:00 to 07:00
$\boldsymbol{\mu_2, \sigma_2}$	Mean and standard deviation over the time period from 07:00 to 12:00
$\boldsymbol{\mu_3, \sigma_3}$	Mean and standard deviation over the time period from 12:00 to 24:00

The clustering analysis is then performed on the 89 remaining profiles. Two parameters have to be set: the number of clusters and a membership threshold. The clustering algorithm calculates the coordinates of the center of each cluster that correspond to the six features (Table IV.3), which are used to minimize the distance between the clusters centers and the individuals. When a new individual is considered for the assignment to a cluster, the similarity for that individual is determined by the Euclidean distance from the individual to the center of each cluster. A high similarity means that the individual is close to the center of the cluster. An individual is included into the cluster for which the calculated similarity is the highest. The membership threshold determines the minimum similarity required by the individual to belong to one of the clusters. If the similarity of an individual is less than the membership threshold, the individual is not assigned to any cluster; this is an atypical individual. For instance, if an individual has a similarity of 55% with cluster #1 and 65% with cluster #2, and the membership threshold is 70%, the individual is said atypical. If membership threshold is 60%, the individual belongs to the cluster #2.

The membership threshold is set to remove the atypical profile from each cluster such as a daily profile with an abnormal peak. The threshold should be high enough to extract the abnormal profiles but not too close to 100% to avoid having too many profiles removed; a trial-error process is performed. The results of the clustering analysis performed with a membership threshold of 75% and two clusters are presented on Figure IV.6. Cluster #1 contains 55 profiles, while cluster #2 contains 27 profiles. Seven atypical profiles are discarded from the dataset. The values of the clusters' centers returned by the clustering algorithm are presented in Table IV.4.

Table IV.4 Values of the features corresponding to the center of each cluster of the supply air flow rate

	Features	Value [%]	Time period
Cluster #1	$\mu_1 \pm \sigma_1$	43.8 ± 1.3	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	48.9 ± 3.6	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	47.6 ± 3.4	12:00 to 24:00
Cluster #2	$\mu_1 \pm \sigma_1$	42.5 ± 1.1	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	43.4 ± 1.5	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	43.3 ± 1.4	12:00 to 24:00

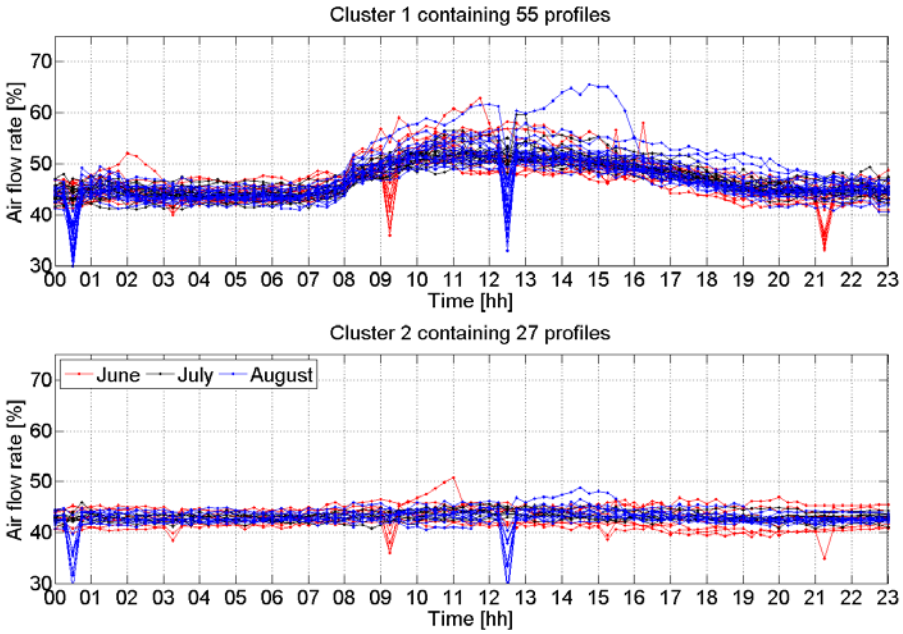

Figure IV.6 Two typical daily profiles corresponding to working days and weekends

The working days are gathered in cluster #1, and weekends and holidays in #2 (Figure IV.7). The Tuesday appearing in the second cluster is Canada Day, on the 1st of July.

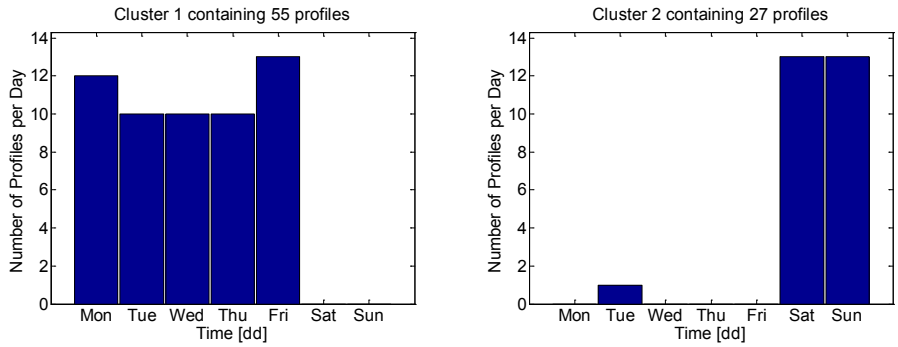

Figure IV.7 Day-type analysis of the two clusters

The schedule of operation of the fans that was first approximately determined from the exploratory analysis (Figure IV.3) is now readjusted. During the occupied period, the fans are operated from 07:00 to 20:00 on working days (Cluster #1 of Figure IV.6) instead of 08:00 to 18:00.

Figure IV.8 displays the atypical profiles that differ from the profiles in cluster #1 and 2. These profiles are atypical because of a system shutdown resulting in no air flow rate supplied (e.g. 26th of June in green) or because of a supplied air flow rate greater than usual (e.g. 23rd of July in yellow).

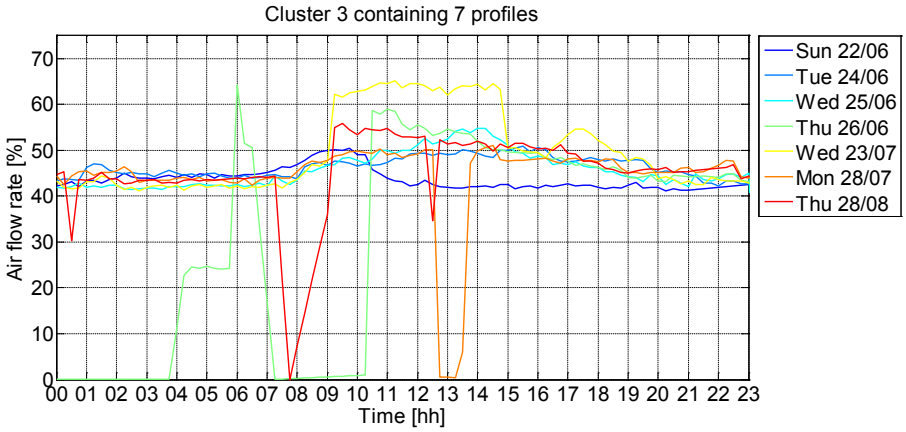

Figure IV.8 Cluster containing the non-classified profiles: atypical profiles

Two forecasting SVR models should now be developed for the supply air flow rate, one for the cluster #1 and another for the cluster #2. These forecasting models use, as regressors, the past and current values of the supply air flow rate and the derived variable labelled “occupation”; those regressors are obtained from the entire data set available.

The same sequence A of pre-processing steps is performed for the other target variables: the cooling coil load, the whole building cooling load and the electric demand of the secondary and primary cooling system. The selected regressors for all target variables are listed in Table IV.18.

IV.3 Pre-processing of the cooling coil load on the air side

The cooling load on the air side of the cooling coil is calculated using the following equation:

$$\dot{Q}_{CCL}[kW] = \dot{v}_{SA} * \rho_A . 10^{-3} * (h_{MA} - h_{CD}) \qquad [IV.5]$$

Where, h_{CD} $[kJ/kg]$ and h_{MA} $[kJ/kg]$ are the cold deck and mixed air enthalpy, respectively calculated as averages of the two AHUs. $\dot{v}_{SA}$ $[L/s]$ is the volumetric air flow rate supplied by the both AHUs of the Genome building; and the density of the air ρ_A is assumed constant at 1.2 $[kg/m^3]$ at 20 $[C]$ and 101.3 $[kPa]$.

The cold deck air enthalpy is calculated based on the measured supply air temperature and relative humidity. The cold deck air temperature is calculated by subtracting the increase of temperature through the fan from the supply air temperature. An increase of the temperature of about 1.8 [°C] through the supply fans is determined from the measurements of the supply and mixed air temperatures over the year 2014 when the heating and cooling coil valves are closed as well as the humidifier (Figure IV.9). The absolute humidity of the air is assumed constant through the fan.

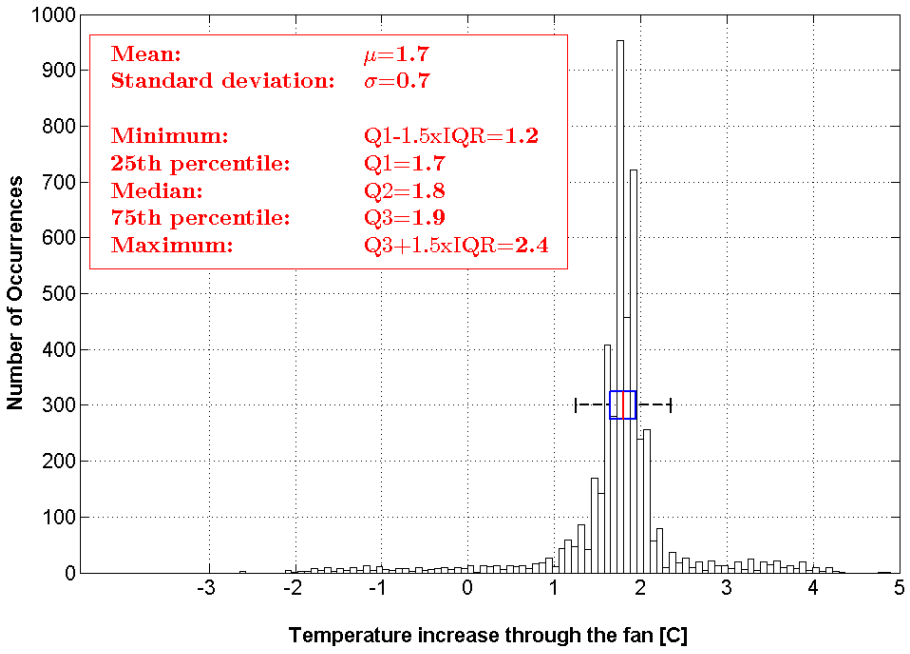

Figure IV.9 Temperature increase through the supply fan

The mixed air enthalpy is calculated based on the measured mixed air temperature and the derived relative humidity of the mixed air (RH_{MA}). The relative humidity of the mixed air is derived from the measured relative humidity of the outdoor (RH_{OA}) and return (RH_{RA}) air (equation [IV.6]).

$$RH_{MA} = \alpha * RH_{OA} + (1 - \alpha) * RH_{RA} \quad [IV.6]$$

Where, α is the ratio of outdoor air flow rate to the total supply air flow rate. When the mixing air dampers are completely closed, α equals 1. When they are completely open, α is derived from the measured supply ($\dot{V}_{SA}$) and return ($\dot{V}_{RA}$) air flow rate (equation [IV.7]).

$$\alpha = (\dot{V}_{SA} - \dot{V}_{RA})/\dot{V}_{SA} \quad [IV.7]$$

The measurements indicate that the mixed air dampers are mostly completely closed (modulation of 100%) or completely open (modulation of 0%) during summer 2014 (Figure IV.10). For the few times when the dampers are partially open (about 2 days out of 92), the ratio of outdoor air is calculated based on the measured mixed (MA), return (RA) and outdoor (OA) air temperatures (equation [IV.8]).

$$\alpha = \frac{T_{MA} - T_{RA}}{T_{OA} - T_{RA}} \quad [IV.8]$$

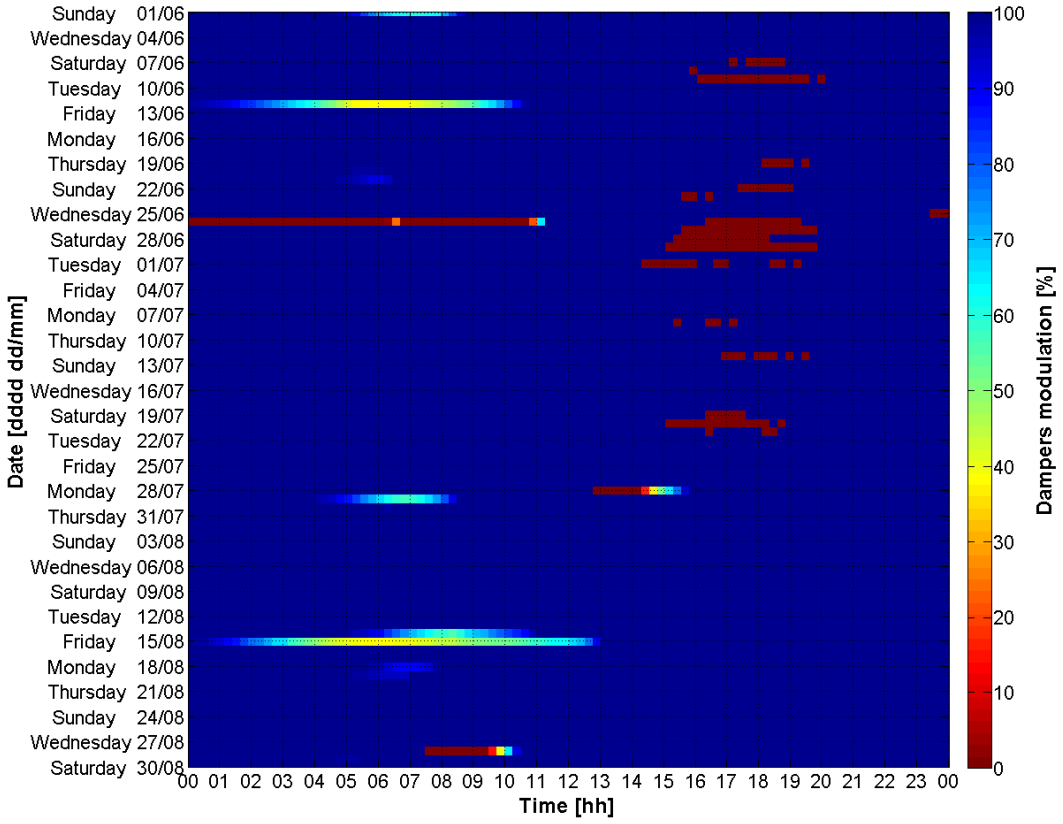

Figure IV.10 Operation of the mixed air dampers during summer 2014

IV.3.1 Exploratory analysis of the cooling coil load

The cooling coil load calculated on the air side of both AHUs over summer 2014 is presented in Figure IV.11. The squares with a color from green to dark blue correspond to a cooling coil load varying from 0 to 800 kW. The warmer colors from yellow to dark red with values from 0 to 150 kW correspond to "free cooling" periods when the air is not conditioned: the cooling coil, heating coil and humidifier are not operated. These "free-cooling" periods will be excluded from the modelling part of the cooling coil load.

The cooling coil load presents a daily pattern most of the days in summer with a higher cooling load from 08:00 (dark blue squares on Figure IV.11). This can be explained by the operation of the supply fans starting around 07:00 (Section Chapter IV). The pattern of the outdoor air temperature through summer 2014 is a second explanatory factor (Figure IV.12), with warmer temperature from 08:00, orange and red squares corresponding to a temperature of about 25 C.

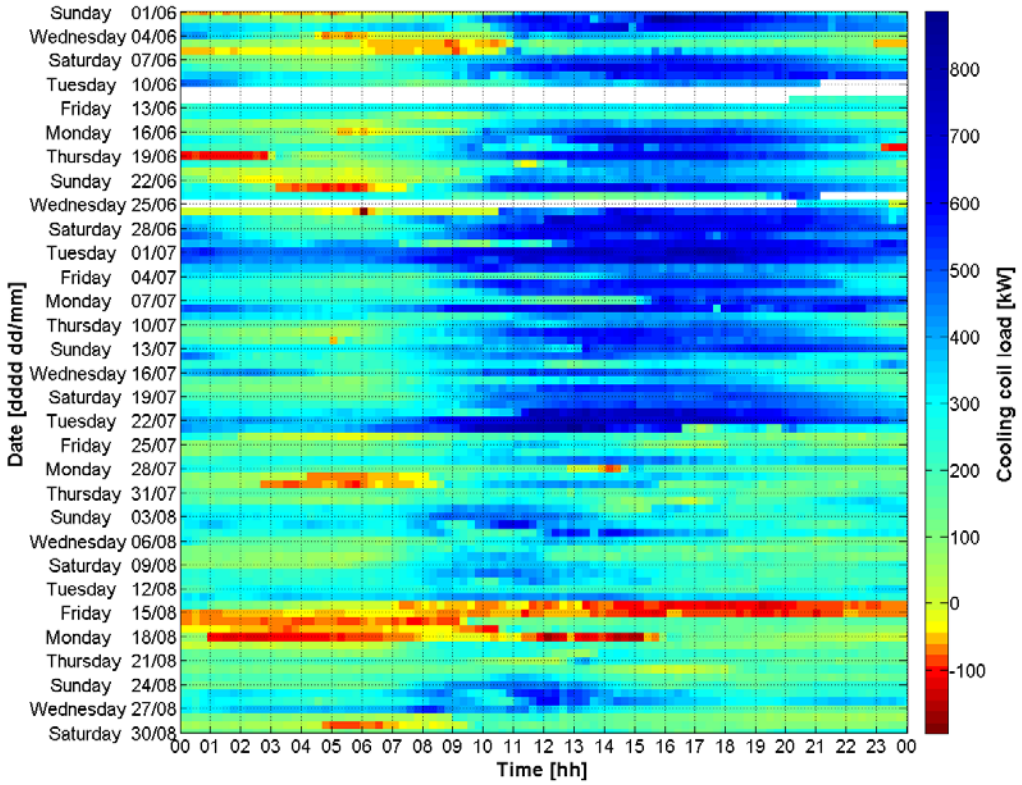

Figure IV.11 Cooling coil load in summer 2014

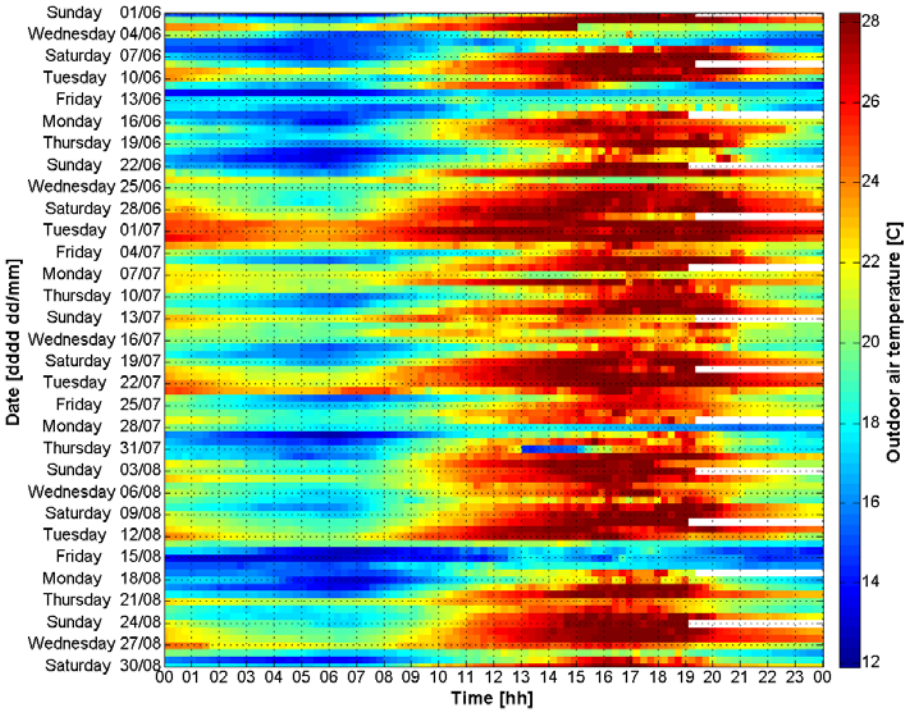

Figure IV.12 Outdoor air temperature in summer 2014

The probability distribution function of the cooling coil load over summer 2014 is displayed on Figure IV.13, for the occupied (in black) and unoccupied (in white) periods. An increase of the median cooling coil load is noticed from around 250 kW, during the unoccupied period, to 330 kW in the occupied period. The cooling coil load goes up to a maximum of about 850 kW in the occupied period.

The "heating" loads up to about 150 kW correspond to a "free-cooling" period when the cooling and heating coil valves are closed as well as the humidifier.

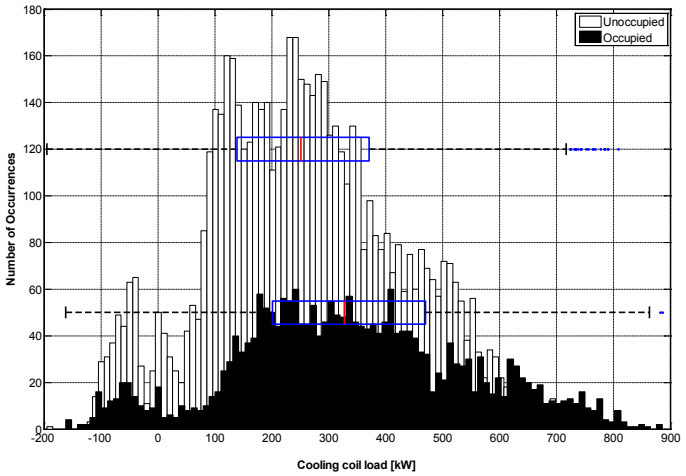

Figure IV.13 Probability distribution of the cooling coil load: occupied versus unoccupied periods

IV.3.2 Selection of the regressors for the forecasting of cooling coil load

IV.3.2.1 Identification of the relevant past values by cross-correlation analysis

The cross-correlation coefficient is calculated between the cooling coil load and the selected measured and derived variables listed in Table IV.1. A cross-correlation coefficient greater than 0.9 indicates a strong linear correlation [102]. The variables presenting a moderate to strong linear correlation with the cooling coil load are highlighted on Figure IV.14 by colors from orange to dark red with a white diamond. The cooling coil load is strongly correlated to its previous values from time t to t-60 min; and moderately correlated to its previous values until time t-210 min. The outdoor air enthalpy from time t to t-75 min is moderately correlated to the cooling coil load, and so is the mixed air temperature in both AHUs from time t to t-60 min.

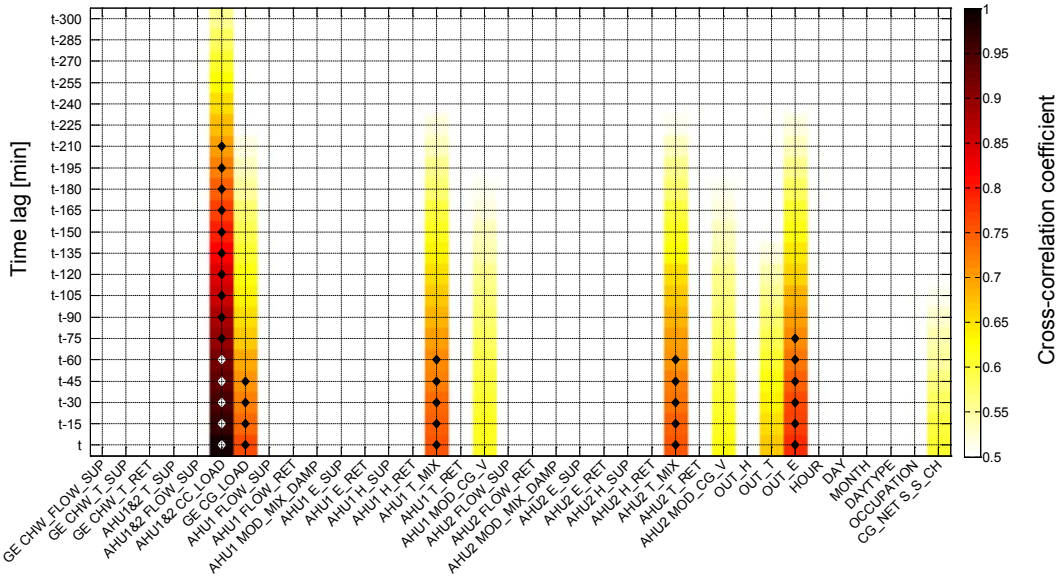

Figure IV.14 Potential regressors of the cooling coil load and their time lag

IV.3.2.2 Filtering of the relevant variables to the cooling coil load

The variables presenting a moderate to strong cross-correlation (greater than 0.7) with the cooling coil load, which are represented by black and white diamonds on Figure IV.14, are retained as regressors. These regressors are ranked in descending order of the cross-correlation coefficient (Table IV.5).

Table IV.5 Ranked regressors retained for the cooling coil load and their cross-correlation coefficient

Rank	Regressors	Description	Time lag (min)	Correlation coefficient
1	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t	0.985
2	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-15	0.968
3	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-30	0.950
4	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-45	0.932
5	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-60	0.914
6	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-75	0.896
7	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-90	0.877
8	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-105	0.858
9	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-120	0.838
10	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-135	0.817
11	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-150	0.795
12	OUT_E	Outdoor air enthalpy	t	0.782
13	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-165	0.772
14	OUT_E	Outdoor air enthalpy	t-15	0.771
15	OUT_E	Outdoor air enthalpy	t-30	0.758
16	AHU1 T_MIX	Mixed air temperature in AHU1	t	0.754
17	AHU2 T_MIX	Mixed air temperature in AHU2	t	0.753
18	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-180	0.750
19	AHU1 T_MIX	Mixed air temperature in AHU1	t-15	0.745
20	OUT_E	Outdoor air enthalpy	t-45	0.745
21	AHU2 T_MIX	Mixed air temperature in AHU2	t-15	0.744
22	AHU1 T_MIX	Mixed air temperature in AHU1	t-30	0.735
23	AHU2 T_MIX	Mixed air temperature in AHU2	t-30	0.734
24	OUT_E	Outdoor air enthalpy	t-60	0.731
25	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-195	0.726
26	AHU1 T_MIX	Mixed air temperature in AHU1	t-45	0.724
27	AHU2 T_MIX	Mixed air temperature in AHU2	t-45	0.723
28	OUT_E	Outdoor air enthalpy	t-75	0.715
29	AHU1 T_MIX	Mixed air temperature in AHU1	t-60	0.712
30	AHU2 T_MIX	Mixed air temperature in AHU2	t-60	0.710
31	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-210	0.703

The supply air flow rate does not appear in the list of regressors which are linearly correlated to the cooling coil load, even if the load is derived from it (Equation [IV.5]). The variation in the difference of supply and return air temperatures impacts more on the cooling coil load than the variation in the supply air flow rate. The supply air flow rate shows a very slight variation over the summer with a coefficient of variation (Equation [IV.9]) of 5% (Figure IV.4) compared to a more significant coefficient of variation of 18% for the difference of air temperature.

$$CV_{\dot{V}_{SA}} = \sigma_{\dot{V}_{SA}} / \mu_{\dot{V}_{SA}} \qquad [IV.9]$$

However, due to the domain knowledge and for the purpose of the cascade-based forecasting method the supply air flow rate at time t is also included as a regressor.

In conclusion, the cooling coil load can be forecasted using the previous values of the cooling coil load, the outdoor air enthalpy and the mixed air temperature in both AHUs:

$$\text{AHU1\&2 } CC\_LOAD(t+15) = \boldsymbol{f}\,(AHU1\&2\ CC\_LOAD(t,\dots,t-210), OUT\_E(t,\dots,t-75), AHU1\ T\_MIX(t,\dots,t-60), AHU2\ T\_MIX(t,\dots,t-60), AHU1\&2\ FLOW\_SUP(t)) \quad [IV.10]$$

IV.3.3 Identification of typical daily profiles of cooling coil load by clustering analysis

The fuzzy C-means clustering algorithm is used to analyse the daily profiles of the target variable and identify similar profiles based on the Euclidean distance. The daily profiles of the cooling coil load are the individuals of interest in this section. The profiles presenting missing values (Figure IV.11: from 10th to 12th and the 24th and 25th of June) are removed from the dataset. 87 daily profiles of the cooling coil load are remaining out of the 92 over summer 2014 (from the 1st of June to the 1st of September), each one composed of 96 values.

The individuals composed of 96 values are reduced to 6 features (Table IV.6). The mean and standard deviation of the cooling coil load over specific portions of the day are used as features to compare the profiles. The specific portions of the day have been determined based on a more detailed analysis. The profiles present three phases (Table IV.6): generally increasing from 07:00 to 12:00 and decreasing the rest of the time. The individuals compared by the clustering algorithm are composed of 6 features characterizing their pattern over the day.

Table IV.6 Features characterizing the cooling coil load daily profiles

Features	Description
$\boldsymbol{\mu_1, \sigma_1}$	Mean and standard deviation over the time period from 00:00 to 07:00
$\boldsymbol{\mu_2, \sigma_2}$	Mean and standard deviation over the time period from 07:00 to 12:00
$\boldsymbol{\mu_3, \sigma_3}$	Mean and standard deviation over the time period from 12:00 to 24:00

The clustering analysis of the cooling coil load is performed with three clusters for the typical profiles and a membership threshold of 70% after the trial-error procedure was used. The 15-min time-interval daily profiles of the three clusters are displayed on Figure IV.15. Forty-four profiles

remain non-classified, they are presented on Figure IV.16. The values of the clusters' centers returned by the clustering algorithm are presented in Table IV.7.

The daily profiles of the cooling coil load present some dissimilarities between each other. These dissimilarities make them difficult to group into clusters. General trends can still be observed: the profiles in cluster #1 present a continuous increase from $300 \pm 100\,kW$ at 06:00 to $600 \pm 200\,kW$ at 12:00 and a decrease the rest of the time. Cluster #2 and 3 present less fluctuating profiles over the day, respectively, around $200 \pm 200\,kW$ and $300 \pm 100\,kW$. Twenty-five profiles were not classified by the clustering algorithm. This shows the dissimilarity between the profiles as well as the variability in each cluster.

Table IV.7 Values of the features corresponding to the center of each cluster of the cooling coil load

	Features	Value [%]	Time period
Cluster #1	$\mu_1 \pm \sigma_1$	262.5 ± 47.1	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	373.0 ± 119.9	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	523.6 ± 93.4	12:00 to 24:00
Cluster #2	$\mu_1 \pm \sigma_1$	77.0 ± 34.9	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	135.5 ± 89.5	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	186.0 ± 63.9	12:00 to 24:00
Cluster #3	$\mu_1 \pm \sigma_1$	233.8 ± 30.4	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	304.4 ± 79.7	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	330.8 ± 83.2	12:00 to 24:00

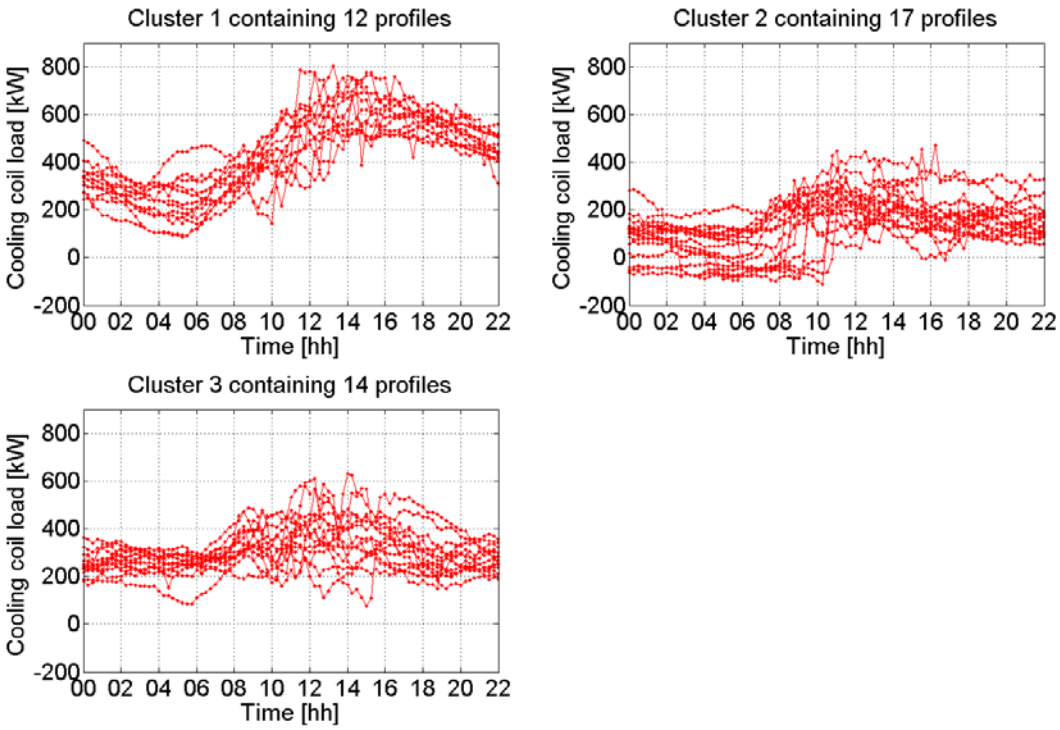

Figure IV.15 Typical daily profiles of the cooling coil load over summer 2014

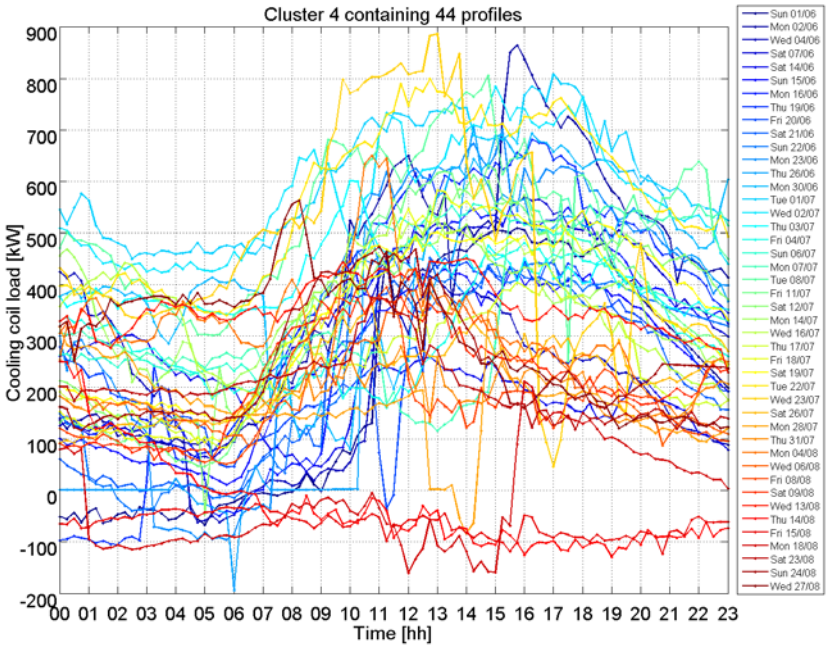

Figure IV.16 Atypical daily profiles of the cooling coil load over summer 2014

Three forecasting models should be developed for the cooling coil load, using as regressors: the previous values of the cooling coil load, the outdoor air enthalpy and the mixed air temperature in both AHU, with sequence A.

IV.4 Pre-processing of the cooling load of Genome building

The building cooling load is calculated using the following equation:

$$\dot{Q}_{GE}[kW] = \dot{V}_{CHW,GE} * \rho_W . 10^{-3} * C_{p_W} * (T_{CHWS} - T_{CHWR}) \qquad [IV.11]$$

Where, T_{CHWS} $[C]$ and T_{CHWR} $[C]$ are the measured temperatures of the chilled water, respectively, supplied to and returned from the Genome building. $\dot{V}_{CHW,GE}$ $[L/s]$ is the measured chilled water flow rate entering the Genome building. The density of the water ρ_W is assumed constant at 1000 $[kg/m^3]$ and the specific heat capacity of the water is 4.2 $[kJ/kg.C]$.

For a better analysis of the building cooling load, in this case, an exploratory analysis of the chilled water flow rate is performed to gather information on its impact on the cooling load. The measured chilled water flow rate supplied to the Genome building over summer 2014 is displayed on Figure IV.17. The flow rate varies from 0 L/s in dark blue to more than 60 L/s in dark red. The chilled water demand presents a general pattern with a lower demand in the occupied period of about 30 L/s in light blue compared to around 40 L/s in orange during the unoccupied period.

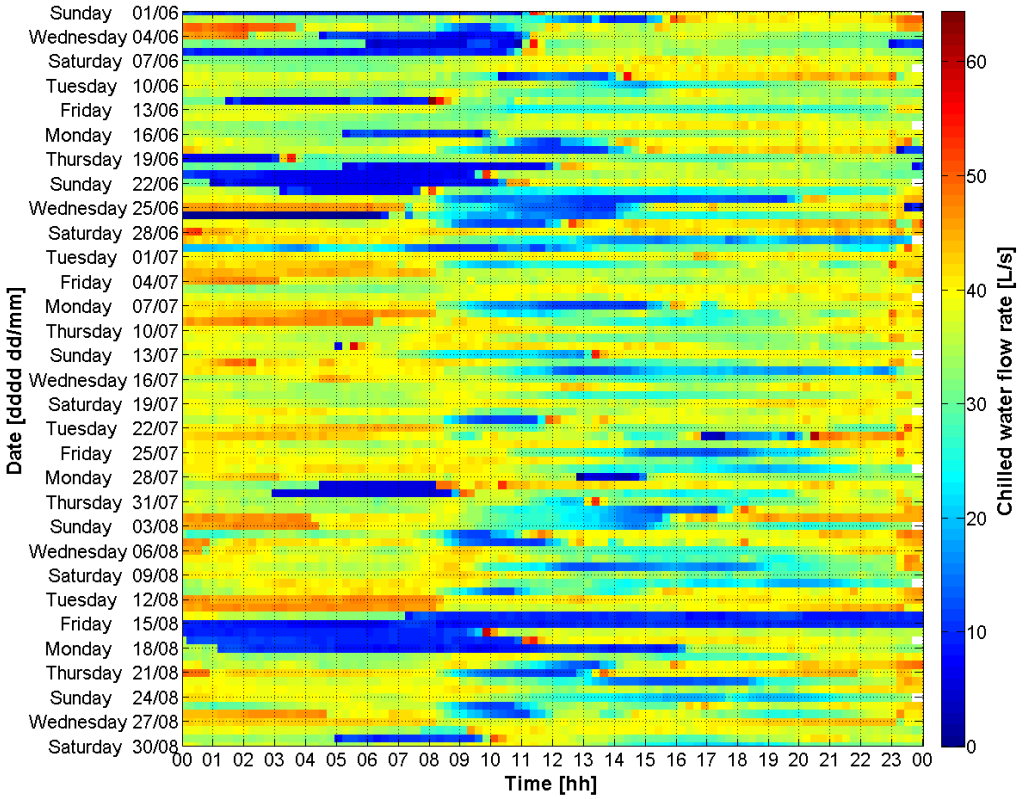

Figure IV.17 Chilled water flow rate over summer 2014

The probability distribution function of the chilled water flow rate supplied to the building in summer 2014 is showed in Figure IV.18. The higher chilled water demand during unoccupied periods is confirmed with a median water demand of 37 L/s compared to 32 L/s during occupied periods. The chilled water demand presents a mean of 29.7 $[L/s]$ with a standard deviation of 10.4 $[L/s]$ during occupied periods and 34.2 ± 9.9 $[L/s]$ during unoccupied periods.

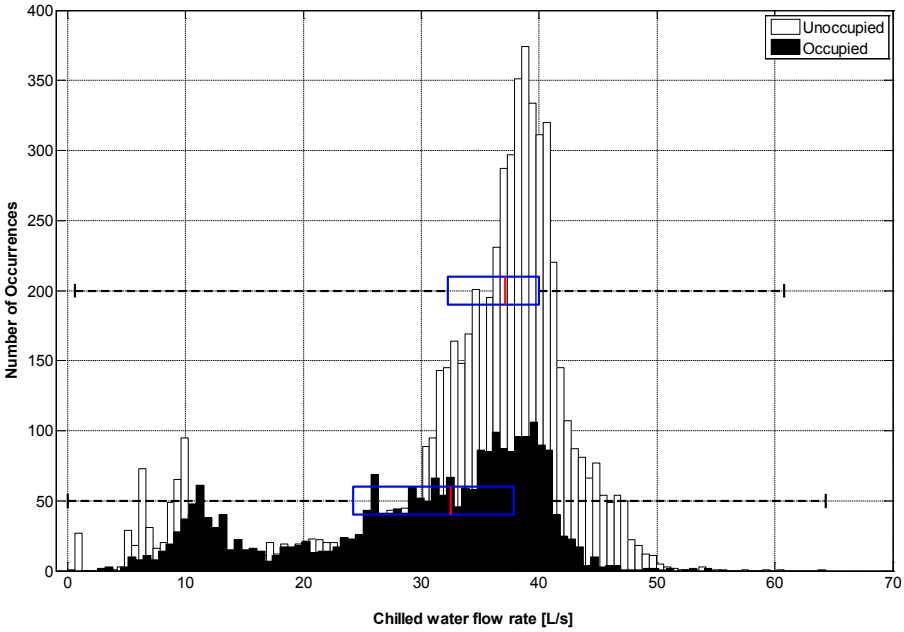

Figure IV.18 Probability distribution of the chilled water flow rate: occupied versus unoccupied periods

IV.4.1 Exploratory analysis of the cooling load of Genome building

The cooling load of the building is calculated over summer 2014 (Equation [IV.11]), from the 1st of June to the 1st of September, and displayed on Figure IV.19. The squares with colors from green to dark blue correspond to a cooling load varying from 0 to 1200 kW. The few points (eight measurement points) with warmer colors from yellow to red correspond to outliers where the calculated load is negative. Some values of the cooling load are missing from the 15th to the 17th of June (white squares).

Figure IV.19 shows a daily general pattern of the cooling load of Genome building; the load starts increasing around 08:00 (color going from blue to darker blue). Some days present no cooling requirement because of a low outdoor air temperature, for instance the cooling load about 12 °C on the 15th of August (Figure IV.12).

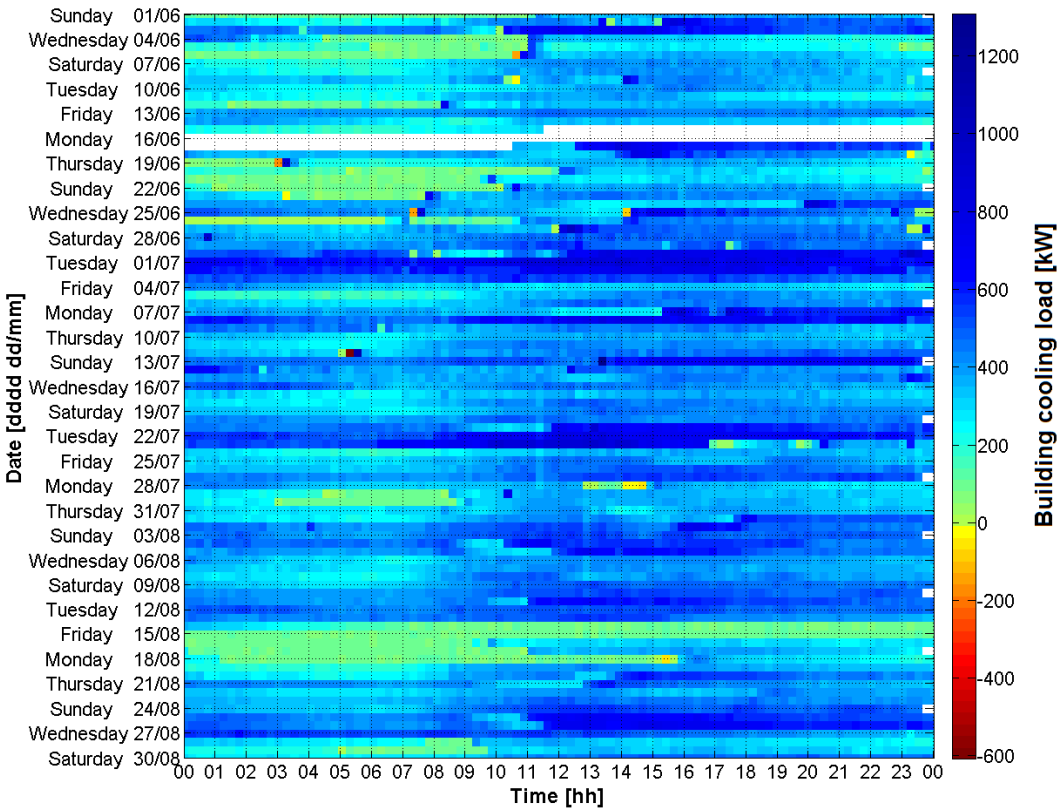

Figure IV.19 Building cooling load over summer 2014

The probability distribution function of the building cooling load over summer 2014 is displayed on Figure IV.20 with the occupied (in black) and unoccupied (in white) periods. The cooling load goes up to a maximum of about 900 kW in the occupied period which is around the maximum cooling coil load of 850 kW. The median building cooling load slightly increases from about 380 kW during unoccupied periods to 390 kW during occupied periods; compared to 250 to 330 kW for the median cooling coil load over the same periods. The difference is explained by the higher chilled water demand during unoccupied periods (Figure IV.18). Two peaks can be observed for both occupied and unoccupied periods.

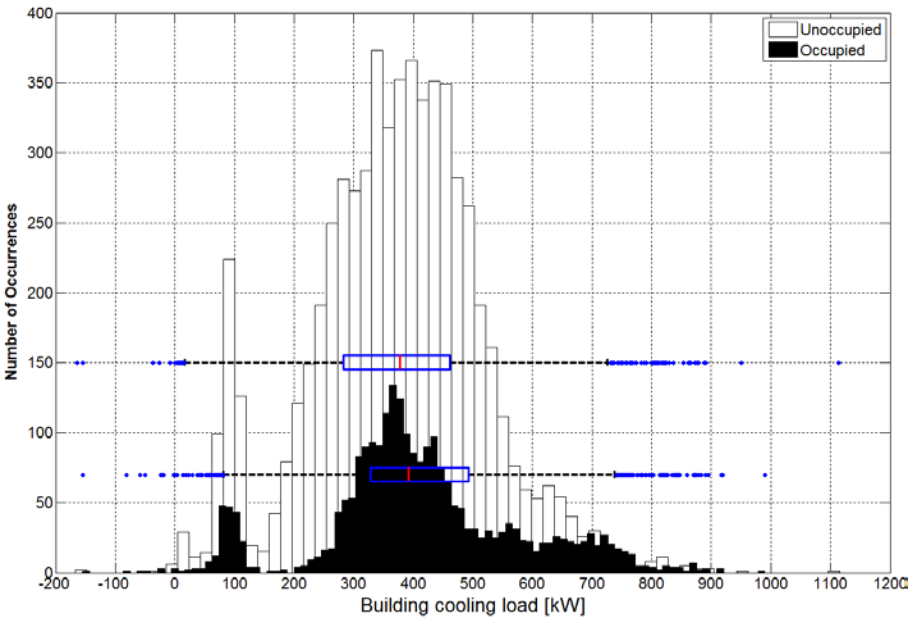

Figure IV.20 Probability distribution of the building cooling load corresponding to occupied and unoccupied periods

IV.4.2 Selection of the regressors for the forecasting of the building cooling load

IV.4.2.1 Identification of the relevant past values by cross-correlation analysis

The cross-correlation coefficient is calculated between the Genome building cooling load and the selected measured and derived variables listed in Table IV.1. The variables presenting a strong and moderate linear cross-correlation (greater than 0.7) to the cooling load of the building are highlighted in Figure IV.21. The cooling load presents a strong correlation to its previous values at time t and t-15 min (white diamonds); and a moderate correlation to its previous values until time t-210 min. A moderate linear correlation of the building cooling load with the cooling coil load, cooling coil valve modulation, and mixed air temperature in both AHUs is also noticed.

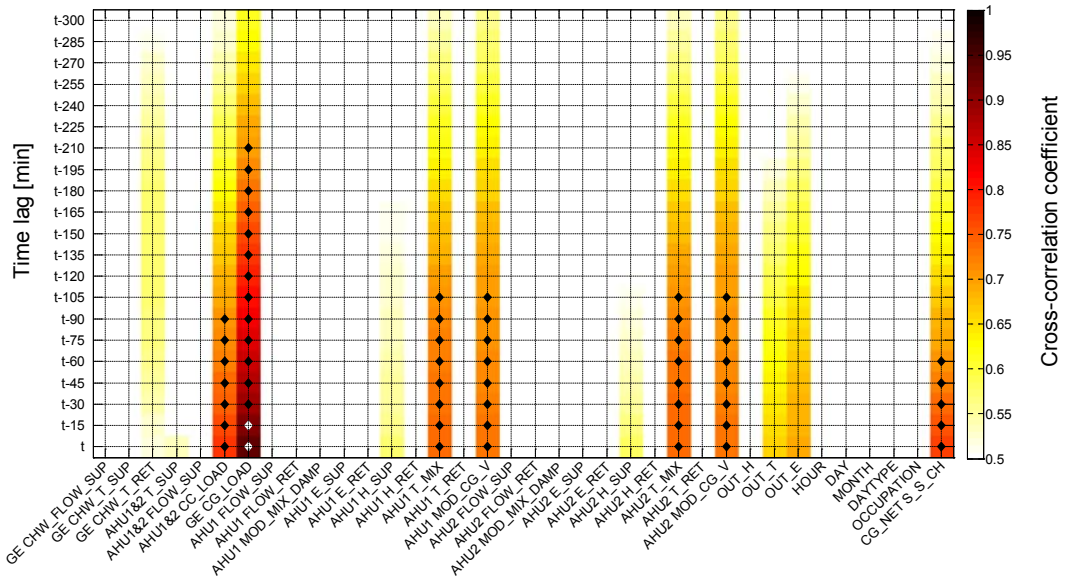

Figure IV.21 Relevant variables to the building cooling load and their time lag

IV.4.2.2 Filtering of the relevant variables to the building cooling load

The variables presenting a moderate to strong correlation with the cooling load of the building are retained as regressor (cross-correlation coefficient greater than 0.7). 54 regressors present a cross-correlation coefficient greater than 0.7. The regressors are ranked in descending order of the cross-correlation coefficient and only the first thirty-five are retained to avoid having too many regressors. These regressors are listed in Table IV.8.

Table IV.8 Ranked regressors retained for the building cooling load and their cross-correlation coefficient

Rank	Regressors	Description	Time lag (min)	Correlation coefficient
1	GE CG_LOAD	Building cooling load	t	0.932
2	GE CG_LOAD	Building cooling load	t-15	0.907
3	GE CG_LOAD	Building cooling load	t-30	0.886
4	GE CG_LOAD	Building cooling load	t-45	0.869
5	GE CG_LOAD	Building cooling load	t-60	0.853
6	GE CG_LOAD	Building cooling load	t-75	0.838
7	GE CG_LOAD	Building cooling load	t-90	0.822
8	GE CG_LOAD	Building cooling load	t-105	0.808
9	GE CG_LOAD	Building cooling load	t-120	0.793
10	GE CG_LOAD	Building cooling load	t-135	0.780
11	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t	0.777
12	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-15	0.763
13	GE CG_LOAD	Building cooling load	t-150	0.763
14	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-30	0.749
15	GE CG_LOAD	Building cooling load	t-165	0.748

16	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-45	0.737
17	AHU2 T_MIX	Mixed air temperature in AHU2	t-15	0.737
18	AHU2 T_MIX	Mixed air temperature in AHU2	t	0.737
19	AHU2 T_MIX	Mixed air temperature in AHU2	t-30	0.735
20	AHU1 T_MIX	Mixed air temperature in AHU1	t-15	0.735
21	AHU1 T_MIX	Mixed air temperature in AHU1	t	0.735
22	AHU1 T_MIX	Mixed air temperature in AHU1	t-30	0.734
23	GE CG_LOAD	Building cooling load	t-180	0.732
24	AHU2 T_MIX	Mixed air temperature in AHU2	t-45	0.732
25	AHU1 T_MIX	Mixed air temperature in AHU1	t-45	0.731
26	AHU2 MOD_CG_V	Cooling coil valve modulation in AHU2	t-15	0.729
27	AHU1 MOD_CG_V	Cooling coil valve modulation in AHU1	t-15	0.729
28	AHU2 MOD_CG_V	Cooling coil valve modulation in AHU2	t	0.728
29	AHU1 MOD_CG_V	Cooling coil valve modulation in AHU1	t	0.728
30	AHU2 T_MIX	Mixed air temperature in AHU2	t-60	0.727
31	AHU1&2 CC_LOAD	Cooling coil load on the air side of both AHUs	t-60	0.726
32	AHU1 T_MIX	Mixed air temperature in AHU1	t-60	0.726
33	AHU2 T_MIX	Mixed air temperature in AHU2	t-75	0.722
34	AHU1 T_MIX	Mixed air temperature in AHU1	t-75	0.721
35	AHU2 MOD_CG_V	Cooling coil valve modulation in AHU2	t-30	0.721

The building cooling load can be forecasted using as regressors the previous values of the building cooling load, the cooling coil load and the cooling coil valve modulation as well as mixed air temperature in both AHUs:

$$\begin{aligned}\text{GE CG\_LOAD}(t+15) = \mathbf{f}\,(&\text{GE CG\_LOAD}(t,\dots,t-180), \text{AHU1\&2 } CC\_LOAD(t,\dots,t-60),\\ &\text{AHU1 } T\_MIX(t,\dots,t-75), \text{AHU2 } T\_MIX(t,\dots,t-75),\\ &\text{AHU1 MOD\_CG\_V}(t,t-15), \text{AHU2 MOD\_CG\_V}(t,\dots,t-30))\end{aligned} \qquad \textit{[IV.12]}$$

IV.4.3 Identification of typical daily profiles of building cooling load by clustering analysis

A clustering analysis is performed on the daily profiles of the Genome building cooling load to group these profiles based on their similarity. Eighty-nine daily profiles are available; the profiles from the 15th to the 17th of June are removed because of missing values (Figure IV.19).

These daily profiles composed of 15-min interval measurements are reduced to 6 features (Table IV.9). The mean and standard deviation of the cooling load over specific portions of the day are used as features to compare the profiles. The portions of the day are identified from the profiles; they are the same as for the cooling coil load (Table IV.9).

Table IV.9 Features characterizing the daily profiles of the building cooling load

Features	Description
$\boldsymbol{\mu_1, \sigma_1}$	Mean and standard deviation over the time period from 00:00 to 07:00
$\boldsymbol{\mu_2, \sigma_2}$	Mean and standard deviation over the time period from 07:00 to 12:00
$\boldsymbol{\mu_3, \sigma_3}$	Mean and standard deviation over the time period from 12:00 to 24:00

A number of four clusters for typical profiles and a membership threshold of 65% is retained after a trial and error procedure. From the 89 available daily profiles, 36 are identified as atypical; they are presented in Figure IV.23. The values of the clusters' centers returned by the clustering algorithm are presented in Table IV.10.

Four types of daily profiles of the building cooling load were identified. In cluster #3, profiles present a low demand around 100 kW in the night and early morning; followed by a peak demand around 600 kW in the morning between 08:00 and 10:00 and varying around 400 kW in the rest of the day. Clusters #1, 2 and 4 gather profiles with a similar trend along the day: with a continuous increase from 06:00 to 12:00 and decreasing the rest of the time. The levels of cooling demand change: from 400 to 600 kW in cluster #2, from 600 to 800 kW in cluster #1 and from 200 to 400 kW in cluster #4.

The daily profiles of the cooling load of the building present dissimilarities between each other which make them difficult to group into cluster similarly to the analysis of the cooling coil load daily profiles. The membership threshold is reduced to 65% to avoid having too many profiles unclassified. 40% of the profiles remain unclassified with a threshold of 65%.

Table IV.10 Values of the features corresponding to the center of each cluster of the building cooling load

	Features	Value [%]	Time period
Cluster #1	$\mu_1 \pm \sigma_1$	506.9 ± 39.1	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	546.7 ± 80.5	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	645.0 ± 80.7	12:00 to 24:00
Cluster #2	$\mu_1 \pm \sigma_1$	403.6 ± 36.7	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	425.5 ± 56.3	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	475.4 ± 61.3	12:00 to 24:00
Cluster #3	$\mu_1 \pm \sigma_1$	134.4 ± 46.5	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	179.4 ± 106.7	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	307.4 ± 40.9	12:00 to 24:00
Cluster #4	$\mu_1 \pm \sigma_1$	283.5 ± 34.4	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	337.7 ± 58.7	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	388.7 ± 49.1	12:00 to 24:00

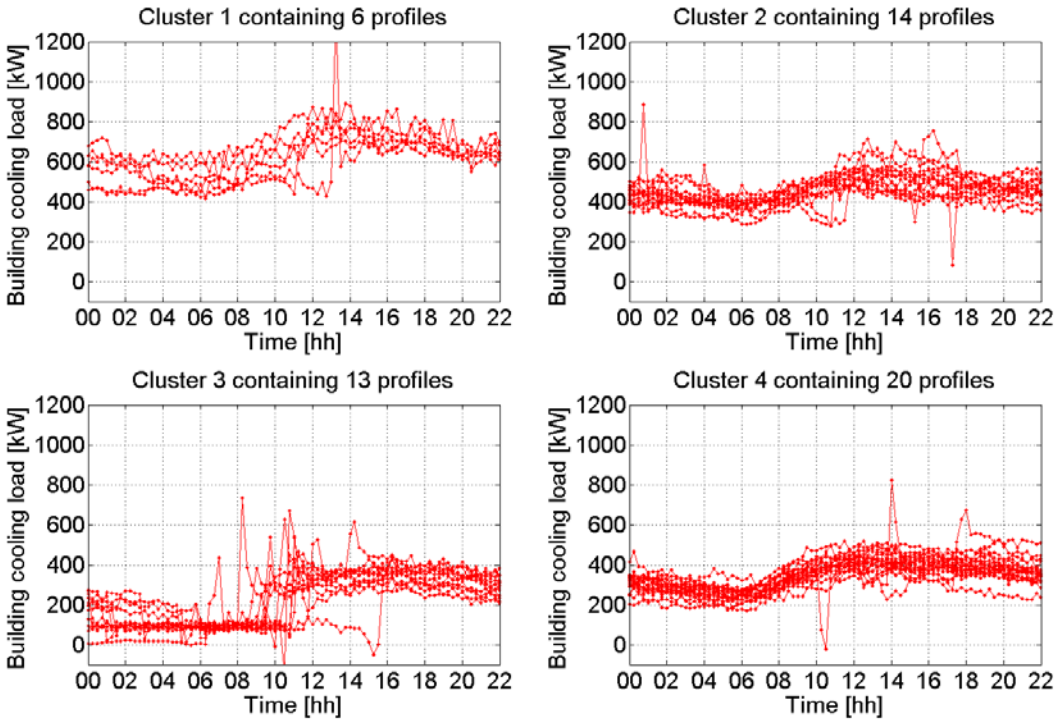

Figure IV.22 Four typical daily profiles of the building cooling load over summer 2014

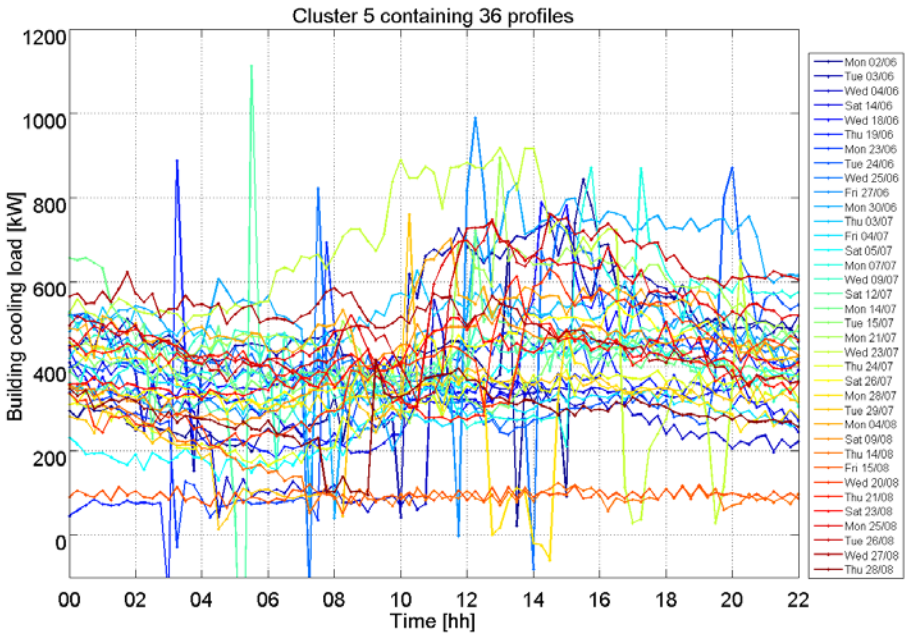

Figure IV.23 Atypical daily profiles of the building cooling load

Three forecasting models should be developed for the cooling load of GE building, using as regressors: the previous values of the building cooling load, the cooling coil load and the cooling coil valve modulation as well as mixed air temperature in both AHUs, with sequence A. A forecasting model is not developed for cluster #1 because it does not gather sufficient amount of profiles to train the model.

IV.5 Pre-processing of the electric demand of the secondary and primary cooling system

Some variables are selected among the 170 variables available in the database of measured and derived variables. This first global selection of variables potentially relevant to the electric demand of the primary cooling system is based on the knowledge of the system operation. The potential regressors considered for the forecast of the electric demand of primary cooling system are listed in Table IV.11.

The electric demand of the secondary cooling system includes four supply fans in two AHUs with a design capacity of 10,618 L/s each and an input electric demand of 29.8 kW (40 HP). The electric demand of the secondary cooling system ($\dot{E}_{sec.syst.}$) is estimated based on the measured supply air flow rate ($\dot{V}_{measured}$) and the input electric demand of the fan at full load given by the

manufacturer (Equation [III.3]). The electric demand of the secondary cooling system is forecasted using the forecasted value of the supply air flow rate in equation [III.3].

The electric demand of each chiller's compressor is measured and gathered through the BAS. The electric demand of the two chillers is summed and represented by the variable labelled "CG_NET CH_POWER_DEMAND".

The electric demand of each cooling tower's fan is estimated using the measured fan modulation (mod_{CT_1}) and the input electric demand ($P_{CT_{in}} = 29.8\ kW$) given by the manufacturer (Equation [IV.13]).

$$\dot{E}_{CT} = P_{CT_{in}}.(mod_{CT_1}/100)^3 + P_{CT_{in}}.(mod_{CT_2}/100)^3 \qquad [IV.13]$$

The electric demand of the four constant speed pumps is estimated based on the state of operation gathered through the BAS and the input electric demand given by the manufacturer. The four constant speed pumps include two circulating pumps for the chilled water loop with a power input of 74.6 kW and two for the condenser water loop with a power input of 55.9 kW.

Table IV.11 Selection of potential regressors for the electric demand of the primary cooling system

Points	Symbol	Description	Measured or Derived (M/D)	Units
Variables of interest				
AHU1&2 F_POWER_DEMAND	$\dot{E}_{sec.syst.}$	Electric demand of the four fans in AHUs	D	kW
CG_NET CH_POWER_DEMAND	$\dot{E}_{CH}$	Electric demand of the two chillers	M	kW
CG_NET CT_POWER_DEMAND	$\dot{E}_{CT}$	Electric demand of the two cooling towers	D	kW
CG_NET P_POWER_DEMAND	$\dot{E}_{P}$	Electric demand of the pumps	D	kW
CG_NET CHW_FLOW_SUP	$\dot{V}_{CHWS,CP}$	Supply chilled water flow rate	M	L/s
CG_NET CHW_T_SUP	$T_{CHWS,CP}$	Supply chilled water temperature	M	°C
CG_NET CHW_T_RET	$T_{CHWR,CP}$	Return chilled water temperature	M	°C
CG_NET CG_LOAD	$\dot{Q}_{CP}$	Central plant cooling load	D	kW
GE CG_LOAD	$\dot{Q}_{GE}$	Genome building cooling load	D	kW
Outdoor air				
OUT_H	RH_{OA}	Outside air relative humidity	M	%
OUT_T	T_{OA}	Outside air temperature	M	°C
OUT_E	h_{OA}	Outside air enthalpy	D	kJ/kg
Time				
HOUR	H	Hour of the day	D	-1,…,1

DAY	d	Day of the week	D	-1,…,1
MONTH	m	Month of the year	D	
DAYTYPE	DT	Working day or weekend	D	0,1
OCCUPATION	Occ	Occupied and unoccupied periods	D	0,1
Other indicators				
CG_NET S_S_CH	CH_{OP}	Operation of the two chillers	D	0, 1, 2

IV.5.1 Pre-processing of the electric demand of the secondary cooling system

IV.5.1.1 Exploratory analysis of the electric demand of the secondary cooling system

The electric demand of the four fans in the two AHUs of the Genome building is displayed over summer 2014, from the 1st of June to the 1st of September, on Figure IV.24. The electric demand varies from 0 kW in dark blue to over 30 kW in dark red.

The same pattern of operation of the fans as in Figure IV.3 is observed: fans operating from 08:00 to about 18:00 on working days.

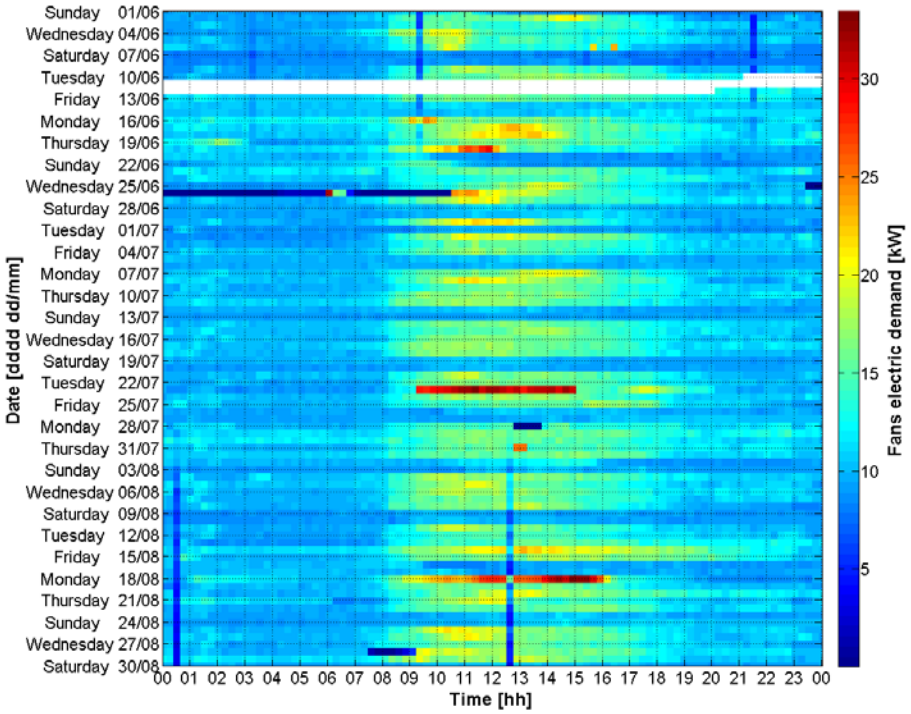

Figure IV.24 Electric demand of the secondary cooling system over summer 2014

The probability distribution function of the electric demand of the fans over summer 2014 is showed during occupied periods in black versus during unoccupied periods in white on Figure

IV.25. The median electric demand increases from 10.0 kW during unoccupied periods to 15.1 kW during occupied periods.

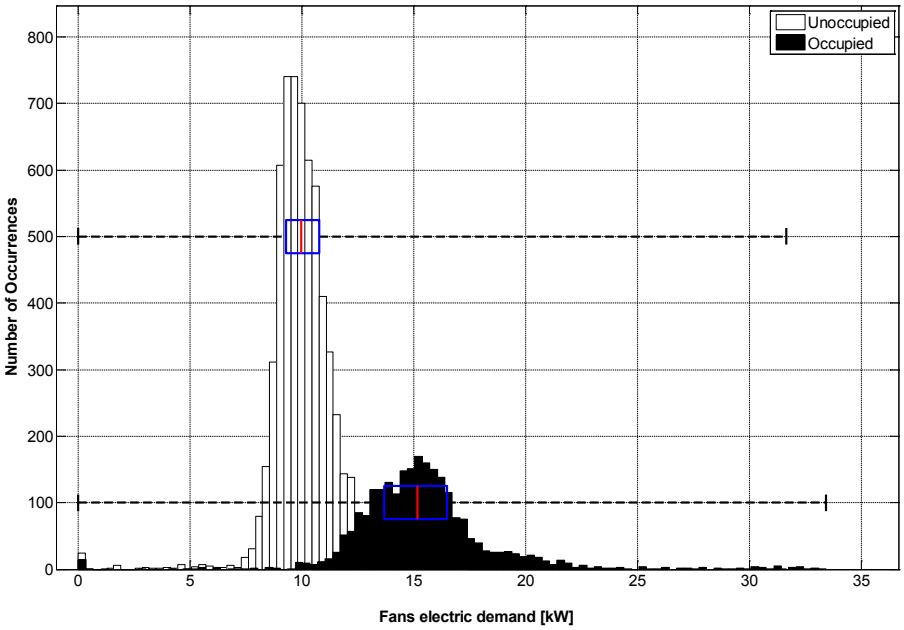

Figure IV.25 Probability distribution of the electric demand of the secondary cooling system over summer 2014

IV.5.1.2 Selection of the regressors for the forecasting of electric demand of the secondary cooling system

The forecast of the electric demand of the secondary cooling system is derived from the analysis of the supply air flow rate in section IV.2. The electric demand of the four supply fans in AHUs of Genome building is forecasted using the forecast of the supply air flow rate in equation [III.3].

$$\text{AHU1\&2 F\_POWER\_DEMAND}(t+15) = \mathbf{f}\ (\text{AHU1\&2 FLOW\_SUP}(t+15)) \qquad [IV.14]$$

IV.5.1.3 Identification of typical daily profiles of the fans' electric demand by clustering analysis

The identification of typical daily profiles of the electric demand of the secondary cooling system is not required. The clustering analysis was already performed on the supply air flow rate; the output of the forecasting model for the supply air flow rate will be used to forecast the electric demand of the secondary cooling system.

IV.5.2 Pre-processing of the chillers electric demand

IV.5.2.1 Exploratory analysis of the chillers electric demand

The electric demand of the two chillers is displayed over summer 2014 on Figure IV.26, from the 1st of June to the 1st of September. The color of the squares is representative of the chillers' electric demand from 0 kW in dark blue to more than 900 kW in dark red.

A general pattern of operation of the chillers can be observed from 08:00 to 00:00, highlighted by the warmer colors (yellow to red) over most of the days in summer.

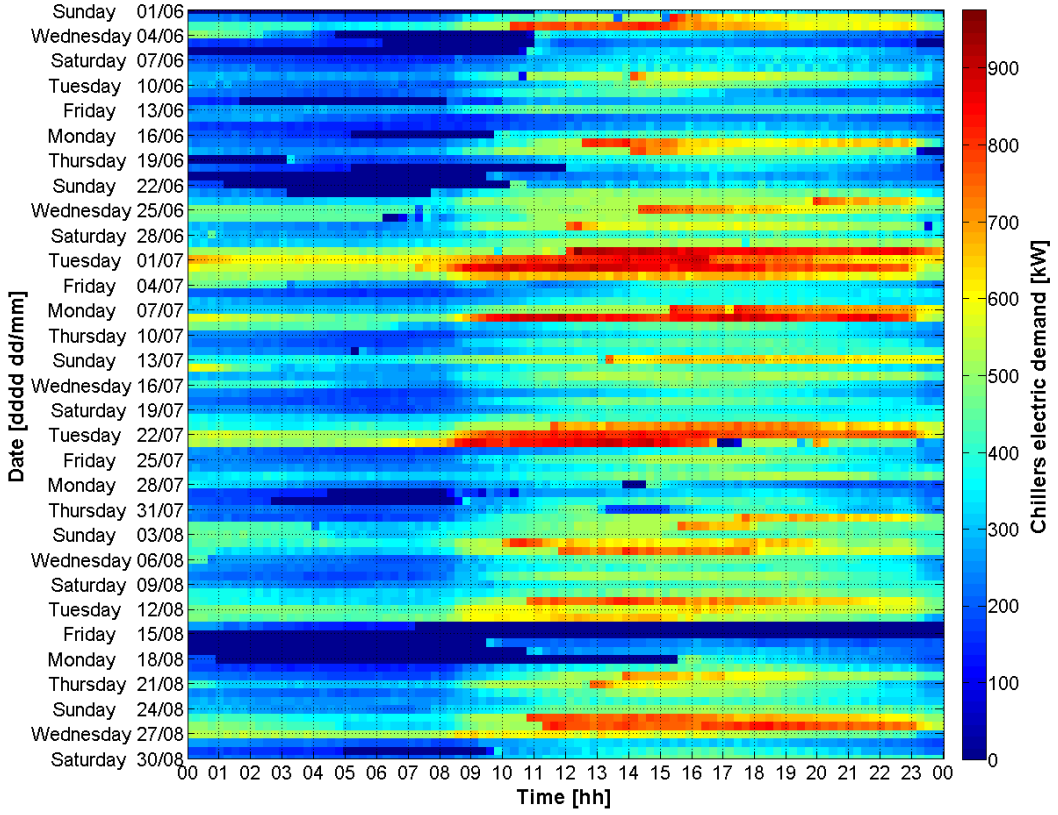

Figure IV.26 Electric demand of the chillers over summer 2014

The probability distribution function of the electric demand with one chiller in operation versus two chillers in operation over summer 2014 is presented in Figure IV.27. The median electric demand increases from about 320 kW, when one chiller is operated, to 600 kW, when both are used. The electric demand reaches a maximum of 975 kW when both chillers are in operation.

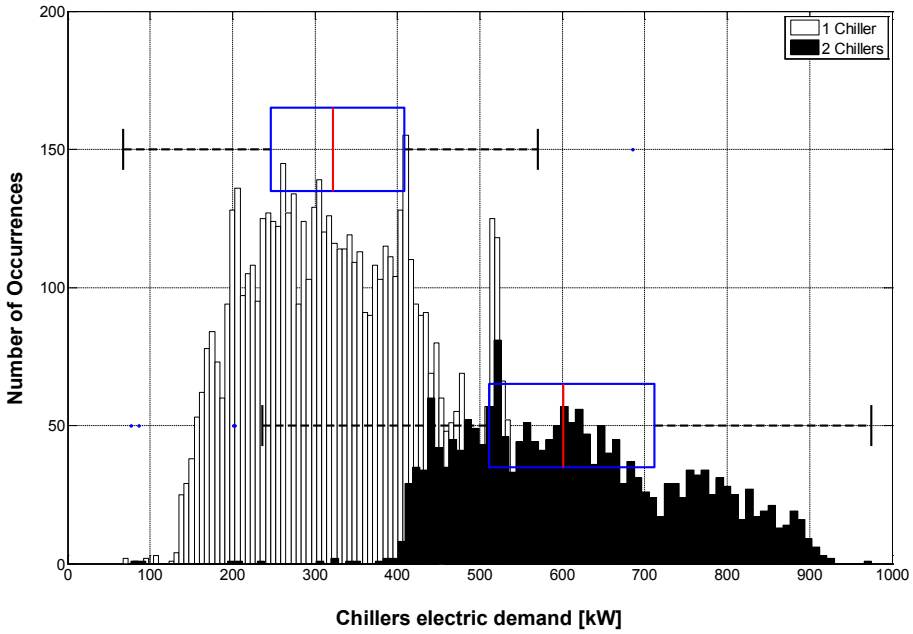

Figure IV.27 Probability distribution of the chillers' electric demand over summer 2014

IV.5.2.2 Selection of the regressors for the forecasting of electric demand of the chillers

IV.5.2.2.1 Identification of the relevant past values by cross-correlation analysis

The cross-correlation coefficient is calculated between the chillers' electric demand and the selected measured and derived variables listed in Table IV.11. The variables presenting a strong and moderate linear correlation to the chillers' electric demand are highlighted in Figure IV.28. The electric demand (CG_NET CH_POWER_DEMAND) at the next time step (t+15 min) presents a strong correlation to its previous values at time t to t-90 min (white diamonds); as well as with the central plant cooling load (CG_NET CG_LOAD) from time t to t-90 min, S.

A moderate linear correlation of the electric demand is also observed to the chilled water flow rate (CG_NET CHW_FLOW_SUP), building cooling load (GE CG_LOAD), outdoor temperature (OUT_T) and enthalpy (OUT_E) as well as chillers operation (CG_NET S_S_CH); it is highlighted by the black diamonds.

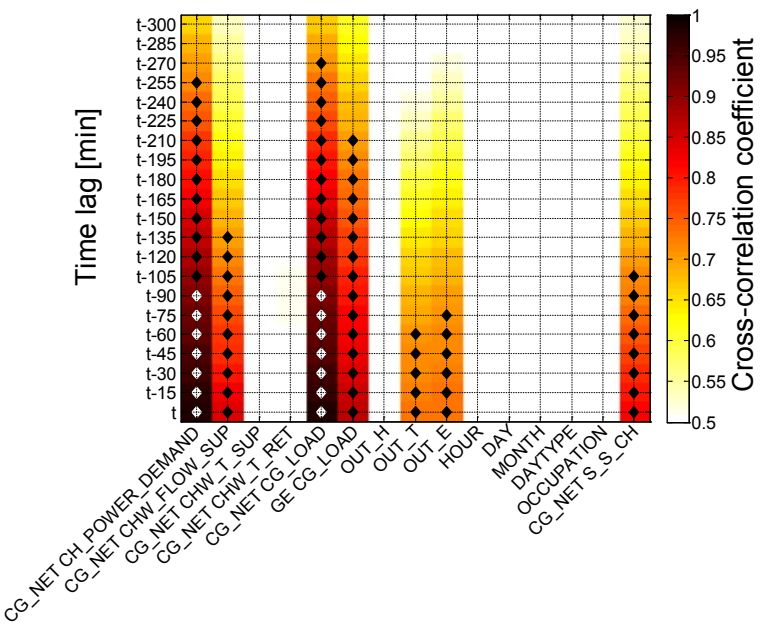

Figure IV.28 Relevant variables to the electric demand of the chillers and their time lag

IV.5.2.2.2 Filtering of the relevant variables to the chillers' electric demand

The variables presenting a moderate to strong correlation with the cooling load of the building are retained as regressor (cross-correlation coefficient greater than 0.7). In this case, 81 regressors present a cross-correlation coefficient greater than 0.7. The regressors are ranked in descending order of the cross-correlation coefficient and only the thirty-five first are retained to avoid having too many regressors. These regressors are listed in Table IV.12.

Table IV.12 Ranked regressors retained for the chillers' electric demand and their cross-correlation coefficient

Rank	Regressors	Description	Time lag (min)	Correlation coefficient
1	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t	0.983
2	CG_NET CG_LOAD	Central plant cooling load	t	0.981
3	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-15	0.972
4	CG_NET CG_LOAD	Central plant cooling load	t-15	0.966
5	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-30	0.958
6	CG_NET CG_LOAD	Central plant cooling load	t-30	0.956
7	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-45	0.948
8	CG_NET CG_LOAD	Central plant cooling load	t-45	0.945
9	CG_NET CG_LOAD	Central plant cooling load	t-60	0.935
10	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-60	0.934
11	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-75	0.923
12	CG_NET CG_LOAD	Central plant cooling load	t-75	0.922
13	CG_NET CG_LOAD	Central plant cooling load	t-90	0.911
14	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-90	0.908
15	CG_NET CG_LOAD	Central plant cooling load	t-105	0.896

16	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-105	0.895
17	CG_NET CG_LOAD	Central plant cooling load	t-120	0.883
18	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-120	0.878
19	CG_NET CG_LOAD	Central plant cooling load	t-135	0.867
20	GE CG_LOAD	Building cooling load	t	0.865
21	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-135	0.862
22	CG_NET CG_LOAD	Central plant cooling load	t-150	0.852
23	GE CG_LOAD	Building cooling load	t-15	0.852
24	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-150	0.845
25	GE CG_LOAD	Building cooling load	t-30	0.842
26	CG_NET CHW_FLOW_SUP	Supply chilled water flow rate	t	0.841
27	CG_NET CG_LOAD	Central plant cooling load	t-165	0.835
28	GE CG_LOAD	Building cooling load	t-45	0.833
29	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-165	0.828
30	GE CG_LOAD	Building cooling load	t-60	0.823
31	CG_NET CHW_FLOW_SUP	Supply chilled water flow rate	t-15	0.822
32	CG_NET CG_LOAD	Central plant cooling load	t-180	0.818
33	CG_NET S_S_CH	Operation of the two chillers	t	0.818
34	GE CG_LOAD	Building cooling load	t-75	0.814
35	CG_NET CH_POWER_DEMAND	Electric demand of the two chillers	t-180	0.810

The chillers' electric demand can be forecasted using, as regressors, the previous values of the electric demand, the chilled water flow rate from the cooling plant, the building and central plant cooling load, as well as the chillers operation:

$$\begin{aligned}\text{CG\_NET CH\_POWER\_DEMAND}(t+15) = \mathbf{f}\,(&\text{CG\_NET CH\_POWER\_DEMAND}(t, \dots, t-180),\\ &\text{CG\_NET CHW\_FLOW\_SUP}(t, t-15), \text{CG\_NET CG\_LOAD}(t, \dots, t-180), \text{GE CG\_LOAD}(t, \dots, t-75),\\ &\text{CG\_NET S\_S\_CH}(t))\end{aligned} \qquad \textit{[IV.15]}$$

IV.5.2.3 Identification of typical daily profiles of the electric demand of chillers by clustering analysis

A clustering analysis is performed on the daily profiles of the chillers' electric demand to group these profiles based on their similarity. Ninety-two daily profiles are available in summer 2014 (1st of June to 1st of September).

The individuals which are compared by the clustering algorithm are the number of chillers in operation at each 15-min during the day. The clustering analysis is performed on the number of chillers in operation instead of the electric demand itself because they are a good indicator of the chillers electric demand. They also are easier to differentiate since they present less variation. The daily profiles of chillers in operation, composed of 15-min interval values, are reduced to 6 features (Table IV.13). The mean and standard deviation of the number of chillers in operation over specific

portions of the day are used as features to compare the profiles. The portions of the day are identified from a detailed analysis of the chillers operation (Figure IV.26). The portions of the day are similar to ones of the cooling coil load and building cooling load (Table IV.13).

Table IV.13 Features characterizing the daily profiles of the chillers' electric demand

Features	Description
$\boldsymbol{\mu_1, \sigma_1}$	Mean and standard deviation over the time period from 00:00 to 08:00
$\boldsymbol{\mu_2, \sigma_2}$	Mean and standard deviation over the time period from 08:00 to 12:00
$\boldsymbol{\mu_3, \sigma_3}$	Mean and standard deviation over the time period from 12:00 to 24:00

A number of four clusters for typical profiles and a membership threshold of 70% is retained after a trial and error procedure. From the 92 available daily profiles, 26 are identified as atypical; they are presented in Figure IV.31. The values of the clusters' centers returned by the clustering algorithm are presented in Table IV.14.

The four clusters presenting the typical daily profiles of chillers' electric demand are presented in Figure IV.29. The corresponding number of chillers in operation over the day is displayed on Figure IV.30. Cluster #1 corresponds to days when one chiller is operated in the night (from 00:00 to 10:00); the electric demand varying from 200 to 500 kW. Both chillers are operated the rest of the time with an electric demand varying from 500 to 800 kW. The second chiller starts at a time varying from 10:00 to 14:00 highlighted by the peak demands in Figure IV.29. Both chillers are operated all day long in cluster #2 with an electric demand between 400 and 900 kW. A step increase in the electric demand is noticeable around 08:00. In cluster #3 none chiller is operated at night (from 00:00 to 08:00); and one is operated over the rest of the day. The peak electric demand appears at the starting time from 08:00 to 11:00; the demand then varies from 200 to 400 kW. Cluster #4 gathers most of the profiles; it corresponds to days with only one chiller in operation over the day. The electric demand varies from about 200 kW at night to 500 kW during the day. The increase in the demand starts from 08:00 and reaches its maximum over the period from 12:00 to 16:00.

Table IV.14 Values of the features corresponding to the center of each cluster of the number of chillers in operation

	Features	Value [%]	Time period
Cluster #1	$\mu_1 \pm \sigma_1$	1.1 ± 0.1	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	1.1 ± 0.2	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	1.8 ± 0.2	12:00 to 24:00
Cluster #2	$\mu_1 \pm \sigma_1$	2.0 ± 0.0	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	1.9 ± 0.0	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	1.9 ± 0.1	12:00 to 24:00
Cluster #3	$\mu_1 \pm \sigma_1$	0.3 ± 0.3	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	0.5 ± 0.4	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	0.9 ± 0.0	12:00 to 24:00
Cluster #4	$\mu_1 \pm \sigma_1$	1.0 ± 0.0	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	1.0 ± 0.0	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	1.0 ± 0.0	12:00 to 24:00

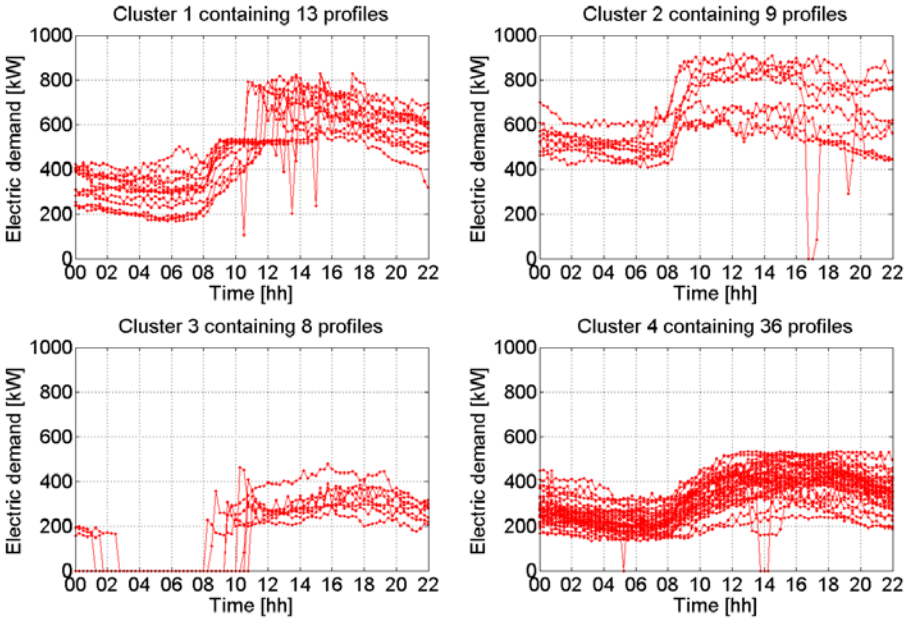

Figure IV.29 Typical daily profiles of the chillers' electric demand over summer 2014

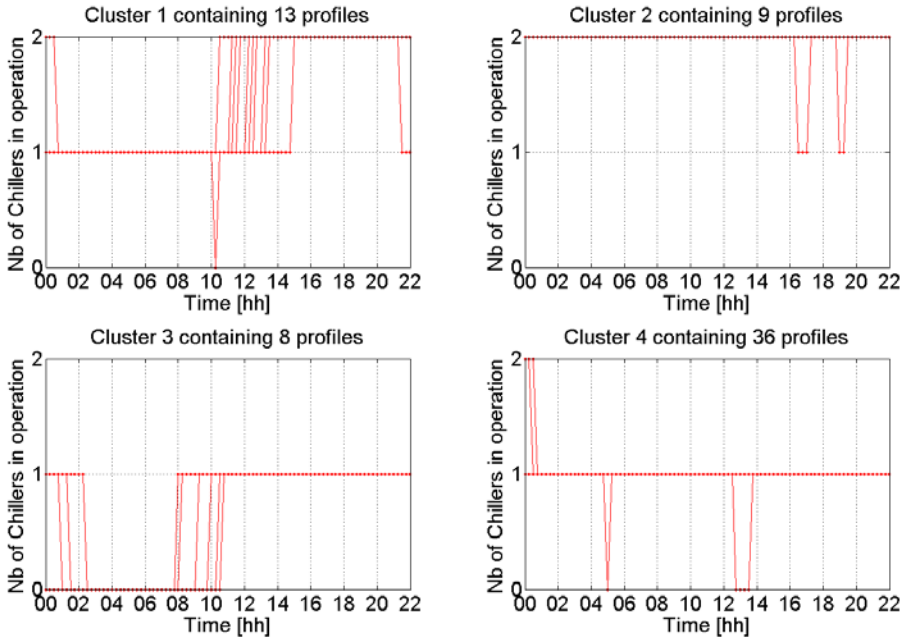

Figure IV.30 Typical daily profiles of chillers' operation over summer 2014

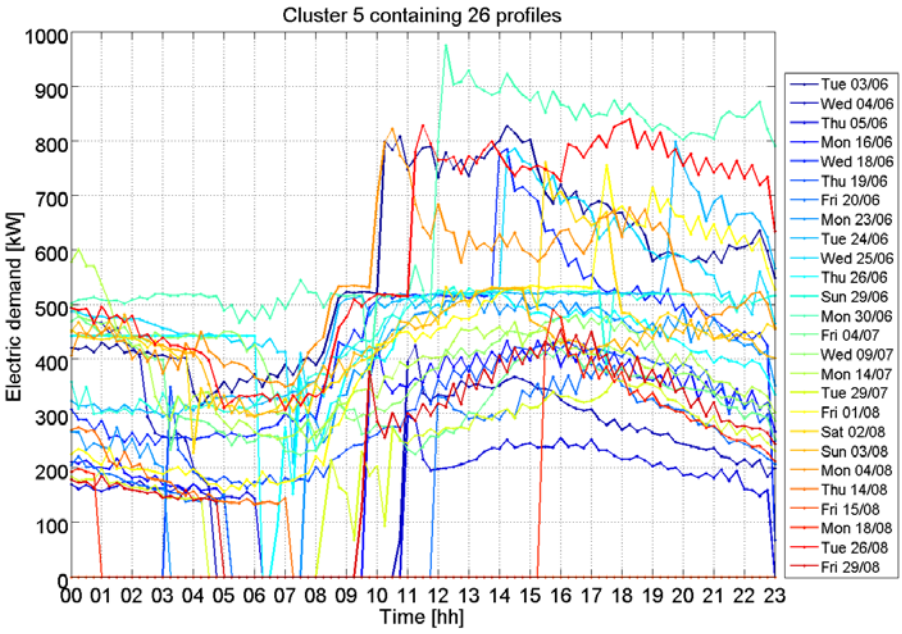

Figure IV.31 Atypical daily profiles of the chillers' electric demand over summer 2014

Four forecasting models should be developed for the electric demand of the chillers, using as regressors: the previous values of the electric demand, the chilled water flow rate from the cooling plant, the building and central plant cooling load, as well as the chillers operation, with sequence A.

IV.5.3 Preprocessing of the electric demand of the cooling towers

IV.5.3.1 Exploratory analysis of the electric demand of the cooling towers

The electric demand of the cooling towers over summer 2014 is displayed on Figure IV.32, from the 1st of June to the 1st of September. The electric demand varies from 0 kW in dark blue to over 55 kW in dark red.

A general pattern of operation of the cooling towers is observed from 08:00 to 00:00, highlighted by the colors from light blue to red over most of the days in summer.

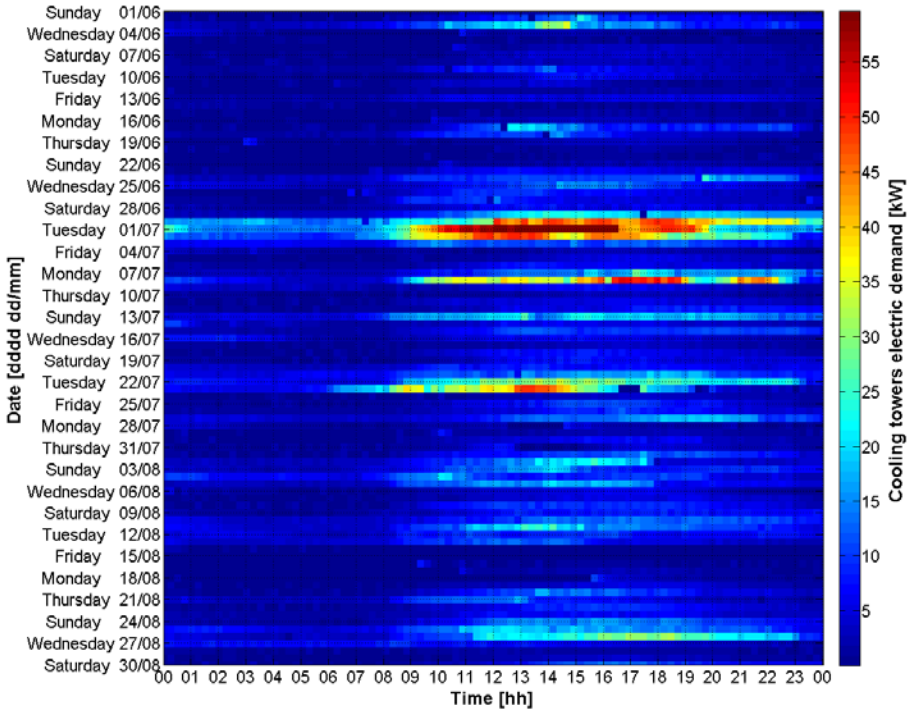

Figure IV.32 Electric demand of the cooling towers over summer 2014

The probability distribution function of the electric demand of the cooling towers over summer 2014 during occupied periods (in black) versus unoccupied periods (in white) is presented in Figure IV.33. The median electric demand increases from about 3 kW, during unoccupied periods, to 6 kW, during occupied periods.

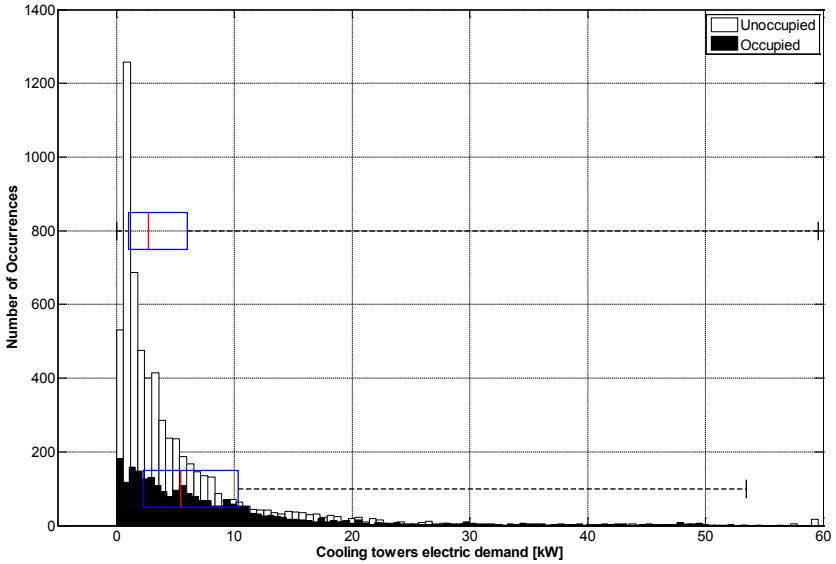

Figure IV.33 Probability distribution of the electric demand of the cooling towers over summer 2014

IV.5.3.2 Selection of the regressors for the forecasting of electric demand of the cooling towers

IV.5.3.2.1 Identification of the relevant past values by cross-correlation analysis

The cross-correlation coefficient is calculated between the electric demand of the cooling towers and the selected measured and derived variables listed in Table IV.11. The variables presenting a strong and moderate linear correlation to the electric demand of the cooling towers are highlighted in Figure IV.34. The electric demand (CG_NET CT_POWER_DEMAND) at the next time step (t+15 min) presents a strong correlation to its previous values at time t to t-75 min (white diamonds). A moderate correlation is observed with the electric demand of the cooling towers from time t-90 min to t-240 min and with the central plant cooling load (CG_NET CG_LOAD) from time t to t-120 min.

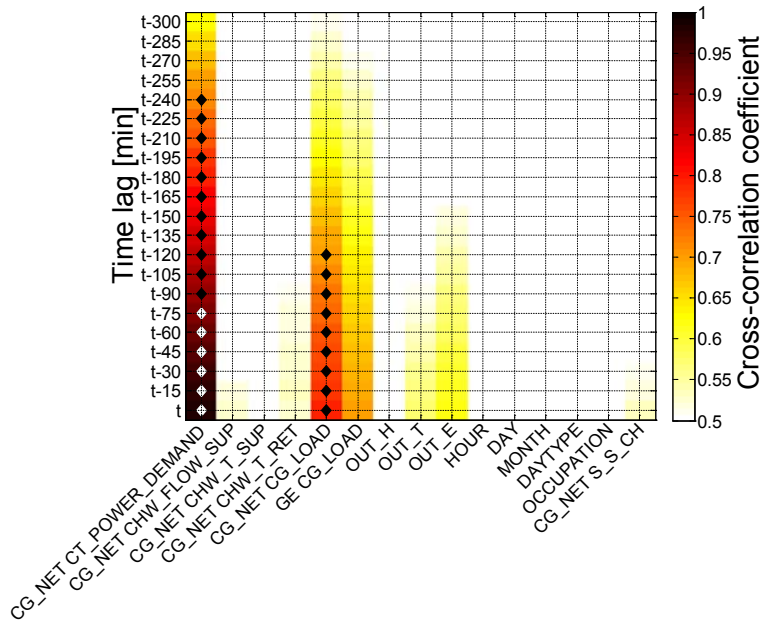

Figure IV.34 Potential regressors of the electric demand of the cooling towers and their time lag

IV.5.3.2.2 Filtering of the relevant variables to the chillers' electric demand

The variables presenting a moderate to strong correlation with the electric demand of the cooling towers are retained as regressor (cross-correlation coefficient greater than 0.7). 26 regressors present a cross-correlation coefficient greater than 0.7. The regressors are ranked in descending order of the cross-correlation coefficient and listed in Table IV.15.

Table IV.15 Ranked regressors retained for the electric demand of the cooling towers and their cross-correlation coefficient

Rank	Regressors	Description	Time lag (min)	Correlation coefficient
1	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t	0.982
2	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-15	0.971
3	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-30	0.957
4	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-45	0.944
5	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-60	0.928
6	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-75	0.913
7	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-90	0.897
8	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-105	0.881
9	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-120	0.864
10	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-135	0.848
11	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-150	0.831
12	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-165	0.814
13	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-180	0.795
14	CG_NET CG_LOAD	Central plant cooling load	t	0.791
15	CG_NET CG_LOAD	Central plant cooling load	t-15	0.780

16	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-195	0.777
17	CG_NET CG_LOAD	Central plant cooling load	t-30	0.770
18	CG_NET CG_LOAD	Central plant cooling load	t-45	0.761
19	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-210	0.758
20	CG_NET CG_LOAD	Central plant cooling load	t-60	0.750
21	CG_NET CG_LOAD	Central plant cooling load	t-75	0.740
22	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-225	0.738
23	CG_NET CG_LOAD	Central plant cooling load	t-90	0.728
24	CG_NET CT_POWER_DEMAND	Electric demand of the two cooling towers	t-240	0.718
25	CG_NET CG_LOAD	Central plant cooling load	t-105	0.716
26	CG_NET CG_LOAD	Central plant cooling load	t-120	0.702

The electric demand of the cooling towers can be forecasted using, as regressors, the previous values of the electric demand and the building cooling load:

$$\text{CG\_NET CT\_POWER\_DEMAND}(t+15) = \mathbf{f}\,(\text{CG\_NET CT\_POWER\_DEMAND}(t, \dots, t-240), \text{CG\_NET CG\_LOAD}(t, \dots, t-120)) \quad \textit{[IV.16]}$$

IV.5.3.3 Identification of typical daily profiles of the electric demand of cooling towers by clustering analysis

A clustering analysis is performed on the daily profiles of the electric demand of the cooling towers to group these profiles based on their similarity. 92 daily profiles are available in summer 2014, from the 1st of June to the 1st of September.

The daily profiles of the electric demand of the cooling towers are compared by the clustering algorithm. The daily profiles of electric demand, composed of 15-min interval values, are reduced to 6 features (Table IV.16). The mean and standard deviation of the electric demand over specific portions of the day are used as features to compare the profiles. The portions of the day are identified from a detailed analysis of the cooling towers operation (Figure IV.32). The portions of the day are the same as for the chillers' electric demand (Table IV.16).

Table IV.16 Features characterizing the daily profiles of the electric demand of the cooling towers

Features	Description
$\boldsymbol{\mu_1, \sigma_1}$	Mean and standard deviation over the time period from 00:00 to 08:00
$\boldsymbol{\mu_2, \sigma_2}$	Mean and standard deviation over the time period from 08:00 to 12:00
$\boldsymbol{\mu_3, \sigma_3}$	Mean and standard deviation over the time period from 12:00 to 24:00

A number of four clusters for typical profiles and a membership threshold of 70% is retained after a trial and error procedure. From the 92 available daily profiles, 29 are identified as atypical; they

are presented in Figure IV.36. The values of the clusters' centers returned by the clustering algorithm are presented in Table IV.17.

The four clusters presenting the typical daily profiles of electric demand of the cooling towers are presented in Figure IV.35. Cluster #1 corresponds to days when the number of cooling towers in operation varies between zero and one from 00:00 to 08:00, corresponding to an electric demand varying from zero to 5 kW. The rest of the day the operation of the cooling towers varies from one to two, with an electric demand reaching a maximum of 20 kW. Cluster #2 gathers daily profiles when one or no cooling tower is operated; the electric demand varies from zero to 10 kW. Both cooling towers are in operation in the daily profiles of cluster #3; with a higher electric demand increasing from about 15 kW at night to about 40 kW and a maximum of 60 kW from 10:00 to 16:00. Cluster #4 presents daily profiles with an electric demand which starts at about 5 kW from 00:00 to 08:00 and increase continuously until 14:00 to reach about 20 kW. It corresponds to days when one cooling tower works at night and both during the day.

Table IV.17 Values of the features corresponding to the center of each cluster of the electric demand of the cooling towers

	Features	Value [%]	Time period
Cluster #1	$\mu_1 \pm \sigma_1$	2.2 ± 0.7	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	4.9 ± 2.4	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	7.9 ± 2.8	12:00 to 24:00
Cluster #2	$\mu_1 \pm \sigma_1$	1.0 ± 0.5	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	1.4 ± 0.9	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	2.6 ± 1.0	12:00 to 24:00
Cluster #3	$\mu_1 \pm \sigma_1$	11.3 ± 2.6	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	28.0 ± 12.2	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	35.6 ± 13.0	12:00 to 24:00
Cluster #4	$\mu_1 \pm \sigma_1$	4.7 ± 1.1	00:00 to 07:00
	$\mu_2 \pm \sigma_2$	10.9 ± 4.7	07:00 to 12:00
	$\mu_3 \pm \sigma_3$	15.5 ± 4.9	12:00 to 24:00

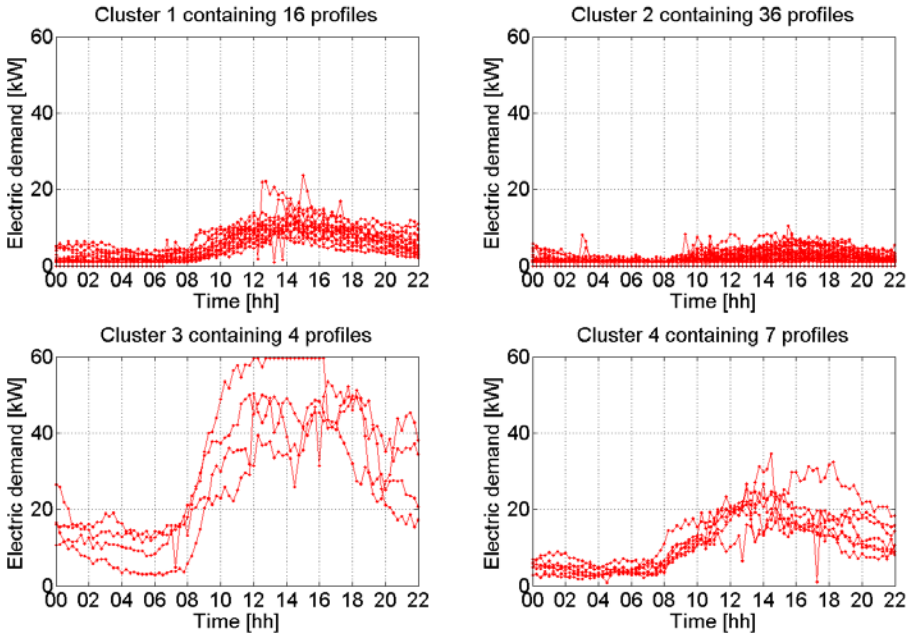

Figure IV.35 Typical daily profiles of the electric demand of the cooling towers over summer 2014

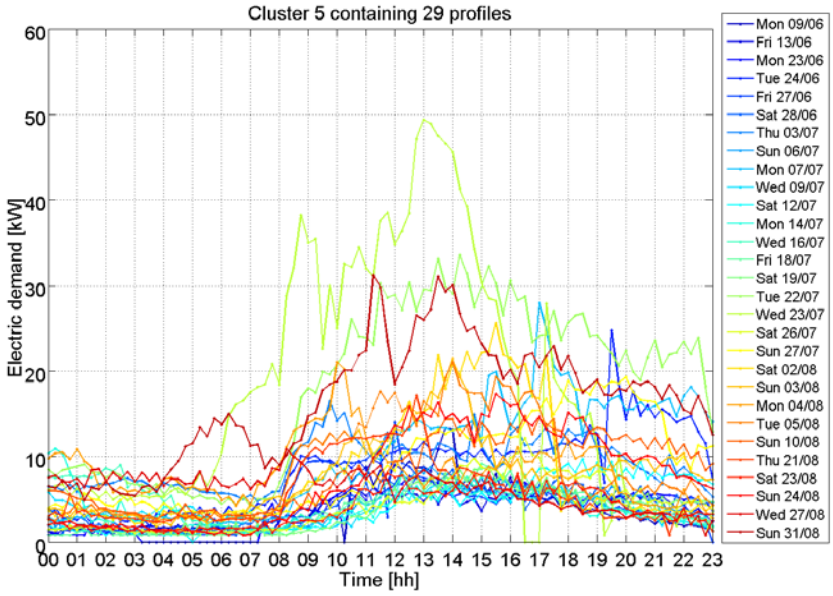

Figure IV.36 Atypical daily profiles of the electric demand of the cooling towers over summer 2014

Two forecasting models should be developed for the electric demand of the cooling towers, using as regressors: the previous values of the electric demand and the building cooling load, with sequence A. Forecasting models are not developed for clusters #3 and 4 because they do not gather sufficient amount of profiles to train the model.

IV.6 Different sequences of pre-processing steps

The list of regressors selected for each target variable according to each sequence of pre-processing steps A, B and C is presented for comparison in Table IV.18. The three different sequences of pre-processing steps A, B and C present a common first step: the exploratory data analysis. The same filtering technique is then applied in all three sequences for the selection of regressors.

1) In the sequence A, the regressors of each target variable are selected by using the whole data set available, followed by identification of different clusters. For each target variable, a forecasting model is developed for each cluster by using the regressors from the entire data set.

2) In the sequence B, the clusters are first identified, followed by the selection of regressors of each target variable and for each cluster. The target variable is forecasted for each cluster by using regressors that are specific to that cluster.

3) In the sequence C, the regressors are selected for each target variable by using the whole data set available (the same regressors as in sequence A). For each target variable, only one forecasting model is developed using the regressors from the entire data set, since there are no clusters.

It is noticed that when the selection of the regressors is performed over a subset of measurements corresponding to similar daily profiles of the target variable, the number of variables retained as regressors is reduced in certain cases. The comparison of selected regressors for the three sequences of pre-processing steps shows that the clusters presenting daily profiles with few variations over the day present only one regressor: the past values of the target variable. This trend is observed on cluster #2 of supply air flow rate (Figure IV.6), clusters #2 and 3 of the cooling coil load (Figure IV.15) and cluster #2 of the building cooling load (Figure IV.22). However in clusters where the daily profiles of the target variable show a large variation, the number of regressors increases.

For each target variable, the main regressor remains the past values of the target variable. When comparing sequences A and C of pre-processing steps to sequence B, in several cases, the hour of the day is selected as a regressor with sequence B (e.g. cluster #1 of the supply air flow rate; clusters #1, 3 and 4 of the building cooling load). The subsets of similar daily profiles of the target variable, obtained by clustering analysis, present daily patterns correlated to the hour of the day that were not detected when merged in the original dataset.

Table IV.18 Selected regressors for each target variable with the three sequences of pre-processing steps

Target variable	Sequences A and C of pre-processing steps	Sequence B of pre-processing steps
Supply air flow rate	**Common regressors:** • AHU1&2 FLOW_SUP(t, … , t − 105) • OCCUPATION(t, … , t − 90)	**Cluster #1** • AHU1&2 FLOW_SUP(t, … , t − 105), • OCCUPATION(t, … , t − 90), • HOUR(t, … , t − 60) **Cluster #2** • AHU1&2 FLOW_SUP(t, … , t − 45),
Cooling coil load	**Common regressors:** • AHU1&2 CC_LOAD(t, … , t − 210) • OUT_E(t, … , t − 75) • AHU1 T_MIX(t, … , t − 60) • AHU2 T_MIX(t, … , t − 60) • AHU1&2 FLOW_SUP(t)	**Cluster #1** • AHU1&2 CC_LOAD(t, … , t − 90) • HOUR(t − 15, … , t − 195) • OUT_T(t) • OUT_E(t) • AHU1 T_MIX(t, … , t − 75) • AHU2 T_MIX(t, … , t − 90) **Cluster #2** • AHU1&2 CC_LOAD(t, … , t − 75) **Cluster #3** • AHU1&2 CC_LOAD(t, … , t − 45)
Building cooling load	**Common regressors:** • GE CG_LOAD(t, … , t − 180) • AHU1&2 CC_LOAD(t, … , t − 60) • AHU1 T_MIX(t, … , t − 75) • AHU2 T_MIX(t, … , t − 75) • AHU1 MOD_CG_V(t, t − 15) • AHU2 MOD_CG_V(t, … , t − 30)	**Cluster #1** • GE CG_LOAD(t, … , t − 75) • AHU1&2 CC_LOAD(t, … , t − 90) • AHU1 T_MIX(t, … , t − 75) • AHU2 T_MIX(t, … , t − 75) • OUT_T(t, t − 15) • OUT_E(t, t − 15) • HOUR(t − 105, . . , t − 150) **Cluster #2** • GE CG_LOAD(t, … , t − 30) **Cluster #3** • GE CG_LOAD(t, … , t − 60) • AHU1&2 CC_LOAD(t, … , t − 30) • GE CHW_FLOW_SUP(t, t − 15) • AHU1&2 T_SUP(t) • AHU1 MOD_CG_V(t, … , t − 60) • AHU2 MOD_CG_V(t, … , t − 60) • HOUR(t − 135, … , t − 180) **Cluster #4** • GE CG_LOAD(t, … , t − 75) • AHU1 T_MIX(t, … , t − 60) • AHU2 T_MIX(t, … , t − 60) • HOUR(t − 90, t − 165)

Chillers' electric demand	**Common regressors:** • CG_NET CH_POWER_DEMAND(t, … t − 180) • CG_NET CHW_FLOW_SUP(t, t − 15) • CG_NET CG_LOAD(t, … , t − 180) • GE CG_LOAD(t, … , t − 75) • CG_NET S_S_CH(t)	**Cluster #1** • CG_NET CH_POWER_DEMAND(t, … , t − 120) • CG_NET CG_LOAD(t, … , t − 120) • CG_NET CHW_FLOW_SUP(t, t − 15) • CG_NET S_S_CH(t) • HOUR(t − 30, . . , t − 225) **Cluster #2** • CG_NET CH_POWER_DEMAND(t, … , t − 105) • CG_NET CG_LOAD(t, … , t − 120) • GE CG_LOAD(t, … , t − 60) • CG_NET T_RET(t, … , t − 60) • OUT_E(t, t − 15) **Cluster #3** • CG_NET CH_POWER_DEMAND(t, … , t − 75) • CG_NET CG_LOAD(t, … , t − 75) • GE CG_LOAD(t, … , t − 45) • CG_NET CHW_FLOW_SUP(t, … , t − 45) • CG_NET S_S_CH(t, … , t − 45) • HOUR(t − 75, . . , t − 225) **Cluster #4** • CG_NET CH_POWER_DEMAND(t, … , t − 120) • CG_NET CG_LOAD(t, … , t − 135) • GE CG_LOAD(t, … , t − 45) • CG_NET CHW_T_RET(t, … , t − 135) • OUT_T(t, t − 15)
Electric demand of the cooling towers	**Common regressors:** • CG_NET CT_POWER_DEMAND(t, … t − 240) • CG_NET CG_LOAD(t, … , t − 120)	**Cluster #1** • CG_NET CT_POWER_DEMAND(t, … , t − 105) • CG_NET CG_LOAD(t, … , t − 60) • CG_NET CHW_T_RET(t, … , t − 105) • HOUR(t, . . . , t − 180) **Cluster #2** • CG_NET CT_POWER_DEMAND(t, … , t − 90) • CG_NET CG_LOAD(t, … , t − 75) • OUT_T(t, . . . , t − 45) **Cluster #3** • CG_NET CT_POWER_DEMAND(t, … , t − 105) • CG_NET CG_LOAD(t, … , t − 60) • OUT_T(t, … , t − 45) • OUT_E(t, … , t − 30) • HOUR(t, . . . , t − 210) **Cluster #4** • CG_NET CT_POWER_DEMAND(t, … , t − 105) • CG_NET CG_LOAD(t, … , t − 75) • HOUR(t, . . . , t − 180)

Chapter V. Cascade-based forecasting method Case study: multistep-ahead forecast of the electric demand

This chapter presents the development of forecasting models as a cascade-based method presented in the flowchart of the previous chapter (Figure IV.39). Forecasting models are developed for each target variable starting from the air flow rate supplied by the AHUs to the zones up to the electric demand of the secondary and primary cooling systems.

For each target variable, a forecasting model is developed based on the outputs of the preprocessing steps (chapter IV): the selected common regressors (table IV.19) and the subsets of measurements (clusters) corresponding to similar daily profiles of the target variable. A forecasting model of each target variable is trained based on measurements of each cluster.

Sections V.1 to V.4 present the forecasting performance of the models for each target variable with the sequence A of pre-processing steps. The forecasting performance of the cascade-based method obtained with the sequence A of pre-processing steps is compared to the forecasting performance of the method when using sequences B and C in section V.6.

V.1 Forecasting of the supply air flow rate

This section presents the forecast of the supply air flow rate over the following six hours, i.e., over 24 time steps. An iterated strategy is applied for the problem of multi-step-ahead forecasting of the target variable (see section III.1.2).

V.1.1 Support Vector Regression

From the literature review we concluded that the support vector machines and neural networks are among the most performant inverse models for the forecast of electric demand [24, 25, 108]. So far the neural networks have been widely used, while the application of Support Vector Regression (SVR) models to HVAC-related studies still require more investigation. For this reason, the SVR models are used in this study.

Support Vector Machines (SVM) are supervised learning models used for classification or regression analyses [57]. In regression analysis, the SVM method approximates an unknown function (f) by mapping, with a nonlinear function (ϕ), the observations of the training set ($\boldsymbol{x}$) into a higher dimension feature space. A linear regression problem is then solved in this feature space.

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \phi(\boldsymbol{x}) \rangle + b \qquad \text{[V.1]}$$

Where, $\boldsymbol{w}$ is the matrix of regression coefficients, b is the intercept term, $\boldsymbol{x}$ is the matrix of regressors and $\langle .,. \rangle$ is the dot product.

An optimization problem was formulated by Vapnik [57] (Equation [V.2]) to identify the function f, which is expressed by the regression coefficients ($\boldsymbol{w}$) and intercept (b), that predicts all the observations of the target vector ($\boldsymbol{y}$) with a precision ϵ.

$$\begin{aligned} & \min_{w,b,\xi,\xi^*} \frac{1}{2}\|\boldsymbol{w}\|^2 + C.\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ \text{Subject to:} \quad & \langle \boldsymbol{w}, \phi(\boldsymbol{x_i}) \rangle + b - y_i \leq \epsilon + \xi_i, \\ & y_i - \langle \boldsymbol{w}, \phi(\boldsymbol{x_i}) \rangle - b \leq \epsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0, i = 1, \dots, l. \end{aligned} \qquad \text{[V.2]}$$

Where, y_i is an observation of the target vector, ξ is a slack variable, and ϵ and C are parameters to be set by the user.

The regression function is reformulated in Equation [V.3] and the regression coefficients ($\boldsymbol{w}$) are given in equation [V.4].

$$f(\boldsymbol{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\, k(x_i, x) + b \qquad \text{[V.3]}$$

$$\boldsymbol{w} = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\, x_i \qquad \text{[V.4]}$$

Where, α_i and α_i^* are the Lagrange multipliers, and k is the kernel function.

The user has to specify the SVM parameters (ϵ and C) and select the kernel function (k). The type of kernel function is selected based on preliminary knowledge of the domain of application [109]. A Radial Basis Function (RBF) is selected since most of the regressors are measured variables presenting a Gaussian distribution (Equation [V.5]). Radial basis functions are used in most studies as kernel functions.

$$k(x_i, x_j) = exp\left(-\gamma . \left\|x_i - x_j\right\|^2\right) \quad [V.5]$$

Where, γ is the "width" parameter that reflects the range of variation of all the regressors in the training set.

Finally the user has to specify the SVM parameters (ϵ and C) and select the kernel function (k). In this study, the SVR model is implemented on Matlab using the LIBSVM library [87]. More details and explanations about the use of support vector machines for regression are given in [110].

V.1.2 Results of multi-step-ahead forecasting of the supply air flow rate

Two forecasting models (one for each cluster) are developed for the supply air flow rate using the regressors selected with sequence A. The regressors are the past values of the supply air flow rate and the derived variable labelled "occupation". The variable "occupation" corresponds to the schedule of operation of the supply fans; it takes a value of 1 when operated and -1 otherwise.

The SVR model maps the value of the supply air flow rate at the next time step (t+15 min) with the current and previous values of the target variable and regressors. The values of three parameters (ϵ, C, γ) are required to be set before training the SVR model. For this purpose, γ and ϵ are varied from 2^{-12} to 2, and C from 2^{-5} to 2^5. For each combination of the three parameters, the SVR model is trained over ten days and tested over the following six hours. The training and test are repeated over ten different samples of the summer dataset. The ten samples have the same training and test lengths, however they differ only by the random starting time and date. The best combination of the three parameters that gives the minimum averaged root-mean-square error (RMSE) of the target variable over the ten testing sets is selected.

The best combination of the three parameters ($\epsilon = 2^{-11}$, $C = 2^{-3}$ and $\gamma = 2$) is obtained when the averaged RMSE of the difference between the forecasts and measurements of supply air flow rate has the minimum value of 562.9 L/s.

The estimation of the optimum set of parameters for the SVR forecasting model of the supply air flow rate lasted 21 min on an Intel Xeon processor E5645 with six cores computing at 2.4 GHz. The averaged RMSE between the forecasted and measured values of supply air flow rate over the ten samples is presented on Figure V.1, for each combination of the three parameters. The best

combination is highlighted by the red squares. The six best combinations of parameters are presented in Table V.1. They show the six smallest values of RMSE of the supply air flow rate varying from 562.9 to 569.8 L/s.

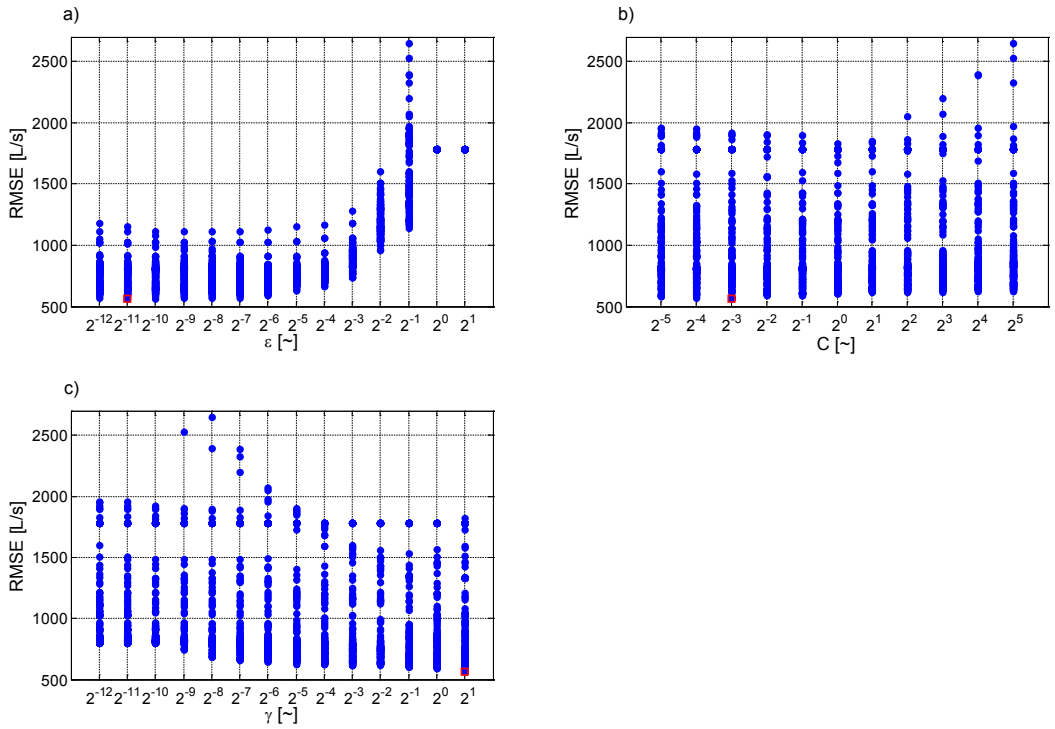

Figure V.1 RMSE of the difference between the forecasts and measurements of the supply air flow rate over the test data set vs combination of parameters: ϵ, C and γ.

Table V.1 Best combination of parameters for the SVR model and kernel function of the supply air flow rate.

Rank	ϵ [-]	C [-]	γ [-]	RMSE [L/s]
1	2^{-11}	2^{-3}	2	**562.9**
2	2^{-10}	2^{-3}	2	**564.3**
3	2^{-12}	2^{-3}	2	**567.4**
4	2^{-7}	2^{-3}	2	**568.9**
5	2^{-9}	2^{-3}	2	**569.0**
6	2^{-8}	2^{-4}	2	**569.8**

Once the parameters (ϵ, C, γ) are set, the SVR model is trained on a dataset covering a period from ten days to two weeks depending on the data available for the supply air flow rate. The regression coefficients of the SVR model are tuned to minimize the prediction error between the measured and forecasted electric demand over the training set. Once the model is trained, it is used to forecast the supply air flow rate over the following six hours.

An example of the forecasting performance of the SVR model is presented on Figure V.2. The SVR forecasting model is trained over ten working days of measurements as training set (from the 3rd to the 18th of June 2014, included) with the best combination of parameters (Table V.1). The forecasting model is tested on the following six hours from 05:00 to 11:00 on 19th of June (Figure V.2). The period selected to test the forecasting model is the beginning of the day. It corresponds to the start-up of the HVAC system adapting to the change in the occupation of the building. The measured supply air flow rate is displayed in red and the forecast in blue on the top part of Figure V.2. The bottom chart presents the forecasting error or residual. On this example, the forecasting model presents a RMSE of 1,053 L/s over the six hours corresponding to CV(RMSE) of 5.4%.

The performance of the forecasting models is presented in terms of RMSE between the forecasted and recorded values of the target variables over the six-hour test. The Coefficient of Variation of the RMSE (CV(RMSE)) is presented as well (Equations [V.6] and [V.7]).

$$RMSE = \sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/n} \quad [V.6]$$

$$CV(RMSE) = RMSE/\bar{y} \quad [V.7]$$

where $\hat{y}_i$ is the forecasted observation i and y_i is the measured one; $\bar{y}$ is the average value of the variable y over the selected period.

The forecasts are displayed with error bounds representing the interval of 95% confidence of the prediction. The error bound is estimated based on the assumption that the forecasting error follows a Laplace distribution with a null mean and a scale parameter (b) is extracted from a five-fold cross-validation on the training dataset [87]. For a 95% confidence interval the prediction error of the supply air flow rate varies between $\pm$3,000 L/s.

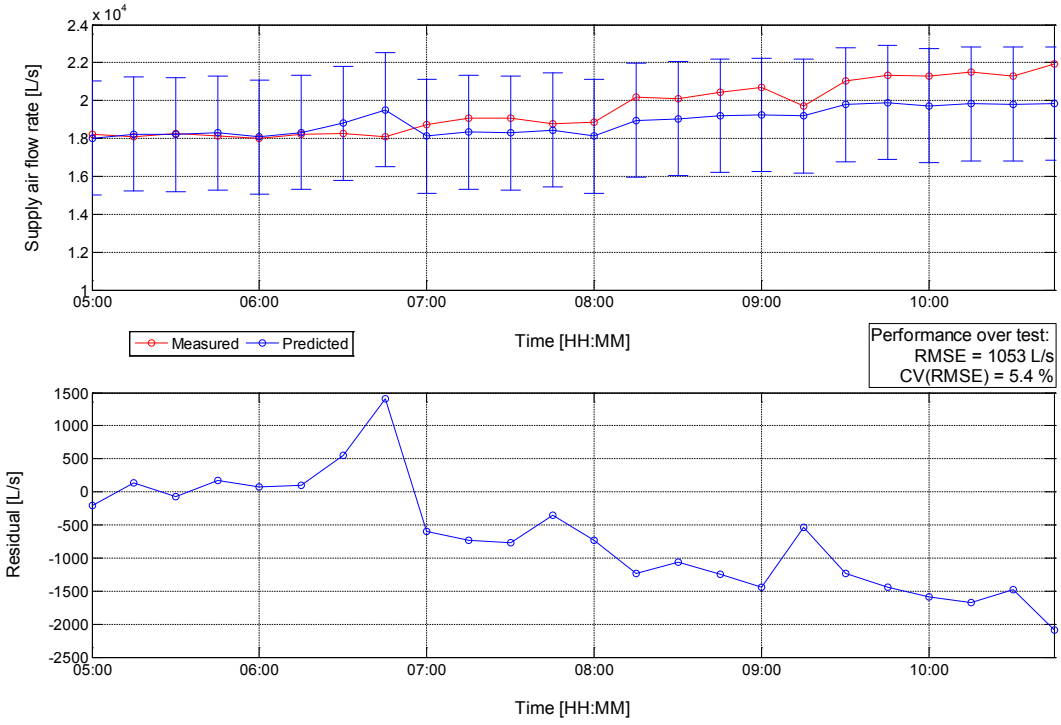

Figure V.2 Forecasted supply air flow rate over the six-hour test set on 19th of June.

V.2 Forecasting of the cooling coil load

The same approach as for the supply air flow rate is used to forecast the cooling coil load of the two air handling units based on the outputs of the sequence A of preprocessing steps. An SVR forecasting model is developed for each one of the three clusters corresponding to typical daily profiles of the cooling coil load identified in the preprocessing step (Section IV.3.3). The SVR model maps the value of the cooling coil load at the next time step (t+15 min) with the current and previous values of the target variable and the regressors identified in the pre-processing steps (Table IV.19: outdoor air enthalpy and mixed air temperature in the two AHUs). The forecast of the cooling coil load at time t+30 min depends on the value of the regressors at time t+15 min. Three other SVR models are also developed to forecast the future value of each regressor based only on its current and past values (See Chapter III).

The best combination of the three parameters (ϵ, C and γ) of the SVR model and kernel function are listed in Table V.2, for each SVR model of each cluster of the cooling coil load with sequence A of preprocessing steps.

The forecasted profiles of the cooling coil load and of the regressors for each cluster are presented respectively in Appendix A and Appendix B in blue against the measurements in red over the same six-hour test period. Their forecasting performance is presented in Table V.4 in terms of RMSE and CV(RMSE) of the measured values against forecasted over the training and testing sets.

Cluster #2 presents the best performance, in this example, for the cooling coil load (RMSE of 36 kW and CV(RMSE) of 19.3%) along with its regressors. By comparing Table V.3 and Table V.4, it is noticed that the forecasting performance of the cooling coil load appears to be more sensible to the mixed air temperature in both AHUs than the outdoor air enthalpy. Indeed the cooling coil load is calculated based on the mixed air enthalpy which depends on the outdoor and return air enthalpy. This forecasting performance of the model on cluster #2 is also drawn by the selected parameters: RMSE of 23.6 kW on cluster #2 compared to 55.6 and 35.5 kW for clusters #1 and 3 (Table V.2).

Table V.2 Best combination of parameters for the SVR model and kernel function of the cooling coil load for each cluster with sequence A.

	ϵ [-]	C [-]	γ [-]	RMSE [kW]
Cluster #1	2^{-3}	2^{5}	2^{-12}	**55.6**
Cluster #2	2^{-5}	2^{-3}	2^{0}	**23.6**
Cluster #3	2^{-4}	2^{2}	2^{-6}	**35.5**

Table V.3 Forecasting performance of the cooling coil load over the six-hour test set for clusters #1 to 3 with sequence A.

	Cluster #1		**Cluster #2**		**Cluster #3**	
	RMSE [kW]	CV(RMSE) [%]	RMSE [kW]	CV(RMSE) [%]	RMSE [kW]	CV(RMSE) [%]
Cooling coil load over training set	52	12.3	26	19.3	31	10.4
Cooling coil load over testing set	76	22.5	36	19.3	45	15.9

Table V.4 Forecasting performance of the regressors over the six-hour test set for clusters #1 to 3 with sequence A.

	Cluster #1		**Cluster #2**		**Cluster #3**	
	RMSE [-]	CV(RMSE) [%]	RMSE [-]	CV(RMSE) [%]	RMSE [-]	CV(RMSE) [%]
Outdoor air enthalpy [kJ/kg]	1.8	3.6	2.5	6.7	1.0	2.2
Mixed air temperature in AHU 2 [C]	1.7	8.0	0.7	3.6	4.3	20.1
Mixed air temperature in AHU 1 [C]	1.4	6.7	0.8	4.3	4.6	21.2

V.3 Forecasting of the cooling load of Genome building

The cooling load of Genome building presents four typical daily profiles corresponding to four clusters (section IV.4.3). Cluster #1 gathers only six daily profiles, which is not enough to train and test a forecasting model. Three SVR forecasting models are developed based on measurements in clusters #2 to 4 of the cooling load of Genome building. According to the sequence A of preprocessing step (Table IV.19), the future value of the cooling load depends on its current and previous values as well as five other regressors (the cooling coil load, the mixed air temperature and modulation of the cooling coil valve in both AHUs). Four autoregressive forecasting models are also developed for the mixed air temperature and the cooling coil valve modulation in both AHUs. The forecasted value of the cooling coil load obtained from the previous sub-section is used as input as suggested in the proposed cascaded-based forecasting method.

The best combination of the three parameters (ϵ, C and γ) of the SVR model and kernel function are listed in Table V.5, for each SVR model of each cluster of the building cooling load with sequence A of preprocessing steps.

Table V.5 Best combination of parameters for the SVR model and kernel function of the cooling load for each cluster.

	ϵ [-]	C [-]	γ [-]	RMSE [kW]
Cluster #2	2^{-5}	2^{5}	2^{-4}	**25.5**
Cluster #3	2^{-3}	2^{-5}	2^{-2}	**20.5**
Cluster #4	2^{-12}	2^{3}	2^{-4}	**28.5**

An example of the forecasted profiles of the cooling load of the whole building for each cluster are presented in Appendix C in blue against the measurements in red. Their forecasting performance is presented in Table V.6 in terms of RMSE and CV(RMSE) over the training and testing sets.

Four autoregressive SVR models were also developed to forecast the regressors based on their current and past values (mixed air temperature and cooling coil valve modulation in both AHUs). Their performance is presented in Table V.7 in terms of RMSE and CV(RMSE) of the measurements against forecasts. The forecasted profiles of the regressors over the same time period from 05:00 to 11:00 are presented in Appendix D.

The SVR models present good forecasting performances with a CV(RMSE) of 10.7% and 12.5% for clusters #2 and 4 on these examples. In cluster #3, the forecasting model presents some difficulties to handle the sudden increase in the cooling load occurring at 09:45 (Figure C.2 in Appendix C). This is mainly because the increase happens at the 20th time step of the horizon prediction and the increase is significant (about 200 kW). Clusters #2 and 4 do not present sharp increase of the building cooling load like in cluster #3 (see the typical daily profiles in Figure IV.22). The accuracy of the forecast of the building cooling load in the example of cluster #3 is also impacted by the poor forecasts of the regressors compared to clusters #2 and 4. The cooling coil load and cooling coil valves modulation present RMSEs of 69.4 kW and both 37%.

Table V.6 Forecasting performance of the building cooling load over the six-hour test set for clusters #2 to 4.

	Cluster #2		**Cluster #3**		**Cluster #4**	
	RMSE [kW]	CV(RMSE) [%]	RMSE [kW]	CV(RMSE) [%]	RMSE [kW]	CV(RMSE) [%]
Building cooling load over training	28	6.3	52	23.4	26	7.5
Building cooling load over testing	48	10.7	82	59.8	38	12.5

Table V.7 Forecasting performance of the regressors over the six-hour test set for clusters #1 to 3.

	Cluster #2		**Cluster #3**		**Cluster #4**	
	RMSE [-]	CV(RMSE) [%]	RMSE [-]	CV(RMSE) [%]	RMSE [-]	CV(RMSE) [%]
Cooling coil load [kW]	25.1	8.9	69.4	476.1	42.4	23.0
Mixed air temperature in AHU2 [C]	0.5	2.2	0.4	2.6	1.4	8.0
Mixed air temperature in AHU1 [C]	0.8	3.6	1.1	6.7	1.3	7.5
Cooling coil valve modulation in AHU1 [%]	3.9	7.3	37.0	191	3.1	6.7
Cooling coil valve modulation in AHU2 [%]	4.0	7.3	37.0	191	3.1	6.7

V.4 Forecasting of the electric demand of the secondary and primary cooling systems

This subsection presents the forecasting of the electric demand of secondary systems of the Genome building, as well as the total electric demand of the primary system of the cooling plant.

V.4.1 Forecasting of the electric demand of the supply fan

The electric demand of the Variable Air Volume (VAV) fans in the AHUs ($\dot{E}_{sec.syst.}^{t+15}$) is estimated using Equation [V.8], where Y is the ratio of the electric demand at part-load to the maximum measured electric demand $\dot{E}_{max}$ over the period considered (equation [V.9]):

$$\dot{E}_{sec.syst.}^{t+15} = \dot{E}_{max} * Y \quad [V.8]$$

$$Y = a_0 + a_1.X + a_2.X^2 + a_3.X^3 \quad [V.9]$$

Where, a_0, a_1, a_2, a_3 are the regression coefficients and X is the part-load ratio of the supply air flow rate: $X = \dot{V}_{measured}/\dot{V}_{max}$. The regression coefficients used in equation [V.9] were identified from spot measurements (Table V.8), and compared with two reference data (ASHRAE [111] and EnergyPlus [112]).

Table V.8 Regression coefficients of the fans performance curves

Model/Coefficients	a_0	a_1	a_2	a_3	R^2
ASHRAE	0.0013	0.1470	0.9506	0.0998	~
Energy Plus	0.0015302446	0.0052080574	1.1086242	-0.11635563	~
Measurements of fan #1	0.0000	0.2530	0.0000	0.7470	0.51
Measurements of fan #2	0.0000	0.1361	0.2006	0.5534	0.61
Measurements of fan #3	0.0000	0.0012	0.1832	0.8156	0.75
Measurements of the 3 fans	0.0000	0.1268	0.0000	0.8732	0.57

The electric demand of three supply fans in the AHUs of the Genome building was measured over two periods of three weeks [104]; from the November 15th to December 5th 2013, and from February 27th to March 20th 2014. The measurements of the supply air flow rate over the same period was gathered through the BAS of the Genome building.

The electric demand and supply air flow rate are normalized according to their maximum values over the set of measurements. A polynomial model of order three is fitted to the data; the regression

coefficients are presented in Table V.8. The intercept a_0 was constrained to a zero value to have a null electric demand when the supply air flow rate is null. The model fits to the measurements of the three fans with a coefficient of determination of 0.57 (Table V.8). The predicted value is displayed in black squares on Figure V.3. The variation of the normalized electric demand is between 0.01 and 0.3 for a normalized supply air flow rate varying between 0.3 and 0.5. Different levels of electric demand are observed at part load for each fan from blue to red in ascending order.

The normalized electric demand of the fan is also estimated using the polynomial equation with regression coefficients given by ASHRAE standard 90.1 [111] and EnergyPlus [112]. They are presented respectively by the blue and green lines on Figure V.3. The points correspond to the normalized values of the measurements of electric demand. The measurements of the three fans are plot in red, blue and green colors. The points corresponding to the fan affinity law, which is without the electric motor efficiency, are also presented, as a black line, for information.

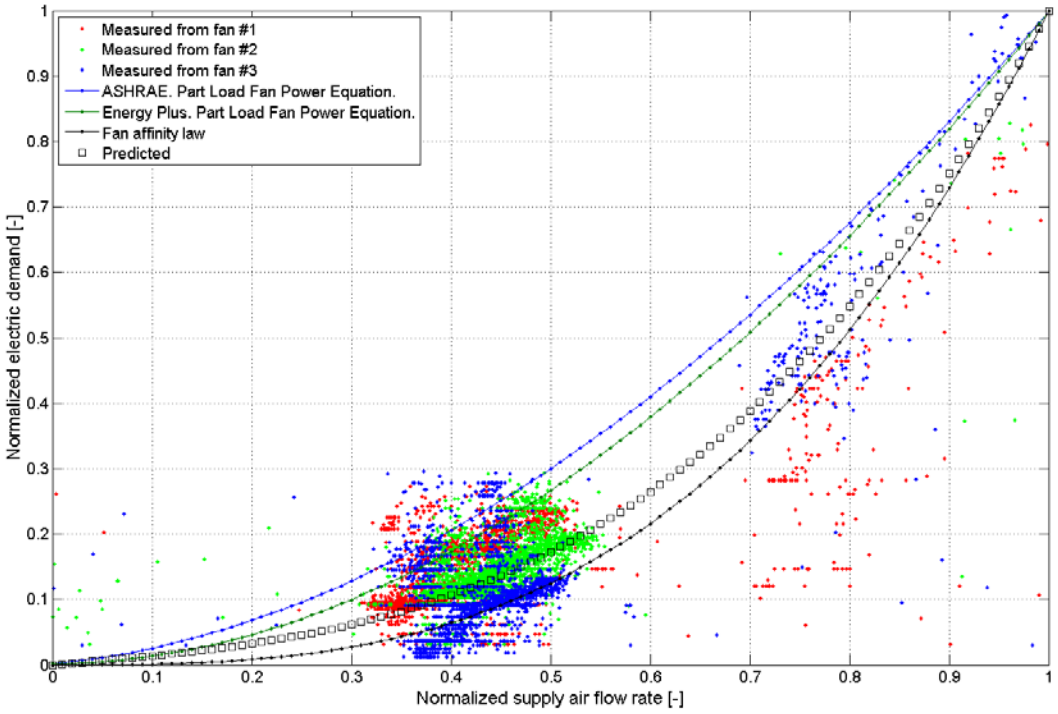

Figure V.3 Average performance curve of the three fans

The future electric demand of the fans is estimated using the polynomial regression model of order three (Equations [V.8] and [V.9]) with coefficients fitted on the three fans (Table V.8). The future value of the supply air flow rate forecasted in section V.1 is used as an input to Equation [V.9].

V.4.2 Forecasting of the electric demand of the chillers

The electric demand of the chillers was partitioned into four clusters of similar daily profiles in the preprocessing step (Section IV.5.2.3). Four SVR forecasting models are developed based on the observations included in clusters #1 to 4. The electric demand of the chillers depends on its current and past values as well as four other regressors. The four regressors selected in sequence A of pre-processing steps are the cooling load and chilled water flow rate supplied by the central plant, the cooling load of the building and the number of chillers in operation. An SVR forecasting model is developed for each regressor based only on its current and previous values, except for the cooling load of the building which was forecasted beforehand in the cascade-based method.

The best combination of the three parameters (ϵ, C and γ) of the SVR model and kernel function are listed in Table V.9, for clusters #1 to 4 of the electric demand of chillers with sequence A of preprocessing steps.

Table V.9 Best combination of parameters for the SVR model and kernel function of the electric demand of the chillers for each cluster.

	ϵ [-]	C [-]	γ [-]	RMSE [kW]
Cluster #1	2^{-4}	2^{5}	2^{-8}	**15.5**
Cluster #2	2^{-5}	2^{4}	2^{-8}	**16.2**
Cluster #3	2^{-10}	2^{4}	2^{-7}	**8.0**
Cluster #4	2^{-6}	2^{5}	2^{-11}	**10.5**

Benchmark values of the electric demand of the chillers are calculated using the values of the ten-day training set. The upper (U_{b,t_i}) and lower (U_{l,t_i}) bounds of the electric demand are calculated at each time step (t_i), using the average value at that time step ($\mu_{\dot{E},t_i}$) plus or minus the standard deviation ($\sigma_{\dot{E},t_i}$) of the electric demand ($\dot{E}$) (Equation [V.10]).

$$U_{b,t_i} = \mu_{\dot{E},t_i} + \sigma_{\dot{E},t_i}$$
$$L_{b,t_i} = \mu_{\dot{E},t_i} - \sigma_{\dot{E},t_i} \qquad [V.10]$$

An example of the forecasted profiles of the electric demand of chillers is given in Appendix F for cluster #1 to #4. The measured values of the electric demand of the chillers is displayed in red on the top part of the graph and the forecasted ones in blue. The upper and lower benchmarks, calculated using the ten-day training data set, are displayed in green. The forecasting performance

of the four SVR models for the electric demand of the chillers are presented in Table V.10 over the training and testing sets.

The forecasts of the four regressors (chilled water flow rate and cooling load supplied by the central plant, number of chillers in operation and cooling load of the Genome building) used as inputs after the first time step of the prediction horizon are presented in Appendix G. Their forecasting performance over these examples are presented in Table V.11 in terms of RMSE and CV(RMSE).

The SVR models present a good forecasting performance on these two examples for the electric demand of the chillers (Table V.10). The two autoregressive models developed for the chilled water flow rate and cooling load supplied by the central plant show a good performance as well over each cluster (Table V.11).

However a closer look at Figure F. in Appendix F shows that the sudden increase of the electric demand at 9:45 is not forecasted by the model. This is increase is also noticed in the measured profile of the chilled water flow rate leaving the central plant and the number of chillers in operation passing from 1 to 2 (Figure G.). The autoregressive models for the forecast of regressors did not predict this increase in terms of flow rate and number of working chillers.

In cluster #4, the cooling load of the Genome building, on the 11th of July, is forecasted in the previous step of the cascade-based method. The good forecasting performance is representative of the one previously presented for cluster #4 of the building cooling load (Table V.6).

The fourth regressor, corresponding to the number of chillers in operation, shows a poor accuracy. This regressor is a categorical variable taking three different values: 0, 1 and 2. The forecast of the number of chillers in operation could be improved by using a classification algorithm instead of regression.

Table V.10 Forecasting performance of the electric demand of the chillers over the six-hour test set for clusters #1 and 4.

	Cluster #1		**Cluster #2**		**Cluster #3**		**Cluster #4**	
	RMSE [kW]	CV(RMSE) [%]	RMSE [kW]	CV(RMSE) [%]	RMSE [kW]	CV(RMSE) [%]	RMSE [kW]	CV(RMSE) [%]
Electric demand of the chillers over training	26	5.3	19	3.0	25	14.4	14	4.9
Electric demand of the chillers over testing	67	15.4	68	13.3	89	48.0	24	9.5

Table V.11 Forecasting performance of the regressors over the six-hour test set for clusters #1 and 4.

	Cluster #1		Cluster #2		Cluster #3		Cluster #4	
	RMSE [-]	CV(RMSE) [%]	RMSE [-]	CV(RMSE) [%]	RMSE [-]	CV(RMSE) [%]	RMSE [-]	CV(RMSE) [%]
Chilled water flow rate [L/s]	20.2	20.0	3.1	2.0	26.2	48.9	1.2	1.3
Cooling load supplied by RF [kW]	419.3	17.2	835.4	31.8	451.6	61.1	157.8	12.3
Cooling load of GE [kW]	62.1	15.0	20.0	4.4	73.9	33.1	41.3	11.7
Number of chillers in operation [-]	0.3	25.3	10^{-4}	0.005	0.4	26.6	0.5	50.0

V.4.3 Forecasting of the electric demand of the cooling towers

Four typical daily profiles of the electric demand of the cooling towers were identified in the preprocessing step. Two clusters gathering four and seven daily profiles are discarded for the forecasting analysis of the electric demand of the cooling towers. Cluster #1 and 2 are remaining.

The future electric demand of the cooling towers depends on its current and past values and the cooling load supplied by the central plant, based on sequence A of pre-processing steps. An autoregressive SVR model is developed for the cooling load supplied by the central plant.

The best combination of the three parameters (ϵ, C and γ) of the SVR model and kernel function are listed in Table V.12Table V.9, for clusters #1 and 2 of the electric demand of cooling towers with sequence A of preprocessing steps.

Table V.12 Best combination of parameters for the SVR model and kernel function of the electric demand of the cooling towers for each cluster.

	ϵ [-]	C [-]	γ [-]	RMSE [kW]
Cluster #1	2^{-3}	2^{0}	2^{-2}	**0.9**
Cluster #2	2^{-4}	2^{-5}	2^{-4}	**0.4**

The forecasting performance of the SVR models for the electric demand of the cooling towers are illustrated over two examples in Appendix H. The forecast are displayed in blue, the measured values in red and the benchmarks in green. Their forecasting performance over these examples are presented in Table V.13 in terms of RMSE and CV(RMSE).

The forecasting performance of the autoregressive model developed for the cooling load supplied by the central plant is presented in Table V.14. The forecasted profiles of the cooling load for clusters #1 and 2 are displayed in Appendix H.

Table V.13 Forecasting performance of the electric demand of the cooling towers over the six-hour test set for clusters #1 and 2.

	Cluster #1		**Cluster #2**	
	RMSE [kW]	CV(RMSE) [%]	RMSE [kW]	CV(RMSE) [%]
Electric demand of the cooling towers over training	1.0	17.6	0.7	47.4
Electric demand of the cooling towers over testing	3.0	41.3	1.0	53.5

Table V.14 Forecasting performance of the regressors over the six-hour test set for clusters #1 and 2.

	Cluster #1		**Cluster #2**	
	RMSE [-]	CV(RMSE) [%]	RMSE [-]	CV(RMSE) [%]
Cooling load supplied by the central plant [-]	789.2	30.1	206.2	21.4

V.4.4 Forecasting of the ratio of cooling loads (α)

The cooling load of the Genome building ($\dot{Q}_{GE}$) is plotted against the cooling load of the central plant ($\dot{Q}_{RF}$) on Figure V.4 over the period of summer 2014 (from the 1st of June to the 1st of September). The color of the points refers to the percentage of opening of the cooling coil valve in the AHUs of the Genome building, from 30% in blue to 95% in red. The cooling load of the Genome building is linearly correlated to the cooling load of the central plant with a coefficient of determination $R^2 = 0.79$, for a cooling load of the central plant varying between 400 kW and 5,000 kW (Equation [V.11]).

$$\dot{Q}_{GE} = 0.12 * \dot{Q}_{RF} + 161.3 \qquad [V.11]$$

The ratio (α) of the cooling load of the Genome building ($\dot{Q}_{GE}$) to the cooling load of the central plant ($\dot{Q}_{RF}$) is identified from the linear regression model: it corresponds to the regression coefficient: $\alpha = 0.12$.

The ratio of cooling loads (α) presents a normal distribution with a mean of 12% and standard deviation of 3% (Figure V.5). Half of the calculated ratios are between 10 and 14%; the limits are

displayed by red dotted lines on Figure V.4. α is assumed constant to its mean value: $\mu(\alpha) = 12\%$ for a cooling load of the central plant between 400 kW and 5,000 kW.

The future electric demand of the primary cooling system corresponding to the requirements of the Genome building ($\dot{E}^{t+15}_{prim.syst.}$) is estimated as a constant multiplied by the future electric demand of the central plant ($\dot{E}^{t+15}_{cooling,RF}$) as presented in Equation [V.12]. The future electric demand of the central plant includes the forecasts of the electric demand of the chillers and cooling towers presented in sub-sections V.4.2 and V.4.3.

$$\dot{E}^{t+15}_{prim.syst.} = 0.12 * \dot{E}^{t+15}_{cooling,RF} \qquad [V.12]$$

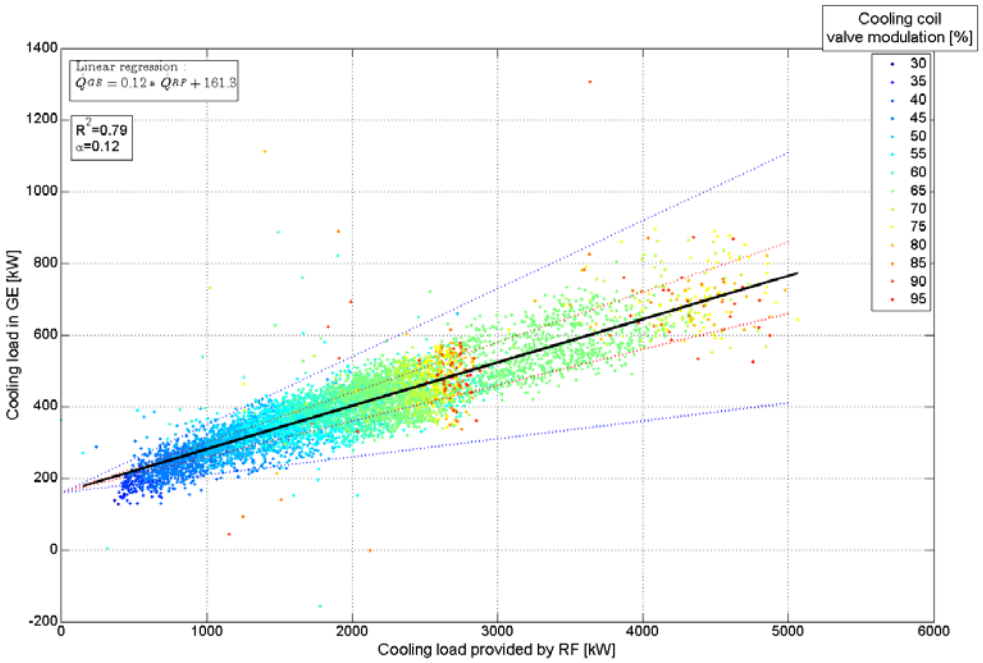

Figure V.4 Linear regression modelling of the ratio (α)

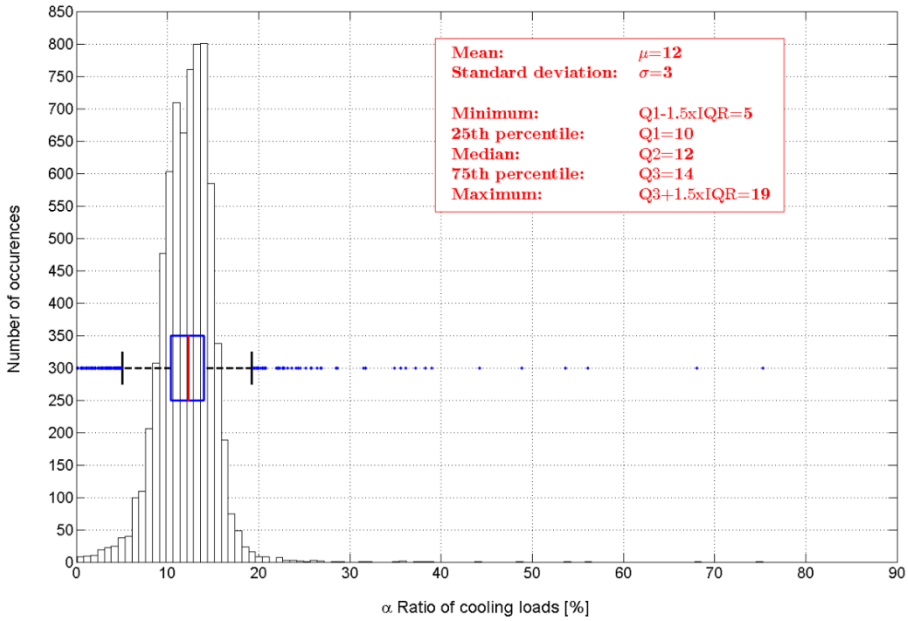

Figure V.5 Probability distribution of the ratio (α)

V.5 Forecasting of the electric demand of the Genome building with the cascade-based method

This section presents the forecasts of each target variable of the cascade-based method using data of Friday, the 30th of July, from 09:00 to 15:00. This test day belongs to cluster #1 of the supply air flow rate, and respectively clusters #2 for the cooling coil load, #3 for the building cooling load, #3 for the chillers and #2 for the cooling towers. The target variables are forecasted one after the other starting with the supply air flow rate from both AHUs, the cooling coil load and the whole building cooling load. The forecast of the first three target variables is displayed in Figure V.6. The forecasts are displayed with error bounds representing the interval of 95% confidence of the prediction. The error bound is estimated based on the assumption that the forecasting error follows a Laplace distribution with a null mean identified the training error: the scale parameter (b) is extracted from the training dataset. The forecast of supply air flow rate with the sequence A has a RMSE of 724 L/s and CV(RMSE) of 3.4%, of the cooling coil load has a RMSE of 88 kW and CV(RMSE) of 23.4%; and of the whole building cooling load has a RMSE of 79 kW and CV(RMSE) of 22.0% (Table V.15).

The electric demand of the secondary and primary cooling system are the last target variables to forecast in this cascade-type method. The electric demand of the supply fans in both AHUs is

estimated using equations [V.8] and [V.9]. The forecast of electric demand of supply fans fits well the measurements (Figure V.7), with a RMSE of 4 kW and CV(RMSE) of 7.3%. The upper and lower values of the benchmarks are displayed as well on Figure V.7 in green. They are calculated based on the values of the training dataset (see Chapter III) including the ten days preceding the test set in the related cluster. These benchmark values reflect the "normal" conditions of operation in the previous days.

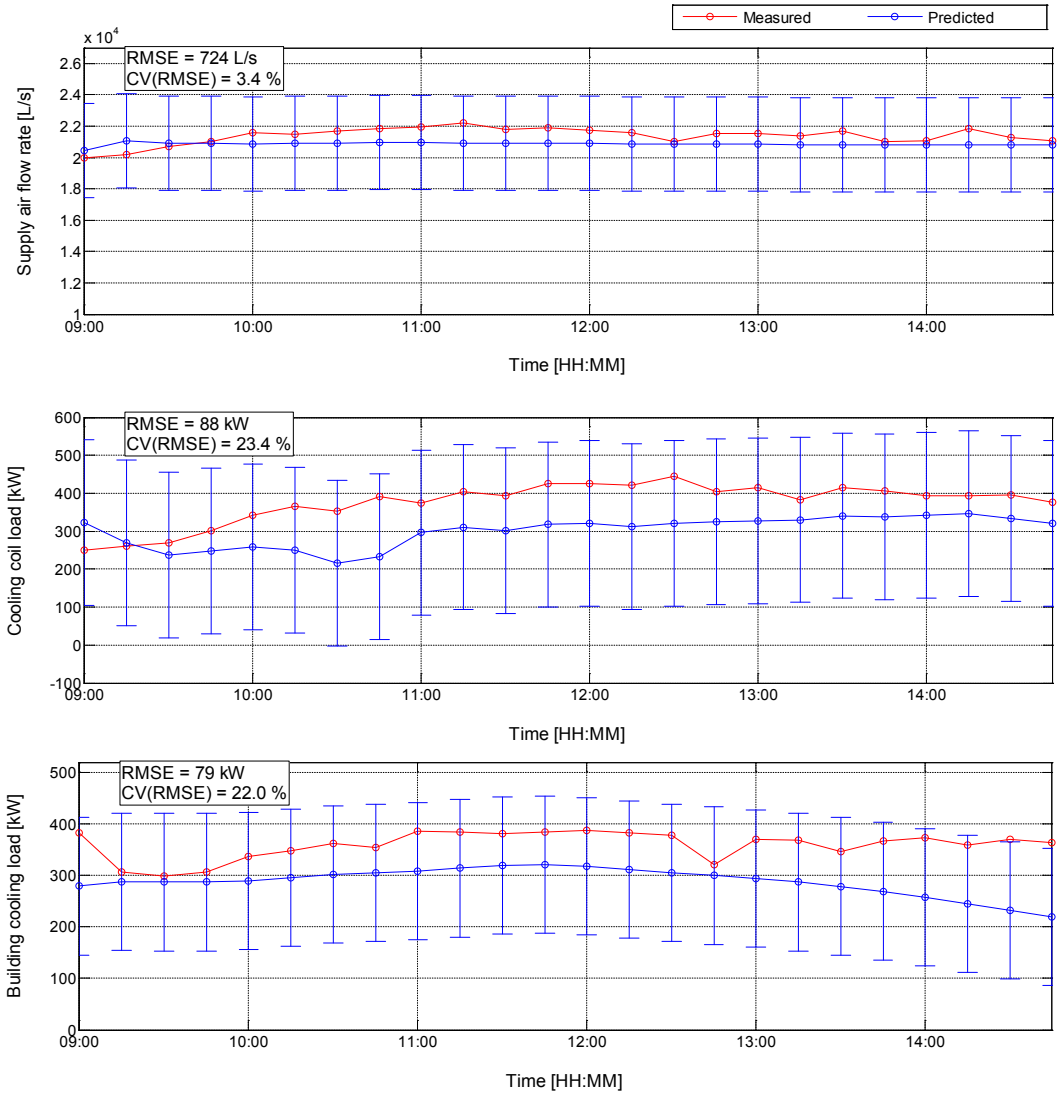

Figure V.6 Forecasting of the first three target variables on Friday, the 30th of July 2014.

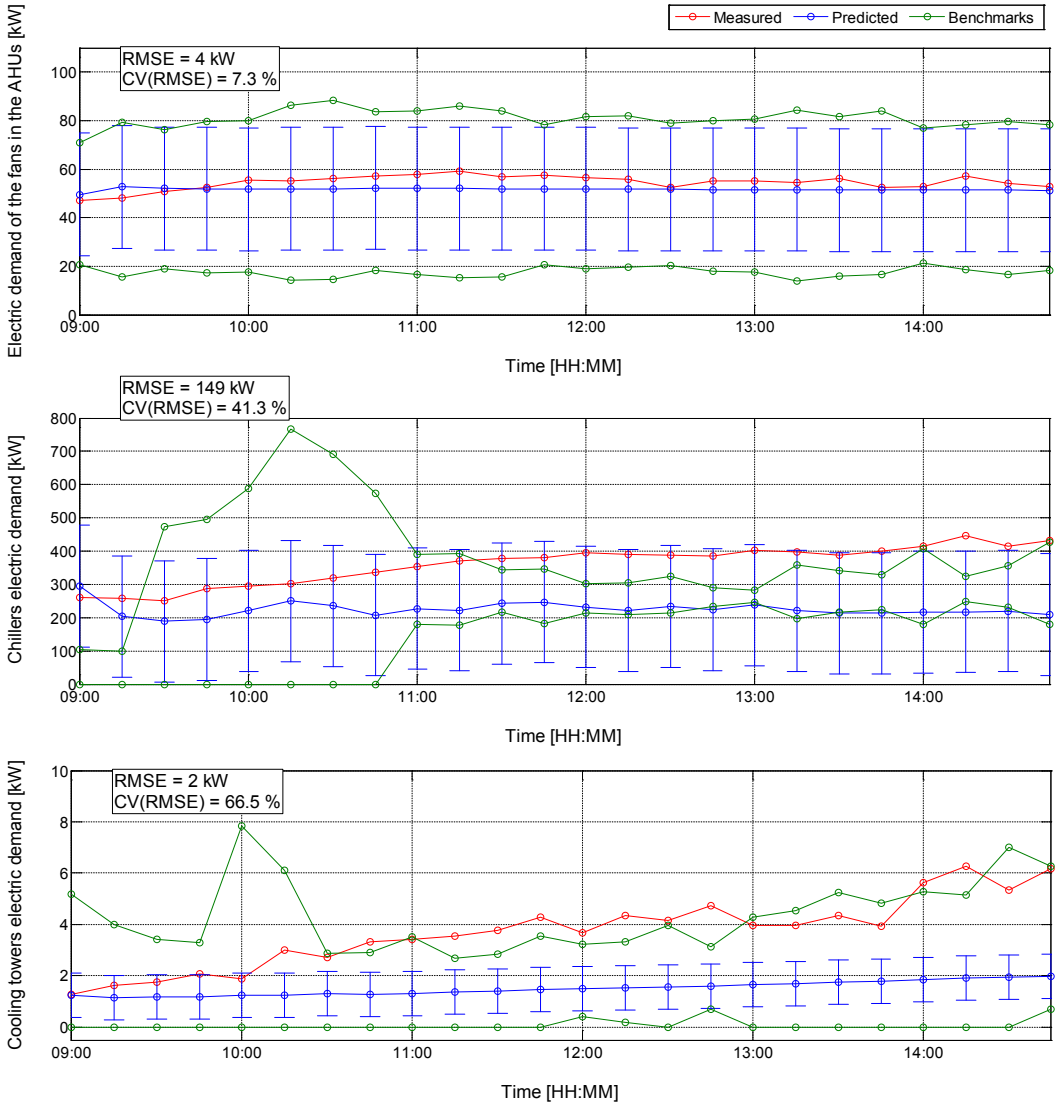

Figure V.7 Forecasting of the electric demand of secondary systems of the GE building and total primary cooling systems of the central plant ($\dot{E}^{t+15}_{cooling,RF}$) on Friday, the 30th of July 2014.

The ratio (α) (equation [V.11]) is identified from the BAS trend data as a constant value of 0.12 over the summer 2014. The total electric demand of the chillers of the central plant is forecasted (Figure V.7) with a RMSE of 149 kW and CV(RMSE) of 41.3%, on that day taken as an example. The total electric demand of the fans in the cooling towers of the central plant is forecasted with a RMSE of 2 kW and CV(RMSE) of 66.5% over the same test period.

The forecast of the total electric demand of the cooling system of the GE building over the next six hours is displayed in Figure V.8. For instance at 10:00 (i.e., one hour ahead), the forecasted electric demand of the supply fans accounts for 66% of total electric demand, the contribution of

GE building to the chillers electric demand is about 35% of the total electric demand, the contribution of cooling towers is negligible (0.2%).

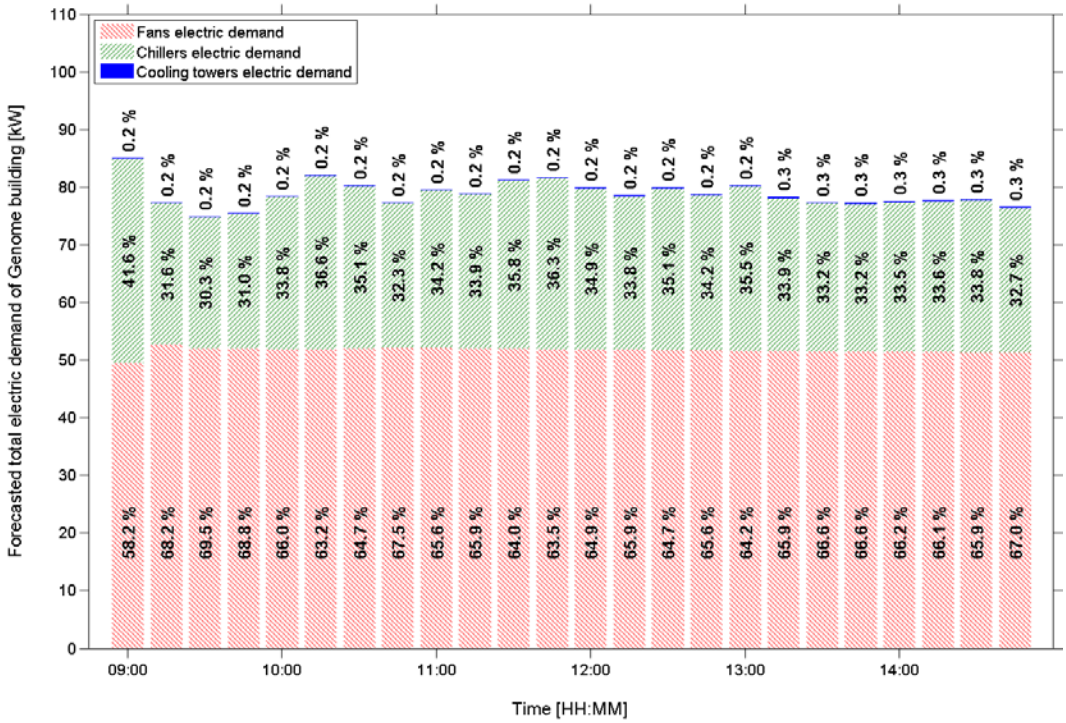

Figure V.8 Contribution of each HVAC system in the forecasted electric demand of GE building on Friday, the 30th of July 2014.

V.6 Comparison of three different sequences of pre-processing steps

The impact of the different sequences of pre-processing steps on the forecasting performance of the cascade-based method is compared for one day (30th of July 2014) in terms of RMSE and CV(RMSE) (Table V.15). The electric demand of the fans in the cooling towers is not presented as it is negligible with an average value of only 4 kW. The method has a good performance for forecasting the total electric demand of the HVAC system for GE building as the CV(RMSE) varies from 14.2% to 22.5% with the three sequences. The forecasted electric demand of the chillers (Table V.15) corresponds to the requirements of the GE building as opposed to the total electric demand of the central plant presented in Figure V.7.

For this particular case, all three pre-processing sequences give good forecasts for all target variables except the cooling coil load (sequence B) and the electric demand of chillers (sequence

A), if we accept the CV(RMSE) of 30% [113] as a benchmark value. We used as benchmark the statistical value proposed for the calibration of whole building energy use, even though it does not refer to the target variables of this study. Furthermore, the benchmark value of 30% applies to the hourly values, while this work uses 15-min time step values; hence, the acceptable CV(RMSE) should be higher.

We expected that the sequence B would give the best forecasts, as the target variables are forecasted for each cluster by using regressors that are specific to that cluster. However, for this case study the sequence C would be preferred because the target variable is forecasted using the regressors from the entire data set, without requiring the clustering analysis.

Table V.15 Comparison of the forecasting performance of the model over the 30th of July 2014 with the three sequences of pre-processing steps.

Description	**RMSE**			**CV(RMSE) [%]**		
Sequence	A	B	C	A	B	C
Supply air flow rate [L/s]	723.6	479.1	723.6	3.4	2.2	3.4
Cooling coil load [kW]	87.7	160.7	109.4	23.4	42.9	29.2
Building cooling load [kW]	79.2	46.4	39.8	22.0	12.9	11.1
Electric demand of fans [kW]	4.0	2.6	4.0	7.3	4.8	7.3
Electric demand of chillers for the GE building [kW]	17.8	11.5	9.8	5.0	3.2	2.7
Electric demand of the cooling system for the GE building [kW]	22.1	14.5	13.9	22.5	14.8	14.2

To obtain a more general comparison of the impact of the three pre-processing sequences on the forecasting performance, a separate analysis is performed, and the results presented in Table V.16. The forecasting model of each target variable is tested over 50 samples of summer 2014 under the same conditions (10-day training set and six-hour test). The starting time and date of the training set are chosen randomly. The overall forecasting performance is compared over the 50 samples in terms of RMSE and CV(RMSE) between the measured and predicted values of the target variable. Table V.16 presents the median value of the RMSE and CV(RMSE) of the forecasted target variable on the six-hour test over the 50 samples for each sequence of pre-processing steps. All three pre-processing sequences give good forecasts for all target variables except the cooling coil load, cluster #2 (sequences A and C), and the building cooling load, cluster #3 (sequences A and B).

Table V.16 Comparison of the forecasting performance of the model over the summer of 2014 with the three sequences of pre-processing steps.

		RMSE			CV(RMSE) [%]		
Sequence		**A**	**B**	**C**	**A**	**B**	**C**
Supply air flow rate [L/s]	Cluster #1	844	716	854	4.5	3.7	4.2
	Cluster #2	289	280		1.6	1.5	
Cooling coil load [kW]	Cluster #1	70	95	82	15.2	20.9	36.4
	Cluster #2	53	49		36.1	29.8	
	Cluster #3	48	50		18.1	16.5	
Building cooling load [kW]	Cluster #1	~	~	60	~	~	16.5
	Cluster #2	44	51		8.7	10.6	
	Cluster #3	78	93		32.3	50.3	
	Cluster #4	39	38		12.5	10.8	
Electric demand of fans [kW]	Cluster #1	4	3	4	8.8	7.7	8.8
	Cluster #2	1	1		3.2	3.1	
Total electric demand of chillers in the central plant [kW]	Cluster #1	51	78	75	10.2	15.8	22.9
	Cluster #2	51	57		9.3	9.7	
	Cluster #3	33	57		11.5	27.5	
	Cluster #4	48	35		13.9	11.9	

V.7 Evolution of the forecasting error after each iteration

At each time step, when new observations become available, the forecasting model is trained again, removing the oldest observations and adding the new ones. A sliding window is used to train the model.

An example of the forecast of the electric demand of the cooling system for the GE building with the updates of the measurements at each time step is presented in Figure V.9. The forecast is performed over three hours starting on the 30th of July at 9:00. The top figure shows the displacement of forecasts towards the measured electric demand as new observation becomes available. The bottom part of the graph shows the evolution of the forecasting error (residual) for each forecasted profile at each time step. The residual of the forecasts decreases at each iteration of the profile forecasted. For instance, the forecasting error of the electric demand of the cooling system thirty minutes ahead, that is at 9:30, is less than the measured demand by 6 kW with the first forecast based on the information known at 8:45 and before. The error is reduced to about -4

kW with the second forecast, when new observation came at 9:00, and to about 1.5 kW with the third forecast, when new observation came at 9:15.

The forecasting performance of the SVR model decreases along with the expansion of the prediction horizon. However the forecasting model can be retrained each time that new measurements are available. This gives a more confident forecast of the "near" future.

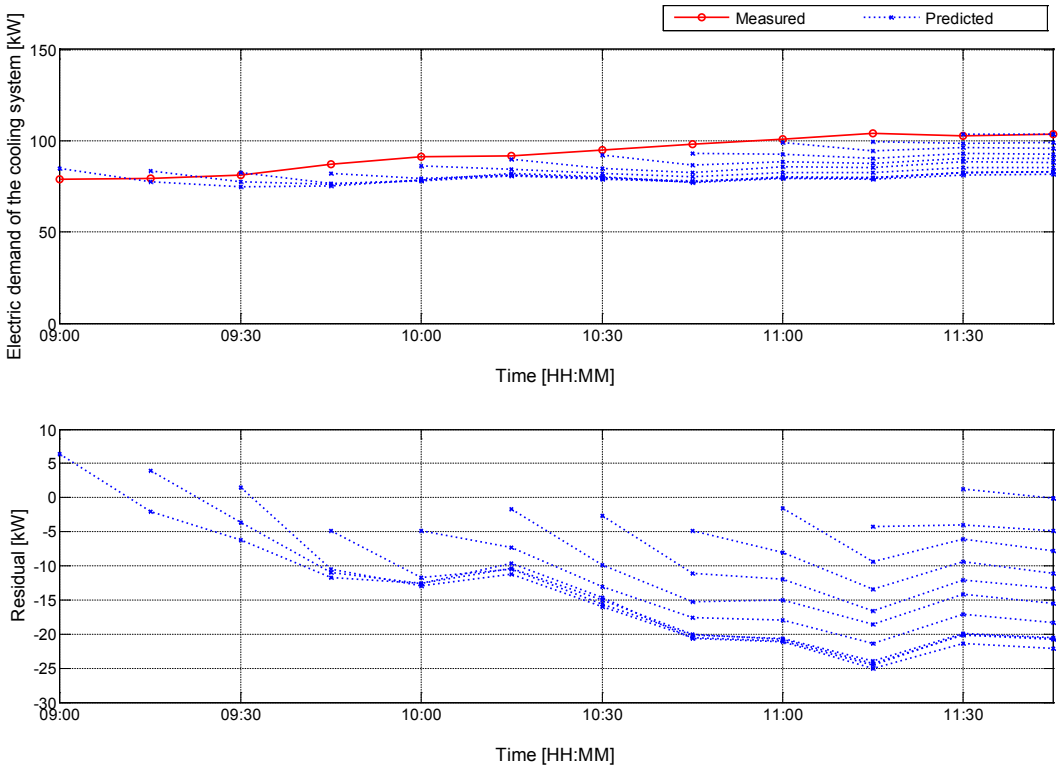

Figure V.9 Forecasting of the electric demand of the cooling system of the GE building with update of the measurements at each time step, on the 30th of July.

V.8 Conclusions

The cascade-based method forecasts successively six target variables of a HVAC system, i.e., the supply air flow rate, the cooling coil load, the whole building cooling load, the electric demand of the secondary and primary cooling systems (chillers and cooling towers). This multi-step forecasting allows to estimate the contribution of some selected HVAC operation variables on the total electric demand of the building. This is a useful feature to assess the impact of some demand response strategies. For instance, the building operator can change the input of mixing air temperature of AHUs (variables AHU1 T_MIX and AHU2 T_MIX), which is used as a regressor

for the forecasting of the cooling coil, and forecast the impact of such a change of the total electric demand of HVAC system.

The short-term forecast over a prediction horizon of six hours was tested by using BAS trend data from a case study research building over the summer of 2014. The results for the 30th of July 2014 show a good forecasting performance.

Although the results from all three pre-processing sequences are good, sequences A and B present slightly better forecasting performance than C over some clusters. The sequence C offers however the advantage of not requiring the partitioning of the data set into typical daily profiles. Hence it can be applied with a smaller data set of measurements of the summer season. However, this conclusion should interpreted with caution as it comes from one single case study. Several similar studies should be undertaken on different HVAC systems to generalize this conclusion.

Chapter VI. Component-based forecasting method: case-study of an office building

This chapter presents the development of forecasting models by following the component-based method. The component-based method uses the same sequence B of pre-processing steps as in the cascade-based method including: (1) exploratory analysis, (2) identification of the typical daily profiles and (3) selection of the regressors in each cluster (for each subset of typical daily profiles).

The main difference with the cascade-based method is that the component-based method focuses on one component of the HVAC system such as a fan, a chilled water pump or a chiller. The pre-processing steps are applied to the variable modulating the HVAC component. This control variable is extracted from the BAS and recorded. The electric demand of the HVAC component is forecasted based on the current and past values of the control variable and the maximum electric demand of the equipment. The selection of the regressors is restrained to the selection of the relevant past values of the control variable to its future value; no other regressor is involved. For the component-based method, the iterated strategy (see chapter III) is employed for multi-step-ahead forecasting of the control variable.

This chapter is structured as follows. Section VI.1 presents the summary of the hybrid forecasting method that follows the process of KDD. The proposed method is applied to the case study of a supply fan in section VI.2. The conclusions are presented in the last section.

VI.1 Method

This section presents the application of the process of knowledge discovery in databases (KDD) [8] for the forecasting, over a time horizon of up to six hours, of the electric demand of the supply fan of an AHU. The method uses trend data from the BAS, which are recorded every 15 minutes. Although the case study is of a supply fan, the proposed method can be used for the forecasting of the operation of any other HVAC component such as heating and cooling coils, and chillers. The hybrid forecasting model is composed of a non-physical model and a physical model. The non-physical model is a time series model implemented in an artificial neural network. This model forecasts the control variable (supply fan modulation) over the next six hours (i.e., 24 time steps).

The result is passed to the physical model that forecasts the fan electric demand over the next six hours.

After the visual exploratory data analysis, the five steps of the process of KDD are followed in this study [8]. The BAS trend data is reduced to subsets by using a clustering algorithm; this is the selection step. Each subset is representative of a typical daily profile of the control variable. The regressors used to forecast the control variable are selected based on an autocorrelation analysis. A hybrid forecasting model is developed to solve the regression problem of mapping the future value of the electric demand of the HVAC component to the current and past values of the control variable. The forecasting results are presented and discussed. The proposed framework for the development of the hybrid forecasting model of the electric demand of an HVAC component is illustrated in Figure VI..

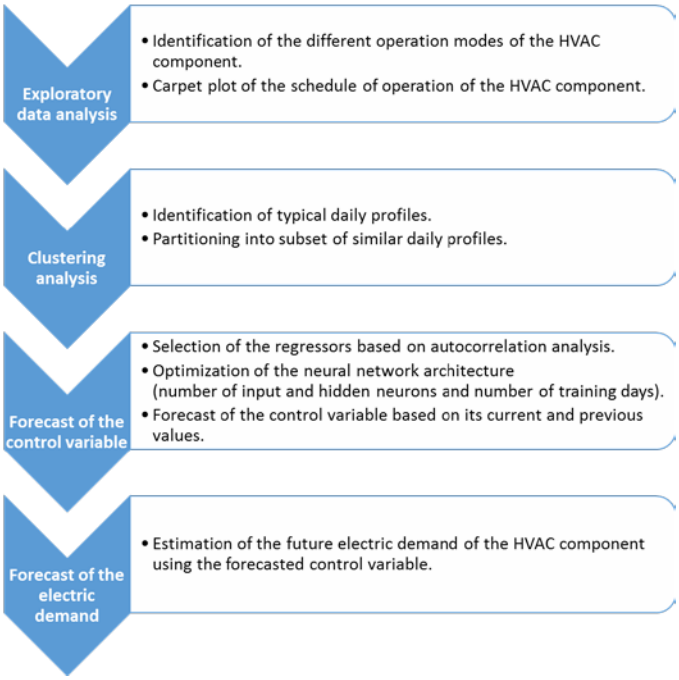

Figure VI.1 Flowchart of the proposed forecasting method.

VI.1.1 Exploratory analysis of the control variable

In this study, the control variable refers to the output of the PI controller modulating the HVAC component, which is recorded by the BAS. A visual exploratory data analysis is first performed to understand the operation of the HVAC system, and to identify patterns in the measurements of the control variable. This exploratory data analysis allows to identify visually different modes of

operation of the HVAC component and also identify evident outliers. Several tools are available such as carpet plots, box plots and histograms [114]. The carpet plot facilitates the visualization of the variation of the control variable with time [115]. For instance, it can present the schedule of operation of an HVAC component over the selected time interval in terms of days and hours of day.

VI.1.2 Clustering analysis of the control variable

The BAS trend data set is partitioned into subsets of similar daily profiles of the control variable. The partitioning helps to improve the performance of the forecasting model that is developed for each subset. Each daily profile is an individual in the entire population of daily profiles. The clustering analysis also identifies atypical daily profiles of the control variable meaning that the HVAC component was operated under unusual conditions.

The clustering analysis identifies similar individuals based on a metric, which is the Euclidean distance in this case. The control variable is recorded by the BAS at a 15-min time step giving a daily profile composed of 96 values. The individuals are reduced to eight features that characterize the pattern over the day to decrease the computing time. For each daily profile of the control variable, its mean and standard deviation over four selected time intervals of the day are calculated. Those time intervals are determined based on the operation profiles observed in the exploratory data analysis; they would vary for each HVAC component.

The incomplete daily profiles, with several missing values, are removed from the dataset; and clustering analysis is performed on the remaining daily profiles.

A fuzzy C-means clustering algorithm is chosen to detect and group the daily profiles of the control variable presenting similar patterns [116]. This algorithm provides a “fuzzy partition”, allowing some individuals to be unclassified if they are too different from the individuals of all the other clusters. They are identified as outliers or unusual individuals.

The fuzzy clustering algorithm requires to set the values of two parameters: the membership threshold and the number of clusters. The clustering algorithm calculates the coordinates of the center of each cluster that correspond to the eight features, which are used to minimize the distance between the clusters centers and the individuals. When a new individual is considered for the assignment to a cluster, the similarity for that individual is determined by the Euclidean distance

from the individual to the center of each cluster. A high similarity means that the individual is close to the center of the cluster. An individual is included into the cluster for which the calculated similarity is the highest. The membership threshold determines the minimum similarity required by the individual to belong to one of the clusters. If the similarity of an individual is less than the membership threshold, the individual is not assigned to any cluster; this is an atypical individual. For instance, if an individual has a similarity of 55% with cluster #1 and 65% with cluster #2, and the membership threshold is 70%, the individual is said atypical. If membership threshold is 60%, the individual belongs to the cluster #2. The membership threshold is set to remove the atypical individual from each cluster such as a daily profile with abnormal peak. This threshold should be high enough to extract the abnormal individuals but not too close to 100% to avoid having too many profiles removed.

The number of clusters is determined using the silhouette index [117] that measures the goodness of partition after clustering. The silhouette index $s(i)$ of an individual i is calculated as stated in equation [VI.1].

$$s(i) = (b(i) - a(i))/\max(a(i), b(i)) \qquad [VI.1]$$

where: $a(i)$ is the compactness (how close are the individuals in each cluster) that is calculated as the average distance between the individual i and all other individuals of the same cluster; and $b(i)$ is the separation (how well separated are the clusters) that is calculated as the average distance between the individual i and the individuals belonging to the closest neighbour cluster.

The clustering analysis aims at finding compact clusters (small value of a) which are well separated (high value of b): the silhouette index has to be as close as possible to one. The clustering algorithm is applied iteratively starting with one cluster, and then increasing the number up to the maximum of ten clusters. The silhouette index is calculated for each individual at each iteration; and then the average of the silhouette indices of all the individuals in each cluster is calculated. The optimal number of clusters is achieved when the overall average silhouette index is the closest to one.

The threshold is obtained by trial-and-error calculations, which results in the least number of abnormal profiles being removed. Finally, the daily profiles of the control variable are partitioned into different subsets corresponding to the optimum number of clusters.

VI.1.3 Time series forecasting model of the control variable

For each subset of similar daily profiles (each cluster), a time series forecasting model is developed for the control variable. The inputs of the time series model are the values at previous time steps of the control variable. An autoregressive time series model is chosen to forecast the value of the control variable based on some of its previous values. A nonlinear ANN model is retained for time series forecasting due to the non-stationary nature of the data. The first important step in the development of the neural network is the determination of the architecture of the network (number of neurons in the input and hidden layers). Two approaches are presented in sections 2.3.1 and 2.3.2 for the selection of the input variables: the filtering method and the hybrid wrapping-filtering method. The architecture of the network is determined by an optimization algorithm which is presented in section 2.3.2. In addition to the selection of input variables, the optimization algorithm tunes the number of hidden neurons and the training data set size.

VI.1.3.1 Input variables selection: filtering method

The value of the control variable at time t+1 is forecasted by using a number of relevant delayed values (values at previous times steps t, t-1, t-2 etc.), which are selected as inputs. The input variables are filtered and ranked based on the autocorrelation coefficient of the control variable. This is a preprocessing step that is less demanding in computing time than a wrapper algorithm [107].

For each cluster of daily profiles, the autocorrelation coefficient is calculated using equation [VI.2], where $\boldsymbol{y^t}$ is a vector of j values, which contains the previous values of $\boldsymbol{y}$ until and including the observation at time $\boldsymbol{t}$; the vector $\boldsymbol{y^{t-i}}$ contains the previous observations until and including the observation at time t-i; $\overline{y^t}$ is the average value of the vector $\boldsymbol{y^t}$.

$$R(i) = \frac{\sum_j (y_j^t - \overline{y^t})(y_j^{t-i} - \overline{y^{t-i}})}{\sqrt{\sum_j \left(y_j^t - \overline{y^t}\right)^2 \sum_j \left(y_j^{t-i} - \overline{y^{t-i}}\right)^2}} \qquad [VI.2]$$

The autocorrelation coefficient is calculated between the vector $\boldsymbol{y^t}$ and a vector of delayed values of the control variable $\boldsymbol{y^{t-i}}$ with i varying from 1 to 112 time steps (equivalent to 28 hours). A time delay in the control variable greater than 28 hours is not considered to limit the number of input variables.

These autocorrelation coefficients of the control variable are ranked in descending order. The values close to one indicate a strong autocorrelation of the variable at time t with those at previous time steps. In this study, only the previous values with an autocorrelation coefficient greater than 0.8 are selected. This threshold value indicates a strong linear correlation to the modulation at time t in terms of the Pearson correlation coefficient [102].

VI.1.3.2 Optimization of the network architecture

The nonlinear ANN model retained for the forecast of the control variable is composed of one input layer, one hidden layer and one output layer. The output layer includes one neuron corresponding to the fan modulation at the future time step. The number of neurons in the input layer equals the number of input variables. The activation function of the hidden layer is a hyperbolic tangent sigmoid transfer function (equation [VI.3]) and the output layer activation function is the identical transfer function (equation [VI.4]).

$$tanh(x) = \frac{exp(2x)-1}{exp(2x)+1} \quad [VI.3]$$

$$id(x) = x \quad [VI.4]$$

The neural network architecture is defined by three parameters: the number of inputs, the number of hidden neurons and the training data set size. For the optimization purpose, the maximum number of inputs is set at 50 for practical reasons, the number of hidden neurons is between 1 and 35, and the training set size is between one day and 25 days.

A hybrid wrapping-filtering method is used to select the input variables. It uses the filtering approach to rank the input variables based on their autocorrelation coefficient (section 2.3.1). The performance of the forecasting model is then iteratively calculated over the test data set with a varying number of the ranked input variables. Two other important parameters are obtained through the optimization: the number of hidden neurons and the training data set size.

For each combination of the three variables selected by the optimization algorithm, five different data sets are used for training and validation, the goal being the robustness of the results. Those five data sets differ only by their starting hour, selected randomly on the first day of the dataset. An example with seven days of training and six hours of validation is given on Figure VI.2. The second data set has the same composition, however, it starts two days later. The neural network is trained and validated over each one of the five data sets. The RMSE is calculated over each

validation set. The average of the five RMSEs is the metric returned by the objective function (F) (equation [VI.5]), where y_{1_i} is the i[th] observation of the control variable in the data set #1, and $\hat{y}_{1_i}$ is the corresponding forecasted value. A Simple Genetic Algorithm (SGA) [118] finds the three parameters that minimize the mean RMSE of the forecasted control variable over the six-hour validation set of the five datasets, corresponding to the time horizon of forecasting. Such an optimization is only performed once to define the best neural network's architecture.

$$F = \frac{\sqrt{\sum_{i=1}^{N}\left(\hat{y}_{1_i}-y_{1_i}\right)^2}+\sqrt{\sum_{i=1}^{N}\left(\hat{y}_{2_i}-y_{2_i}\right)^2}+\sqrt{\sum_{i=1}^{N}\left(\hat{y}_{3_i}-y_{3_i}\right)^2}+\sqrt{\sum_{i=1}^{N}\left(\hat{y}_{4_i}-y_{4_i}\right)^2}+\sqrt{\sum_{i=1}^{N}\left(\hat{y}_{5_i}-y_{5_i}\right)^2}}{5} \quad [VI.5]$$

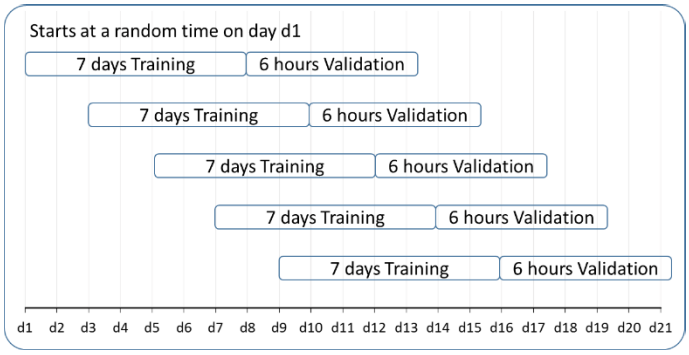

Figure VI.2 Example of five different data sets used for the architecture optimization.

The SGA used 20 individuals, which are the potential solutions, over twenty generations. The optimization approach using a SGA presents the advantage of performing a global search, with different potential solutions searched in parallel, and of avoiding stopping in local minima.

VI.1.4 Forecast of the control variable with a closed-loop ANN

Let us suppose that the optimum architecture of the ANN is composed of 15 inputs (for the current and past 14 observations), two hidden neurons and one output (the control variable). The dataset corresponding to one of the cluster is divided into three subsets in sequence, for training, validation and test of the forecasting model. In this example, the training set for the forecasting is composed of 10 days (960 observations) and the validation and test sets are six hours long each. The closed-loop ANN is trained with the Levenberg-Marquardt algorithm to forecast the value of the output variable over multiple time-steps.

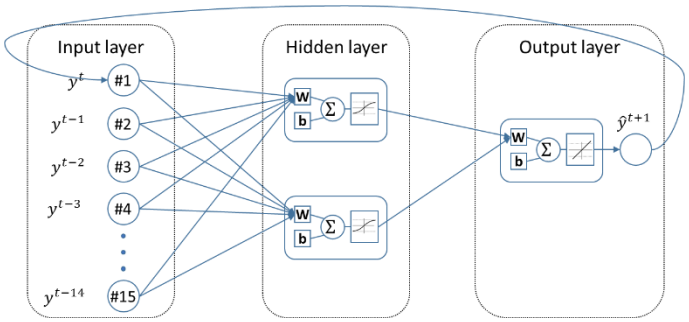

Figure VI.3 Closed-loop nonlinear autoregressive neural-network.

Figure VI.3 presents the optimum architecture of the closed-loop ANN model that uses one input layer with 15 neurons, one for each relevant delayed value of the control variable. The hidden layer includes 2 neurons and the output layer presents one scalar output: the forecast of the control variable. Starting with inputs of known values, for instance, from time t to t-14, the neural network forecasts the value of the target variable at time t+1 ($\hat{y}^{t+1}$). The value of $\hat{y}^{t+1}$ is then used as an input (along with values at times t, t-1, …, t-13) to forecast the value at the next time step ($\hat{y}^{t+2}$). In order to keep constant the number of inputs, once the value at time t+1 is used as an input, the last value at time t-14 is removed from the list of inputs and so on. The closed-loop network is also called dynamic neural network, where the output depends on the current input but also the previous inputs and outputs.

The Levenberg-Marquardt algorithm tunes the weights and biases of the network (**W** and **b** on Figure VI.3) to minimize the Mean Squared Error (MSE) over the training set. The training of the network is stopped by one of the two criteria: (i) if the magnitude of the gradient is below 1.0e-7 or (ii) if the performance on the validation set did not decrease after 10 successive iterations.

Once the closed-loop ANN is trained and validated, the non-physical model can then forecast the control variable over the following six hours. Let us assume as an example that the starting point is at time t, from which the control variable is forecasted over the following four time steps (Figure VI.4) using the current and past 14 values of the control variable as regressors. In the first forecast, the value of the control variable at time t+1 is forecasted using known (measured) values at time t, t-1, t-2, …, t-14. The value at time t+2 is forecasted using known (measured) values at time t, t-1, …, t-13 and the forecasted value at time t+1. The calculations are repeated until the forecast of the value at time t+4 is obtained. The forecast of the control variable over the testing time interval is shown with square markers, while the circles on red line correspond to measured values.

As soon as a new measurement is available (i.e., the value at time t+1), the second forecast takes place. The ANN is retrained over the past ten days starting with the previous weights. The second forecast of the control variable is performed over the following four time steps (t+2, t+3, t+4, t+5) by using the measured values at time t+1, t, t-1,…t-14; it is presented with diamond markers on Figure VI.4.

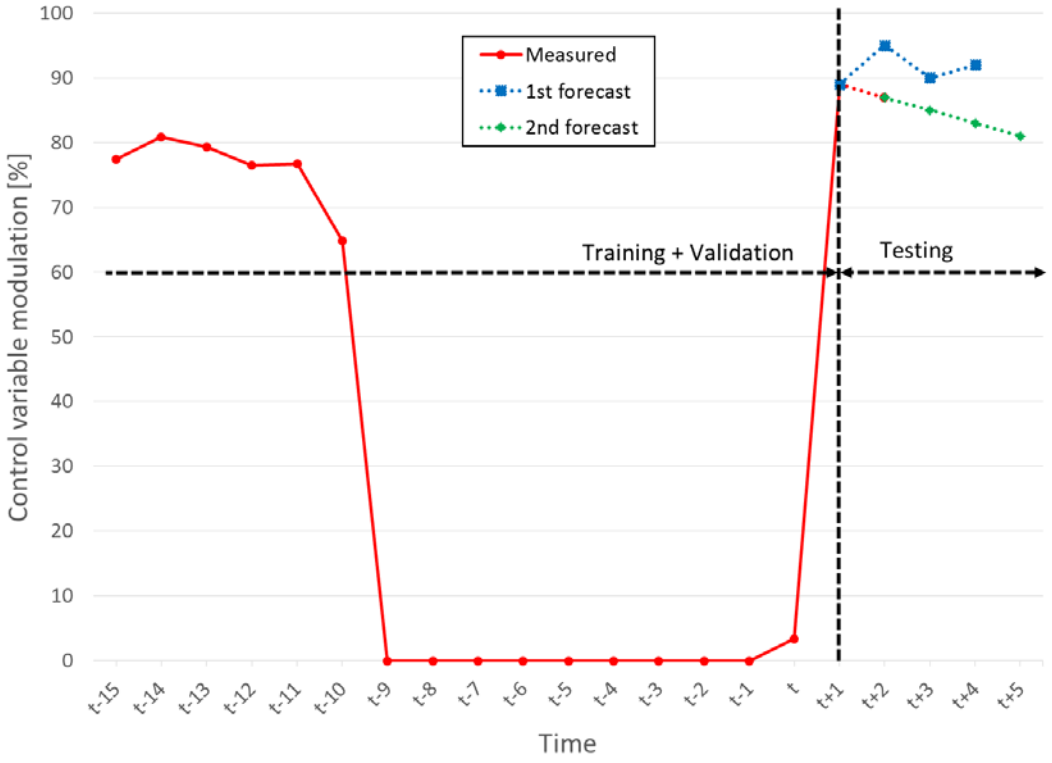

Figure VI.4 Multistep forecasting of the control variable with a sliding window training set.

VI.1.5 Forecast of the electric demand of the HVAC component

The forecasted value of the control variable is passed to the physical model which estimates the electric demand of the HVAC component over the same test set. In this study, the HVAC component under investigation is a supply fan with a capacity of 4720 L/s. In this example, the electric demand of the supply fan from time t+1 to t+4 is estimated using the forecasted control variable (supply fan modulation: mod_{fan}^{t+1}) from time t+1 to t+4, and the electric demand of the supply fan at full load (equation [VI.6]). The input electrical power demand of the supply fan at full load ($P_{in\ Full\ load}$) is estimated based on the manufacturer data of the brake horsepower of the fan ($P_{BHP} = 19\ kW$) and its efficiency ($\eta = 93.6\%$) both at full load.

$$P_{fan}^{t+1} = P_{in\ Full\ Load} . \left(mod_{fan}^{t+1}/100\right)^3 \quad [VI.6]$$

VI.2 Case study: fan electric demand

This section presents the forecasting of the electric demand of a supply fan of 4720 L/s of an AHU that serves an office building in Shawinigan-Sud, Quebec.

VI.2.1 Exploratory analysis of the supply fan modulation

A visual exploratory analysis of the supply fan modulation is first performed to identify the different modes of operation of the supply fan. The schedule of operation of the supply fan is presented by the carpet plot in Figure VI.5.

The supply fan modulation is defined as the ratio of the rotation speed of the fan over its maximal rotation speed, presented in percentages. The carpet plot presents the selected days on the Y-axis (from March 10th to August 7th, 2014), and the hours of the day on the X-axis. The supply fan modulation value of each cell is indicated by using the color legend, from 0% in dark blue to 100% in dark red. The white cells correspond to missing values.

The carpet plot shows that the supply fan is used only on working days, from Monday to Friday, from 00:00 to 02:30 and then from 05:00 to 18:45, except Monday (see the vertical red dotted lines). On Monday, the supply fan works from 3:45 to 18:45. The time intervals when the fan is off are showed in dark blue highlighted by the pink horizontal lines over weekend and vacations: April 18-21 and May 19.

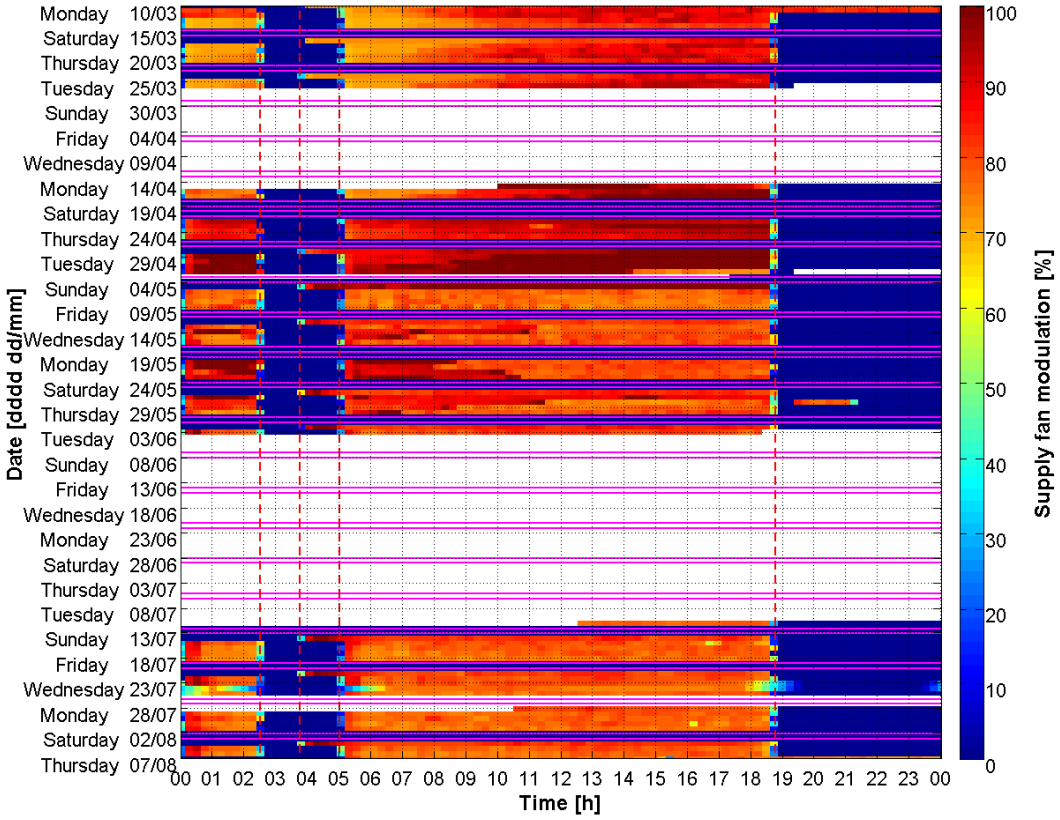

Figure VI.5 Carpet plot of the supply fan modulation value.

VI.2.2 Clustering analysis of the supply fan modulation

The BAS trend data set is partitioned into subsets of similar daily profiles of the fan modulation. The partitioning helps to improve the performance of the forecasting model that is developed for each subset. The clustering analysis also identifies atypical daily profiles of the control variable meaning that the supply fan was operated under unusual conditions.

A fuzzy C-means clustering algorithm is selected to group the similar daily profiles of the supply fan modulation into clusters, based on the Euclidean distance between the profiles. The fan modulation is recorded at a 15-min time step giving a daily profile composed of 96 values. In the trend data of March 10th to August 7th, one can extract 151 daily profiles of 96 values each. The daily profiles of the fan modulation are reduced to eight features that characterize the pattern over the day (Table VI.) to decrease the computing time. The mean and standard deviation of the fan modulation over four selected time intervals of the day are used as features. Those time intervals are determined based on the operation profiles identified in Figure VI.5. The features are calculated based on values, from the same sample space, of the fan modulation of summer 2014.

Table VI.1 Features characterizing each daily profile of the supply fan modulation

Features	Description
$\boldsymbol{\mu_1, \sigma_1}$	Mean and standard deviation over the time period from 00:00 to 03:00
$\boldsymbol{\mu_2, \sigma_2}$	Mean and standard deviation over the time period from 03:00 to 05:00
$\boldsymbol{\mu_3, \sigma_3}$	Mean and standard deviation over the time period from 05:00 to 19:00
$\boldsymbol{\mu_4, \sigma_4}$	Mean and standard deviation over the time period from 19:00 to 24:00

The incomplete daily profiles, with several missing values, are removed from the dataset; and clustering analysis is performed on the 86 remaining daily profiles.

The fuzzy clustering algorithm requires to set the values of two parameters: the number of clusters (k) and the membership threshold. The number of clusters is determined using the silhouette index [117] that measures the goodness of partition after clustering. The silhouette index (presented in section 2.2) has to be as close as possible to one. The clustering algorithm is applied iteratively starting with one cluster, and then increasing the number up to a maximum of ten clusters. The silhouette index is calculated for each individual at each iteration; and then the average of the silhouette indices of all the individuals in each cluster is calculated. The optimal number of clusters is achieved when the overall average silhouette index is the closest to one. Figure VI.6 shows the resulting value of the silhouette index for this case study when comparing the number of clusters k from 2 to 5. The X-axis presents the number of clusters and the Y-axis is the silhouette index. The quality of clustering is given by the overall average width of the silhouette over the total number of cluster. For instance, in the case of three clusters (k=3), the overall average silhouette index $\bar{s}(k) = 0.9811$; the silhouette index of each individual (represented by the length of blue line) is greater than 0.8. In the case of k=4, the overall average is smaller: $\bar{s}(k) = 0.8920$, with smaller silhouette indices for individuals in clusters 1 and 4. Hence the optimal number of clusters is three (k=3) since the average silhouette index of 0.9811 is the closest to one.

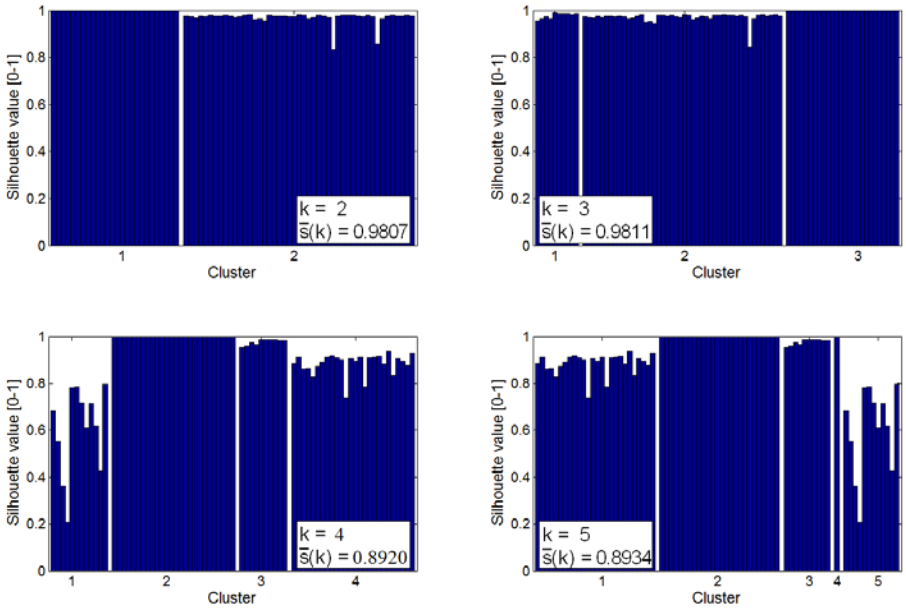

Figure VI.6 Silhouette index for different number of clusters.

The daily profiles of the supply fan modulation are partitioned into three clusters, based on the clustering analysis, with a membership threshold of 75%. This threshold was obtained by trial-and-error calculations, which resulted in only four abnormal profiles being removed. The 15-min time-step daily profiles are displayed in red on Figure VI.7. The black doted profiles correspond to the centers of the clusters calculated by the clustering algorithm, and represent the typical daily profile of the supply fan modulation. The coordinates of the center of clusters #1 and #3, are presented in Table VI.2 by the mean and standard deviation.

Table VI.2 Values of the features corresponding to the center of each cluster of the supply fan modulation

	Features	Value [%]	Time period
Cluster #1	$\mu_1 \pm \sigma_1$	65.8 ± 29.1	00:00 to 03:00
	$\mu_2 \pm \sigma_2$	0.1 ± 0.1	03:00 to 05:00
	$\mu_3 \pm \sigma_3$	77.4 ± 18.8	05:00 to 19:00
	$\mu_4 \pm \sigma_4$	1.4 ± 0.8	19:00 to 24:00
Cluster #3	$\mu_1 \pm \sigma_1$	0.7 ± 0.3	00:00 to 03:00
	$\mu_2 \pm \sigma_2$	42.1 ± 43.5	03:00 to 05:00
	$\mu_3 \pm \sigma_3$	82.3 ± 13.5	05:00 to 19:00
	$\mu_4 \pm \sigma_4$	4.0 ± 0.2	19:00 to 24:00

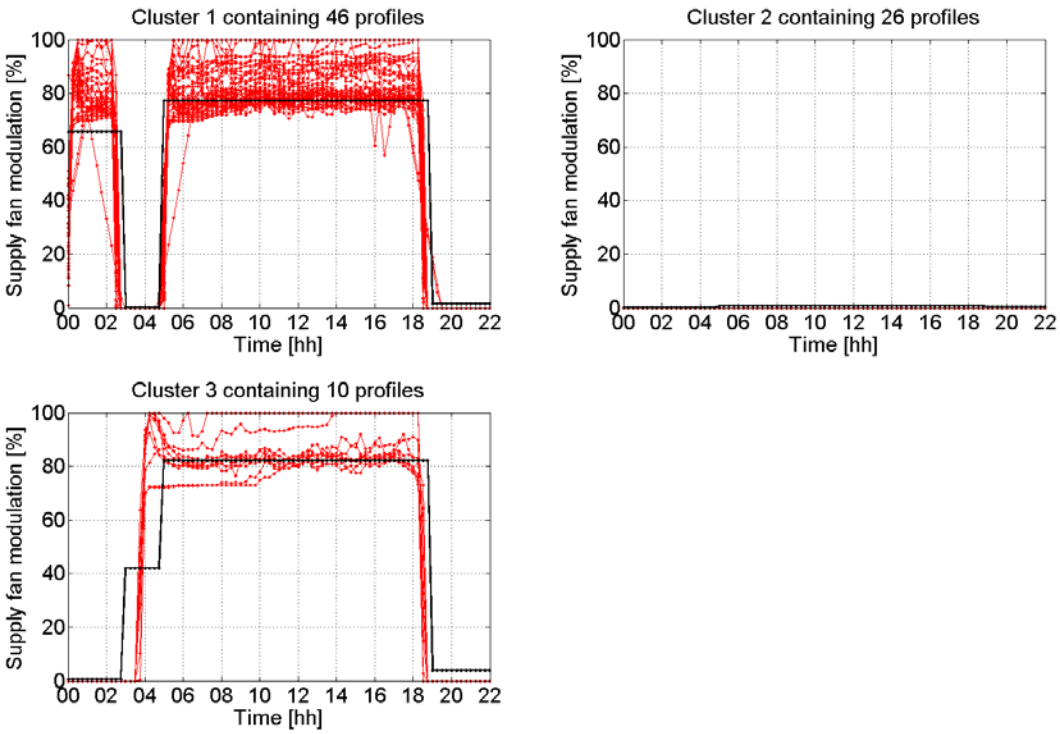

Figure VI.7 Three clusters of supply fan modulation profiles.

Four daily profiles are classified as atypical by the clustering algorithm (Figure VI.8). For instance, the daily profiles of March 10 and 11 are considered as atypical because the fan does not stop around 18:00, but continues to run until 23:00.

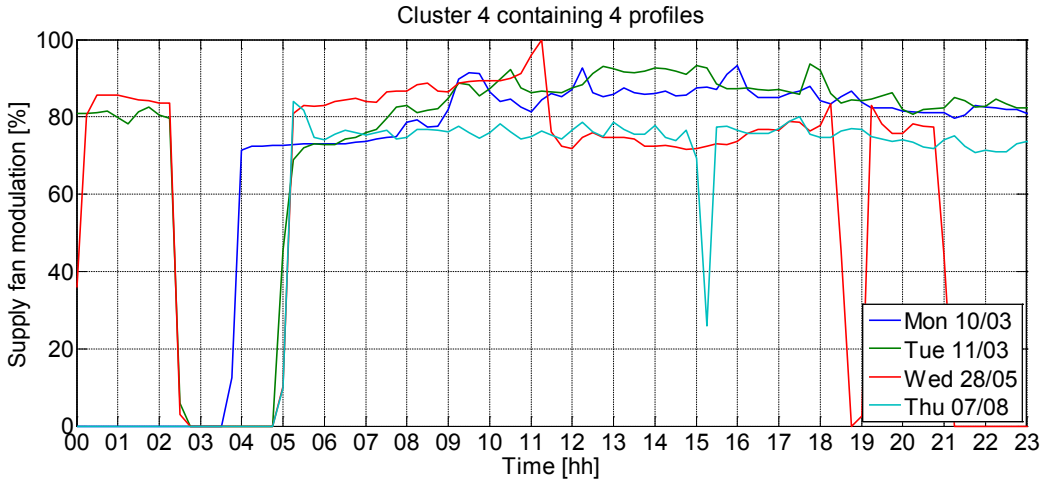

Figure VI.8 Cluster 4 with atypical modulation profiles.

A day-type analysis is performed on the daily profiles of each cluster (Figure VI.9). The results reveal that cluster #2 includes only non-working days: 23 weekends as well as holidays. Cluster #1 and #3 contain only the working-days: Tuesdays to Fridays for cluster #1 and Mondays for

cluster #3. The peak recorded early in the morning (cluster #3 in Figure VI.7) is therefore due to fan start-up on Mondays after weekends.

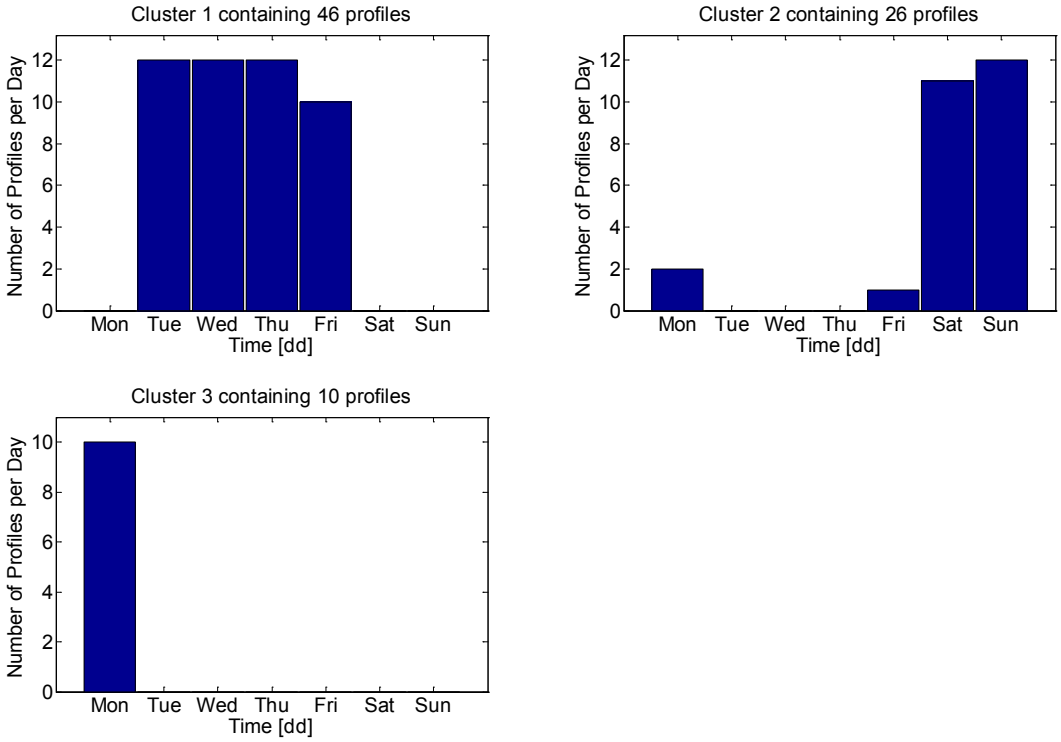

Figure VI.9 Day-type of the modulation daily profiles.

A time series forecasting model should be developed for each cluster of daily profiles, while the atypical profiles are removed from the dataset.

VI.2.3 Time series forecasting model of the supply fan modulation

An autoregressive time series model is chosen to forecast the value of the control variable based on some of its previous values. A nonlinear ANN model is retained for time series forecasting and it is implemented in the Matlab Neural Network Toolbox [119]. Two approaches are presented in sections 3.3.1 and 3.3.2 for the selection of the input variables: the filtering method and the hybrid wrapping-filtering method. The architecture of the network is determined by an optimization algorithm which is presented in section 3.3.2. In addition to the selection of input variables, the optimization algorithm tunes the number of hidden neurons and the training data set size.

VI.2.3.1 Input variables selection: filtering method

With the filtering method, the input variables of the forecasting model are selected, for each subset of daily profiles (cluster), by calculating the autocorrelation coefficient (equation [VI.2]). Figure VI.10 presents, as an example, the autocorrelation coefficients of the supply fan modulation for the cluster #1, which are then ranked in descending order in Table VI.3.

The values close to one indicate a strong autocorrelation of the variable at time t with those at previous time steps. In this study, only the previous values with an autocorrelation coefficient greater than 0.8 are selected. The supply fan modulation at time t is correlated with the measurements of the previous 90 min (6 measurements) as well as the thirteen measurements occurring about 24 hours before (blue rectangles on Figure VI.10). They correspond to the first 19 variables of Table VI.3. Therefore, the future value of the supply fan modulation at time $t+1$ could be expressed as stated in equation [VI.7].

$$mod_{fan}^{t+1} = f(mod_{fan}^{t}, \dots, mod_{fan}^{t-5}, mod_{fan}^{t-89}, mod_{fan}^{t-101}) \qquad [VI.7]$$

Table VI.3 Autocorrelation coefficients of the supply fan modulation, in descending order, using measurements of cluster #1

Rank	Autocorrelation coefficient of the supply fan modulation	Delay [HH:MM]	Delay [time step]
1	0.9794	00:15	t
2	0.9701	24:00	t-95
3	0.9548	23:45	t-94
4	0.9546	24:15	t-96
5	0.9478	00:30	t-1
6	0.9249	23:30	t-93
7	0.9245	24:30	t-97
8	0.9167	00:45	t-2
9	0.8949	23:15	t-92
10	0.8943	24:45	t-98
11	0.8856	01:00	t-3
12	0.8648	23:00	t-91
13	0.8641	25:00	t-99
14	0.8544	01:15	t-4
15	0.8344	22:45	t-90
16	0.8336	25:15	t-100
17	0.8229	01:30	t-5
18	0.8038	22:30	t-89
19	0.8028	25:30	t-101
20	0.7916	01:45	t-6
21	0.7733	22:15	t-88
22	0.7723	25:45	t-102

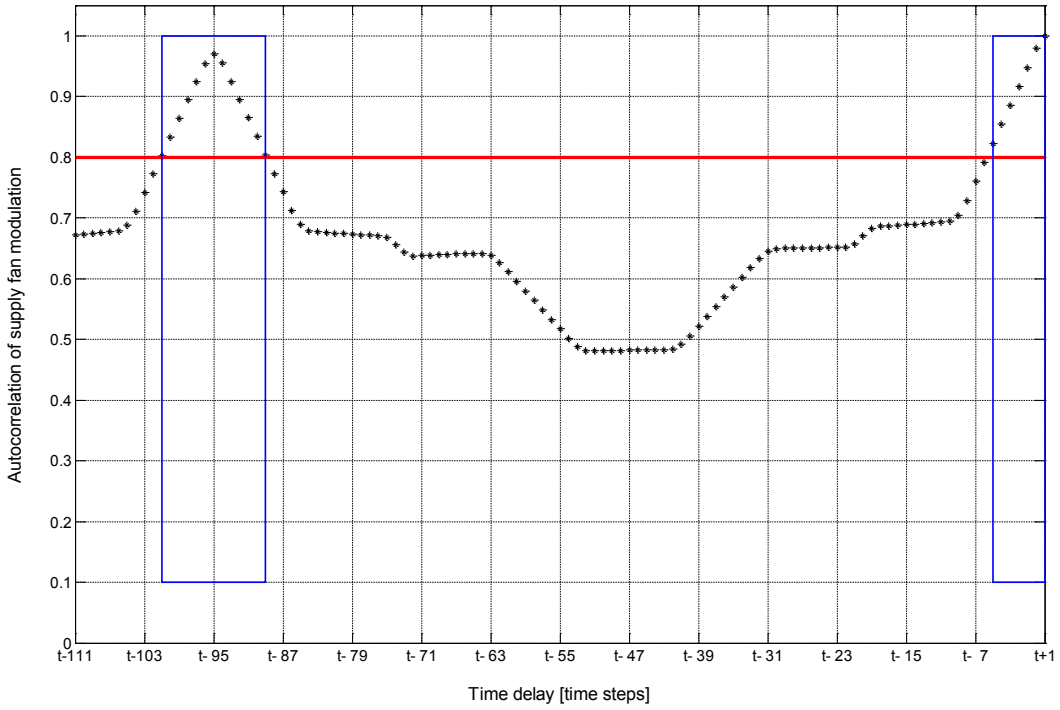

Figure VI.10 Autocorrelation coefficients of the supply fan modulation versus previous values of the measurements of cluster #1.

VI.2.3.2 Optimization of the network architecture

The supply fan modulation (control variable) is forecasted by a nonlinear ANN model. The neural network architecture is defined by three parameters: the number of inputs, the number of hidden neurons and the training data set size. For the optimization purpose, the maximum number of inputs is set at 50 for practical reasons, the number of hidden neurons is between 1 and 35, and the training set size is between one day (96 observations) and 25 days.

In this second approach, a hybrid wrapping-filtering method is used to select the input variables. It uses the filtering approach to rank the input variables based on their autocorrelation coefficient. The first 22 inputs are presented in Table VI.3. The performance of the forecasting model is then iteratively calculated over the test data set with a varying number of the ranked input variables. Two other important variables are obtained through the optimization: the number of hidden neurons and the training data set size.

This chapter presents the use of cluster #1 that contains data of 46 days, from March 12 to August 6. Thirty-six days are used to optimize the ANN's architecture, from March 12 to July 17. The ten days left, from July 17 to August 6, are kept to test the forecasting model. The ANN developed for cluster #3 is not presented because of space limitation.

The optimum architecture of the ANN, given by the SGA optimization algorithm, has 13 input variables, 1 hidden neuron and 23-day training set size. The RMSE between the measurements and forecasts over the time horizon of six hours is equal to 7%. It is worth mentioning that those 13 inputs were the first 13 inputs among the 19 selected by the filtering method (Table VI.3).

VI.2.4 Forecast of the fan modulation with a closed-loop ANN

Once the optimum ANN architecture was defined, the dataset corresponding to cluster #1 is divided into three subsets in sequence, for training, validation and test of the forecasting model. The close-loop ANN is trained over the 23-day period starting from April 25 to July 17, extracted from the dataset of cluster #1. The validation is performed from 17:00 to 23:00 on July 17. The test is performed and presented over the period of the day the most difficult to predict, starting at 23:00 on July 17th (Figure VI.11), when the fan starts and stops. The validation and test are performed over the remaining subset of cluster #1 that was not used to optimize the ANN architecture.

Once the closed-loop ANN model is trained and validated, it gives the first forecast over the next six hours. This first forecast, starting at 23:00 on July 17 and ending at 05:00, is displayed in blue with squares markers on the top part of Figure VI.11. The network is retrained each time a new observation is available. A sliding window of 23 days is used for training: each time a new observation is included in the training set, the first observation is removed. At each iteration, a new forecast is provided over the following six hours. Only the first forecast is presented on Figure VI.11. The recorded fan modulation is also displayed in red on Figure VI.11. The performance of the forecasting model is presented in terms of Root Mean Squared Error (RMSE) between the forecasted and recorded values of the fan modulation over the six-hour test. The Coefficient of Variation of the RMSE (CV(RMSE)) and the Absolute Value of the Forecasting Error (AVFE) are presented as well (Equations [VI.8],[VI.9] and [VI.10]).

$$RMSE = \sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/n} \quad [VI.8]$$

$$CV(RMSE) = RMSE/\bar{y} \quad [VI.9]$$

$$AVFE(i) = |\hat{y}_i - y_i| \quad [VI.10]$$

where $\hat{y}_i$ is the forecasted observation i and y_i is the measured one; $\bar{y}$ is the average value of the variable y over the selected period. A RMSE of 5.5% and CV(RMSE) of 17.6% is reached over

the testing period for the first forecast. The lower part of Figure VI.12 presents the AVFE, in the units of fan modulation [%] over the six hour testing step.

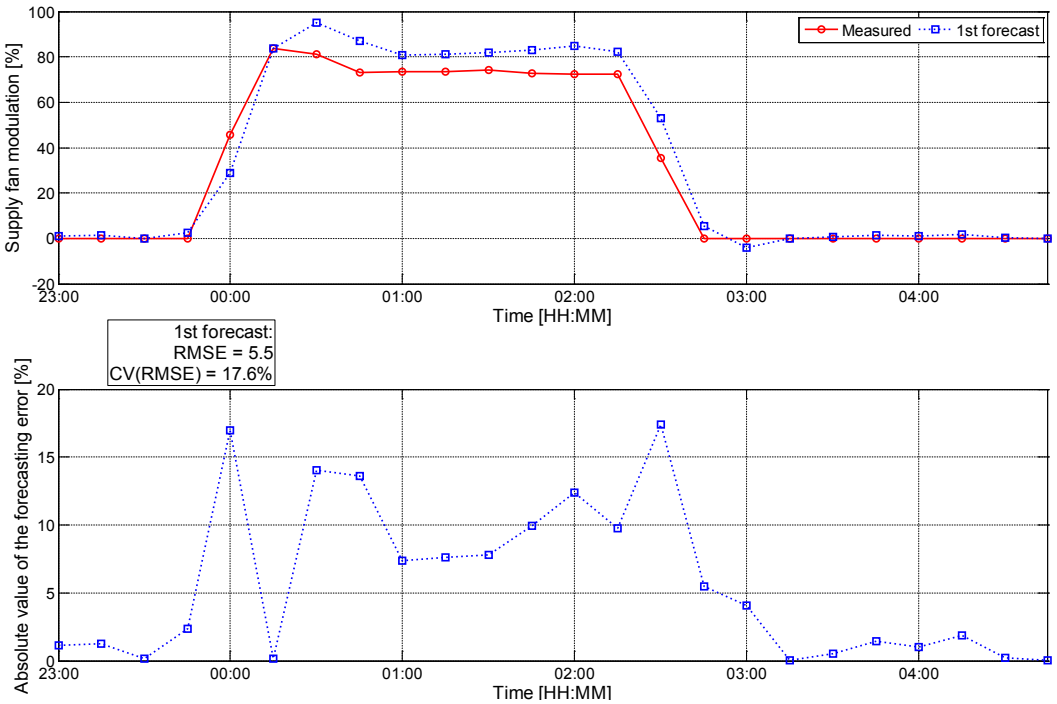

Figure VI.11 Multistep forecasting of the supply fan modulation over the test set on July, the 17th.

The evolution of the AVFE over the time horizon of six hours (24 time steps) is presented in the boxplots in Figure VI.12. The AVFE of the twenty-four successive forecasts of the fan modulation is analysed in terms of the time-step of the prediction horizon; the degradation of the forecasting performance is observed along with the prediction horizon. Some statistical indices of the absolute value of the forecasting error of the twenty-four consecutive forecasts are presented.

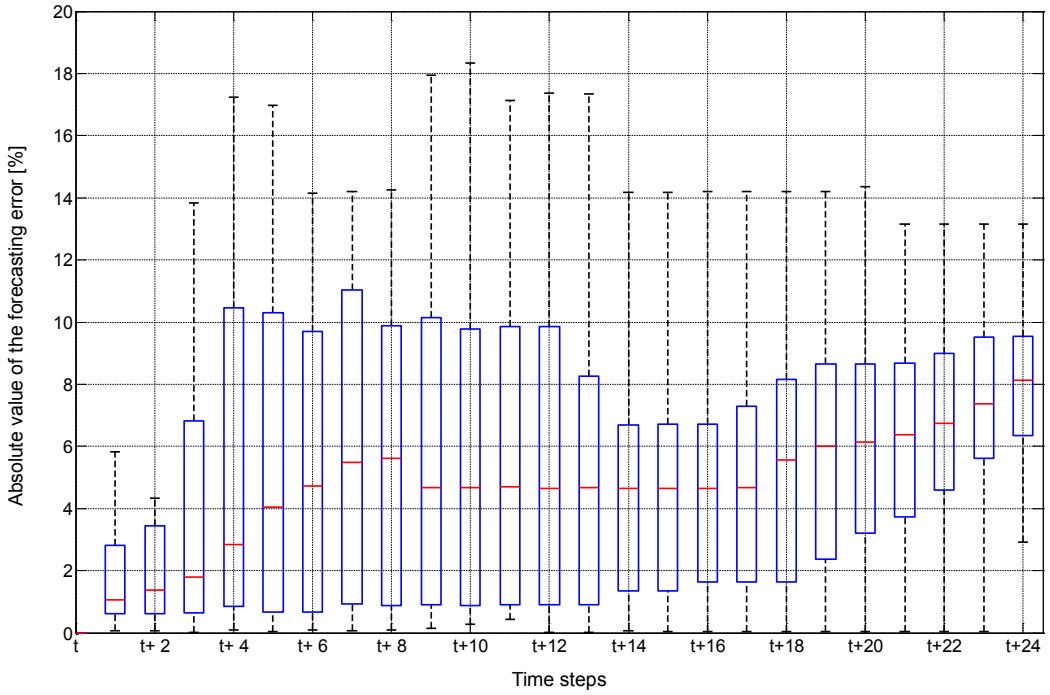

Figure VI.12 AVFE of the fan modulation with clustering over the time horizon of six hours.

At each time step, the horizontal red line indicates the median value; the top and bottom edges of the box are respectively the 25th and 75th percentiles; and the black "whiskers" indicate the upper (U_b) and lower (L_b) bound, corresponding to the median plus or minus 2.7 times the standard deviation. They are respectively calculated as presented in [114] (see Equation [VI.11]).

$$U_b = Q_3 + 1.5 \times IQR$$
$$L_b = Q_1 - 1.5 \times IQR \qquad [VI.11]$$

where Q_1 and Q_3 are the first and third quartiles and IQR is the interquartile range.

The values exceeding the bounds are considered as outliers. About two to three outliers are identified at each forecasted time step; they are not displayed on Figure VI.12. The median of the absolute forecasting error gradually increases but does not exceed 6% over the first 18 time steps, and reaches about 8% after six hours. The mean absolute percentage error (MAPE) of 8.45% is calculated when the fan works (i.e., fan modulation different from 0%), from 00:00 to 02:30 on July 17th. The MAPE is reduced to 4.5% when the first and last measurements, corresponding to the start and stop of the fan, are excluded.

VI.2.5 Forecast of the supply fan electric demand

The electric demand of the supply fan is forecasted by the physical model using the forecasted control variable (supply fan modulation), given by the time series model, and the electric demand of the supply fan at full load (equation [VI.6]). The physical model presented in equation [VI.6] is used to forecast the fan electric demand over the six-hour test set (Figure VI.13). The measured profile is displayed in red and the forecast in blue on the upper part of the graph. The lower part shows the AVFE. The forecasting model presents a RMSE of 1.4 kW and CV(RMSE) of 30% over the six hour test set.

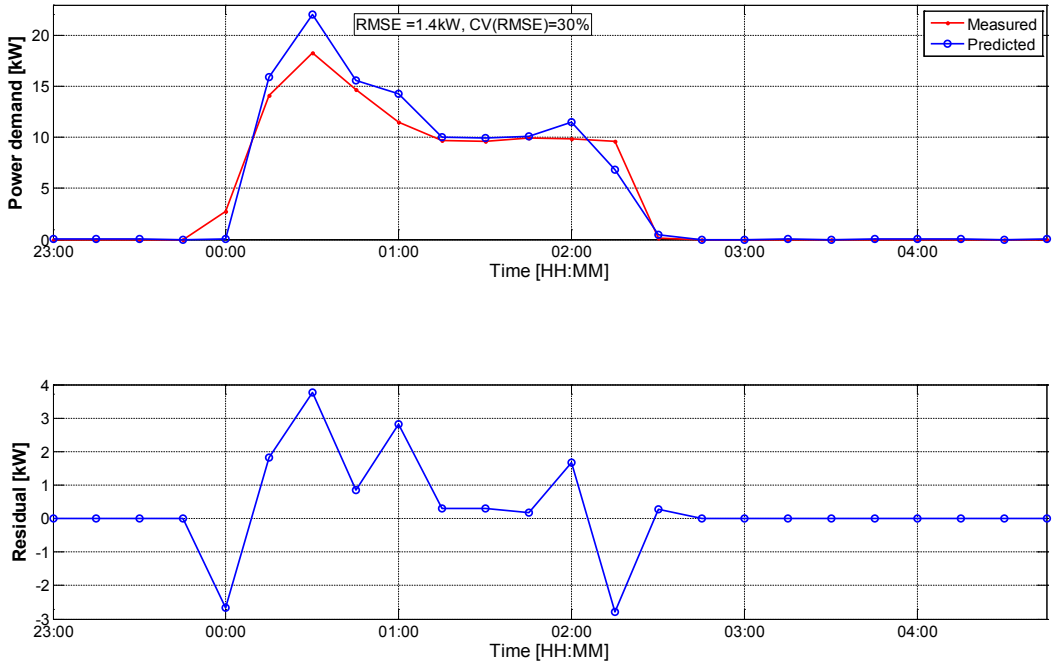

Figure VI.13 Supply fan electric demand forecasting over July, the 17th with the modulation forecast of the closed-loop network.

The forecasting error of the twenty-four consecutive forecasts is also analyzed. The evolution of the AVFE with the increase of time horizon is displayed on Figure VI.14. The median of the AVFE gradually increases to reach about 2 kW after 19 time-steps. When the fan is in operation, the MAPE is 22.9%, and it is reduced to 14.8% when the measurements corresponding to the start and stop of the fan are excluded.

The forecasting model presents a better performance during working hours from 05:00 to 18:45. A MAPE of 8.6% was calculated over a six-hour horizon prediction from 06:00 to 12:00 on July 17th.

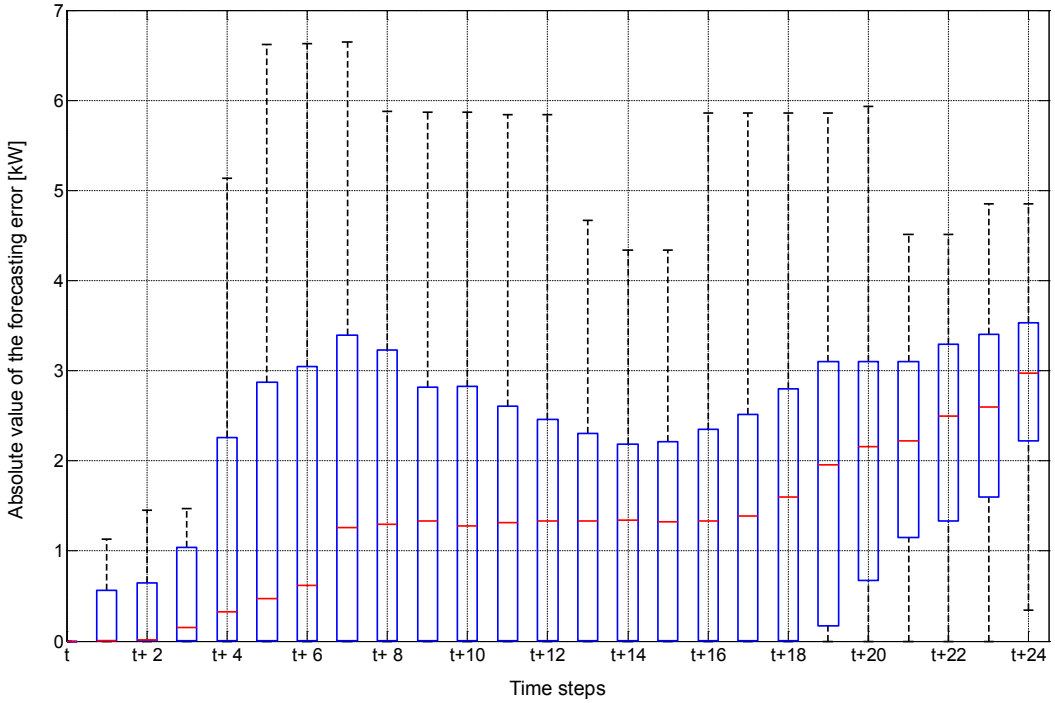

Figure VI.14 Absolute value of forecasting error of the fan electric demand over the time horizon of six hours.

VI.2.6 Discussion

The hybrid forecasting model was tested on the six-hour period the most difficult of the day presenting a start and stop of the fan. The performance of the first forecast after training and validation of the hybrid model is summarized in Table VI.4 and Table VI.5 for the modulation and electric demand of the fan. The RMSE and CV(RMSE) over the six-hour test period (with fans on and off, Figure VI.11 and Figure VI.13) are displayed in Table VI.4. Table VI.5 shows the results only when the fan are on, over a 2.5 hour time interval. The MAPE is also listed in Table VI.5 when the fan is operated and when the observations corresponding to the start and stop of the fan are discarded.

The forecasting model is said to be calibrated regarding the only existing guideline to the author's knowledge: ASHRAE guideline 14 [113]. The model presents a CV(RMSE) of 30.0 % over six hours at a 15-min time-step compared to the threshold of 30 % for hourly values.

Table VI.4 Results of the forecasts of supply fan modulation and electric demand

	RMSE	**CV(RMSE)**
Supply fan modulation	5.5 %	17.6 %
Electric demand of the fan	1.4 kW	30.0 %

Table VI.5 Results of the forecasts of supply fan modulation and electric demand when the fan is operated

	RMSE	CV(RMSE)	MAPE	MAPE excluding start and stop times
Supply fan modulation	11.7 %	16.9 %	8.5 %	4.5 %
Electric demand of the fan	2.0 kW	27.3 %	22.9 %	14.8 %

The hybrid forecasting model presents a median AVFE smaller than 0.5 kW on the first hour for the twenty-four successive forecasts (Figure VI.14). The maximum value of the AVFE is smaller than 1.5 kW on the first 45 min. The proposed hybrid model presents a good forecasting accuracy on the very short-term.

VI.2.6.1 Sensitivity analysis

A 23-day sliding window for training of ANN, as obtained from the architecture optimization in section 3.3.2, can be restrictive for the forecasting of fan modulation in an existing system. Therefore two sensitivity analyses are performed by reducing the number of training days from 23 days to 12, 8 and 4 days, respectively. The calculations are performed by using 13 input neurons and one hidden neuron as resulted from the architecture optimization.

(1) In the first sensitivity analysis, the average RMSE (Equation [VI.5]) between the forecasts and measurements of the fan modulation is calculated over the validation data sets of six hours long each, over the 36-day period from March 12 to July 17. The average RMSE does not vary significantly over the validation set when using 4, 8 or 12 training days (Table VI.6). When the number of training days was increased from 12 days to 23 days, the average RMSE is reduced from 9.63% to 6.27%.

(2) In the second sensitivity analysis, the impact of the training dataset size on the forecasting performance is also of interest. The RMSE between the forecasts and measurements of fan modulation is calculated over the validation data sets of six hours long each, from July 17 to August 6. In this analysis compared to (1), fifty trial runs are performed for each training set size of 4, 8, 12 and 23 days, respectively. Each validation set has a randomly chosen starting time. The training and testing sets are defined as the time intervals directly preceding, and respectively following, the validation set. The median RMSE shows a negligible decrease over the validation set with the increase of the training set size, from about 5% (with 4 days) to 4% (with 23 days) (Figure VI.15). The median RMSE over the six-hour testing set that follows the validation set remains

approximately constant at around 5% with the increase of increase of the training set size (Figure VI.16).

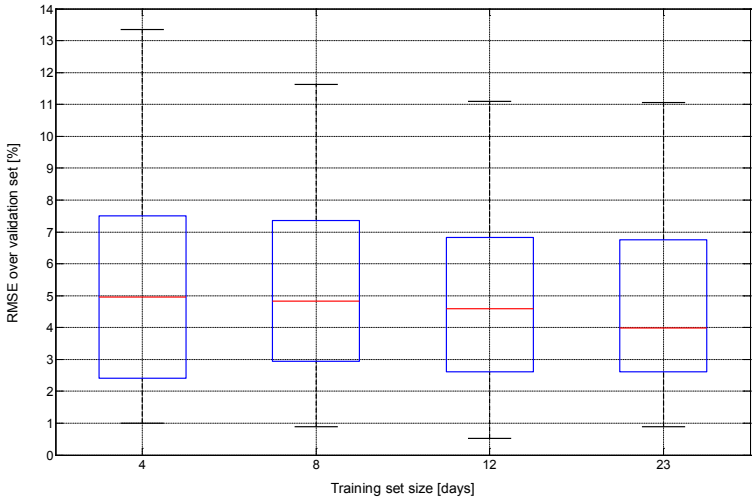

Figure VI.15 RMSE over the validation set between July 17 to August 6 when using 4, 8, 12 and 23 training days.

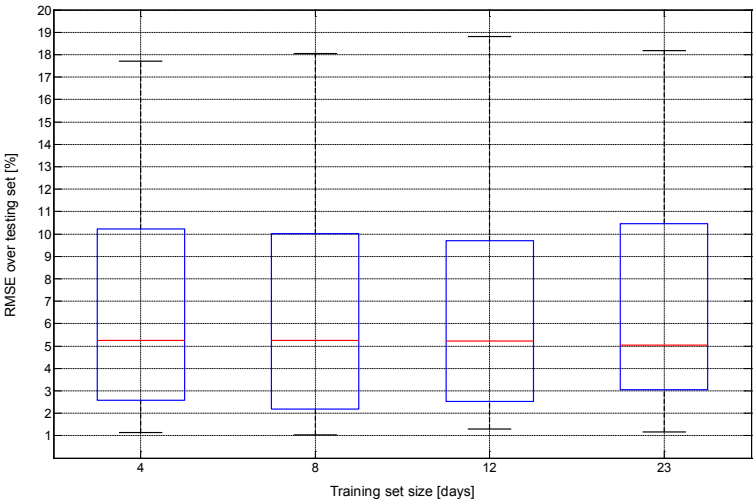

Figure VI.16 RMSE over the testing data set between July 17 and August 6 when using 4, 8, 12 and 23 training days.

The sensitivity analyses shows that the reduction of training data set size from 23 days to 4 or 8 days does not have a negative impact on the value of RMSE.

Table VI.6 Variation of average RMSE with different training data set sizes, over validation data sets of six hours long each, returned by the objective function F; with 13 inputs and one hidden neuron.

Number of training days	Average RMSE [%]
4	9.93
8	9.61
12	9.63
23	6.27

The median AVFE of the fan modulation over the six-hour time interval (24 time steps) is between 0% and 8% (Figure VI.12) when the cluster #1 is used for validation, compared with 10% to 40% (Figure VI.17) when the whole original data set is used. This result highlights the advantage of using clusters for the forecasting purposes.

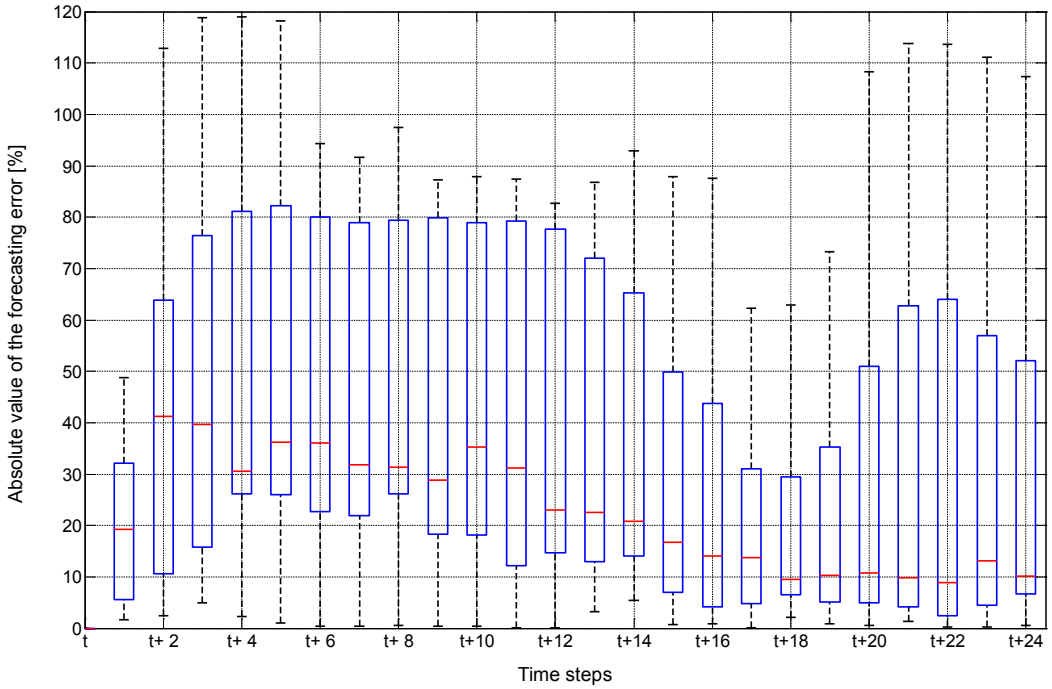

Figure VI.17 Absolute value of forecasting error of the fan modulation without clustering over a six-hour testing set.

VI.3 Conclusions

This study presented the use of data mining techniques applied to the BAS trend data for the development of forecasting models of the supply fan modulation and electric demand over a time

horizon of six hours. The proposed hybrid forecasting model combines a nonlinear ANN time series model that forecasts the fan modulation, with a physical model that forecasts the supply fan electric demand. The originality of this work consists in the development of a framework for the multistep forecasting of electric demand over a horizon of up to six hours of an existing HVAC sub-system based on sub-hourly BAS trend data.

First the ANN architecture was optimized with a genetic algorithm, followed by the use of that architecture for the forecasting of the fan modulation. The fan modulation was forecasted over the testing period with RMSE of 5.5% and CV(RMSE) of 17.6%. The mean absolute percentage error (MAPE) of 8.45% was calculated when the fan is on. The fan electric demand was forecasted with MAPE of 22.9%, RMSE of 2.0 kW, CV(RMSE) of 27.3 % when the fan is on.

The sensitivity analysis indicated that the RMSE between the forecasts and measurements of fan modulation does not change significantly, for this specific application, if shorter training data sets are used instead of the optimum of 23 days. The advantage of applying the hybrid forecasting model on cluster #1 of daily profiles rather than using the whole BAS trend data set was revealed.

The proposed forecasting method presents the advantage of having a good generalization skill, which allows its application to diverse HVAC components. Two distinct models are combined: a closed-loop nonlinear autoregressive neural network (ANN) model for the forecasting of fan modulation with a physical model for the forecasting of electric demand of the fan. The two models are coupled in this sequence in a hybrid model, allowing for more flexibility in the assessment of different demand response strategies. The ANN model enables the forecast of control variable (the fan modulation), while the physical model enables the forecast when fan performance curves, which link the electric demand to the air flow rate, are changed.

In this case study, the forecast of the control variable uses as regressors only its current and past values. The use of other regressors in the forecasting model of the control variable might bring more robustness to the model and help in the forecast of unexpected events.

Chapter VII. Main contributions and future work

This thesis investigated the issue of short-term forecasting of the electric demand of HVAC systems in the context of demand response. Two methods were proposed for forecasting the electric demand over a horizon of prediction up to six hours: a cascade-based method and a component-based method.

The cascade-based method is a global approach that forecasts successively six target variables of a HVAC system, i.e., the supply air flow rate, the cooling coil load, the whole building cooling load, the electric demand of the secondary and primary cooling systems. This multi-step forecasting method, based on SVMs, allows to estimate the contribution of some selected HVAC sub-systems on the total electric demand of the building.

The component-based approach presents a hybrid model forecasting the electric demand of a HVAC component based on the future value of its control variable. The hybrid model combines a nonlinear time series model based on ANN that forecasts the control variable (e.g. fan modulation), with a physical model that forecasts the electric demand of the HVAC component.

Both methods implement data mining techniques as pre-processing steps to help in the development of the forecasting models and improve their accuracy. The data mining techniques make use of the information embedded in the historical dataset of sub-hourly trend data collected from the BAS.

The main contributions of this thesis are listed as follows:

1) Multistep forecasting of the electric demand of HVAC cooling system at a 15-min time-step with a prediction horizon up to six hours. The inverse forecasting models are based on measurements; they reflect the real operating conditions. The forecasting models are developed using state-of-the-art SVM and ANN models known for their adaptability. The proposed method can be applied to a majority of existing building using a BAS. This work contributes to the research field of demand response by investigating the forecast of power demand on the demand side for HVAC systems of large CI buildings.

a. The cascade-based method forecasts the electric demand of some HVAC components. It gives the impact of some target variables (e.g. supply air flow rate, cooling coil load) on the electric demand.
b. The component-based method forecasts the electric demand of a HVAC component as a function of the control variable only. It facilitates the test of different demand response strategies.

2) Application of data mining techniques to help in the development of forecasting models. The DM techniques give better understand the operation of HVAC systems. A hybrid wrapping-filtering approach is developed for the selection of regressors and compared to a filtering method. DM techniques are used to identify typical daily profiles of the target variable and partition the training set. The impact of different sequences of the pre-processing steps on the forecasting performance of the model is analyzed.

3) Validation of the proposed methodology on two case studies. The cascade-based method has been tested for forecasting the electric demand of the cooling system of the Genomic research center on Loyola Campus of Concordia University. The component-based method has been validated on one component (the supply fan) of an office building located in Shawinigan-Sud.

The limitations and difficulties faced in this thesis are discussed in the following paragraphs.

The forecasting performance of the inverse models used in this work (ANN and SVM) is significantly impacted by the selection of regressors and the pre-processing steps. One hybrid wrapping-filtering approach for selecting the regressors was proposed in this study but more investigations should be further performed. The three parameters of the SVM model are of great importance on the prediction accuracy of the model. They must be identified with caution: a grid-search approach was performed in this work. A heuristic-based optimization algorithm can be tested to tune the parameters on a training set and reduce the computing time. More difficulties were faced when developing an accurate ANN-based forecasting model. There is no general rule to elaborate the optimum architecture of an ANN (number of hidden layers and neurons, type of activation function). The process of elaboration by trial and error is time-consuming.

Regarding the cascade-based method, some constraints appeared in terms of data available when partitioning the whole dataset into subsets based on the clusters of typical daily profiles. Some clusters presented a small number of observations making the training of the model less efficient. The component-based approach uses an autoregressive time series model for the forecast of the control variable. The forecasting accuracy of this model is impacted by its only one regressor. Adding other regressors would improve its performance but neutralize its ease in testing DR strategies.

The scope of this study was limited to the forecast of the electric demand of HVAC systems. The forecasting models should be implemented in a further work into a model predictive control approach. In such framework, the forecasting models could be used to estimate and minimize the electric peak demand or energy consumption of HVAC system over demand response events.

Regarding the noticed limitations impacting the performance of the data-driven forecasting models, further research is recommend on the pre-processing of the models. It includes the test of other approaches for the selection of regressors using for instance the entropy index of information theory. Another critical point in the pre-processing steps requiring more investigation is the selection of the adequate training set. This thesis showed that the forecasting models are significantly simplified when selecting a training set of similar daily profiles; their performance is improved as well. The role of data sample size in the training of data-driven models was recently investigated in [120]. The study also pointed out that the selection of a proper training set allows to reduce the complexity of the model. Further investigation are required in the selection of an adequate training set; covering the size of the training set but also the variability of the conditions and the age of the measurements (adaptive training have shown better performance than static training).

In the proposed component-based forecasting method, the control variable modulating the HVAC component is forecasted using as regressors only its current and past values. The use of other regressors in the forecasting model of the control variable should be investigated to assess the impact on the robustness of the model and help in the forecast of unexpected events.

References

[1] IEA. Key world energy statistics. International Energy Agency; 2015.

[2] Buildings Energy Data Book. Energy Efficiency & Renewable Energy, U.S. Department of Energy; 2012.

[3] OEE. Energy Efficiency Trends in Canada 1990 to 2010. Ottawa: Office of Energy Efficiency, Natural Resources Canada; 2013.

[4] Lee K-h, Braun JE. Development of methods for determining demand-limiting setpoint trajectories in buildings using short-term measurements. Building and Environment. 2008;43(10):1755-68.

[5] Lee K-h, Braun JE. Development and application of an inverse building model for demand response in small commercial buildings. SimBuild. Boulder, Colorado, USA2004. p. 1-12.

[6] Xu P, Haves P. Case Study of Demand Shifting with Thermal Mass in Two Large Commercial Buildings. ASHRAE transactions. 2006;112(1).

[7] Kurgan LA, Musilek P. A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review. 2006;21(01):1-24.

[8] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine. 1996;17(3):37.

[9] Piatetsky-Shapiro G. Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. AI magazine. 1990;11(4):68.

[10] Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in knowledge discovery and data mining: AAAI Press, 1996.

[11] Buchheit RB, Garrett Jr JH, Lee SR, Brahme R. A knowledge discovery framework for civil infrastructure: a case study of the intelligent workplace. International Conference on Computing in Civil and Building Engineering. Stanford, California, United States: American Society of Civil Engineers; 2000. p. 914-21.

[12] Kusiak A, Li M, Zheng H. Virtual models of indoor-air-quality sensors. Applied Energy. 2010;87(6):2087-94.

[13] Yu ZJ, Haghighat F, Fung BC, Morofsky E, Yoshino H. A methodology for identifying and improving occupant behavior in residential buildings. Energy. 2011;36(11):6596-608.

[14] Du Z, Fan B, Jin X, Chi J. Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. Building and Environment. 2014;73:1-11.

[15] Ramos S, Duarte JM, Duarte FJ, Vale Z. A data-mining-based methodology to support MV electricity customers' characterization. Energy and Buildings. 2015;91:16-25.

[16] Kusiak A, Li M, Tang F. Modeling and optimization of HVAC energy consumption. Applied Energy. 2010;87(10):3092-102.

[17] Fan C, Xiao F, Wang S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Applied Energy. 2014;127:1-10.

[18] Kusiak A, Li M, Zhang Z. A data-driven approach for steam load prediction in buildings. Applied Energy. 2010;87(3):925-33.

[19] Feinberg EA, Genethliou D. Load forecasting. Applied Mathematics for Restructured Electric Power Systems: Springer US; 2005. p. 269-85.

[20] Kyriakides E, Polycarpou M. Short term electric load forecasting: A tutorial. Trends in Neural Computation. 2007;35:391-418.

[21] Li X, Wen J. Review of building energy modeling for control and operation. Renewable and Sustainable Energy Reviews. 2014;37:517-37.

[22] Sallam AA, Malik OP. Load Demand Forecasting. Electric Distribution Systems. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2011. p. 33-65.

[23] Penya YK, Borges CE, Fernández I. Short-term load forecasting in non-residential buildings. Africon. Livingstone, Zambia: IEEE; 2011. p. 1-6.

[24] Zhao HX, Magoulès F. A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews. 2012;16(6):3586-92.

[25] Suganthi L, Samuel AA. Energy models for demand forecasting—A review. Renewable and Sustainable Energy Reviews. 2012;16(2):1223-40.

[26] Hahn H, Meyer-Nieberg S, Pickl S. Electric load forecasting methods: Tools for decision making. European Journal of Operational Research. 2009;199(3):902-7.

[27] Crawley DB, Hand JW, Kummert M, Griffith BT. Contrasting the capabilities of building energy performance simulation programs. Building and Environment. 2008;43(4):661-73.

[28] Yezioro A, Dong B, Leite F. An applied artificial intelligence approach towards assessing building performance simulation tools. Energy and Buildings. 2008;40(4):612-20.

[29] Fels MF. PRISM: an introduction. Energy and Buildings. 1986;9(1-2):5-18.

[30] Flouquet F. Local weather correlations and bias in building parameter estimates from energy-signature models. Energy and Buildings. 1992;19(2):113-23.

[31] Hadley DL. Daily variations in HVAC system electrical energy consumption in response to different weather conditions. Energy and Buildings. 1993;19(3):235-47.

[32] Wand MP, Jones MC. Kernel smoothing. London, UK: Chapman & Hall, 1994.

[33] Rosenblatt M. Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics. 1956;27(3):832-7.

[34] Parzen E. On estimation of a probability density function and mode. The annals of mathematical statistics. 1962;33(3):1065-76.

[35] Brown M, Barrington-Leigh C, Brown Z. Kernel regression for real-time building energy analysis. Journal of Building Performance Simulation. 2011;5(4):263–76.

[36] Kramer O, Satzger B, Lässig J. Power Prediction in Smart Grids with Evolutionary Local Kernel Regression. Hybrid Artificial Intelligence Systems: Springer Berlin; 2010. p. 262-9.

[37] De Gooijer JG, Hyndman RJ. 25 years of time series forecasting. International journal of forecasting. 2006;22(3):443-73.

[38] Cuaresma JC, Hlouskova J, Kossmeier S, Obersteiner M. Forecasting electricity spot-prices using linear univariate time-series models. Applied Energy. 2004;77(1):87-106.

[39] Abdel-Aal R, Al-Garni A. Forecasting monthly electric energy consumption in eastern Saudi Arabia using univariate time-series analysis. Energy. 1997;22(11):1059-69.

[40] Darbellay GA, Slama M. Forecasting the short-term demand for electricity: Do neural networks stand a better chance? International Journal of Forecasting. 2000;16(1):71-83.

[41] Taylor JW, McSharry PE, Buizza R. Wind power density forecasting using ensemble predictions and time series models. IEEE Transactions on Energy Conversion. 2009;24(3):775-82.

[42] Dhar A, Reddy T, Claridge D. A Fourier series model to predict hourly heating and cooling energy use in commercial buildings with outdoor temperature as the only weather variable. Journal of solar energy engineering. 1999;121(1):47-53.

[43] Kalman RE. A new approach to linear filtering and prediction problems. Journal of basic Engineering. 1960;82(1):35-45.

[44] Markoulakis SA, Stavrakakis GS, Nikolaou TG. Short-term load forecasting based on the Kalman filter and the neural-fuzzy network (ANFIS). International Conference on Energy & Environmental Systems. Chalkida, Greece2006. p. 189-93.

[45] Dounis AI. Artificial intelligence for energy conservation in buildings. Advances in Building Energy Research. 2010;4(1):267-99.

[46] Krarti M. An overview of artificial intelligence-based methods for building energy systems. Journal of solar energy engineering. 2003;125(3):331-42.

[47] Kalogirou SA. Artificial neural networks in energy applications in buildings. International Journal of Low-Carbon Technologies. 2006;1(3):201-16.

[48] Zadeh LA. Fuzzy sets. Information and Control. 1965;8(3):338-53.

[49] Jang J-SR. ANFIS: adaptive-network-based fuzzy inference system. Systems, Man and Cybernetics, IEEE Transactions on. 1993;23(3):665-85.

[50] Booker LB, Goldberg DE, Holland JH. Classifier systems and genetic algorithms. Artificial intelligence. 1989;40(1):235-82.

[51] Čongradac V, Kulić F. Recognition of the importance of using artificial neural networks and genetic algorithms to optimize chiller operation. Energy and Buildings. 2012;47:651-8.

[52] Rosenblatt F. Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms. Buffalo, NY, US: DTIC Document; 1961.

[53] Yang J, Rivard H, Zmeureanu R. On-line building energy prediction using adaptive artificial neural networks. Energy and Buildings. 2005;37(12):1250-9.

[54] González PA, Zamarreño JM. Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. Energy and Buildings. 2005;37(6):595-601.

[55] Escrivá-Escrivá G, Roldán-Blay C. New artificial neural network prediction method for electrical consumption forecasting based on building end-uses. Energy and Buildings. 2011;43(11):3112-9.

[56] Karatasou S, Santamouris M, Geros V. Modeling and predicting building's energy use with artificial neural networks: Methods and results. Energy and Buildings. 2006;38(8):949-58.

[57] Vapnik VN. The nature of statistical learning theory. New York: Springer Science & Business Media, 1995.

[58] Li Q, Meng Q, Cai J, Yoshino H, Mochida A. Applying support vector machine to predict hourly cooling load in the building. Applied Energy. 2009;86(10):2249-56.

[59] Li Q, Meng Q, Cai J, Yoshino H, Mochida A. Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. Energy Conversion and Management. 2009;50(1):90-6.

[60] Xuemei L, Jin-hu L, Lixing D, Gang X, Jibin L. Building Cooling Load Forecasting Model Based on LS-SVM. Asia-Pacific Conference on Information Processing: IEEE; 2009. p. 55-8.

[61] Pai PF, Hong WC. Support vector machines with simulated annealing algorithms in electricity load forecasting. Energy Conversion and Management. 2005;46(17):2669-88.

[62] Dong B, Cao C, Lee SE. Applying support vector machines to predict building energy consumption in tropical region. Energy and Buildings. 2005;37(5):545-53.

[63] He W. Forecasting electricity load with optimized local learning models. International Journal of Electrical Power & Energy Systems. 2008;30(10):603-8.

[64] Niu D, Wang Y, Wu DD. Power load forecasting using support vector machine and ant colony optimization. Expert Systems with Applications. 2010;37(3):2531-9.

[65] Yang X. Comparison of the LS-SVM based load forecasting models. International Conference on Electronic & Mechanical Engineering and Information Technology. Harbin, Heilongjiang, China: IEEE; 2011. p. 2942-5.

[66] Kusiak A, Li M. Cooling output optimization of an air handling unit. Applied Energy. 2010;87(3):901-9.

[67] Braun JE, Chaturvedi N. An Inverse Gray-Box Model for Transient Building Load Prediction. HVAC&R Research. 2002;8(1):73-99.

[68] Zhou Q, Wang S, Xu X, Xiao F. A grey-box model of next-day building thermal load prediction for energy-efficient control. International Journal of Energy Research. 2008;32(15):1418-31.

[69] Wang S, Xu X. Simplified building model for transient thermal performance estimation using GA-based parameter identification. International Journal of Thermal Sciences. 2006;45(4):419-32.

[70] Bourdouxhe JPH, Saavedra C, Grodent M, Silva KL, Lebrun JJ. Toolkit for primary HVAC system energy calculation part II: reciprocating chiller models. ASHRAE Transactions. 1994;100(2):774-86.

[71] Lemort V, Lebrun J, Felsmann C. Testing and validation of simulation tools of HVAC mechanical equipment including their control strategies part III: validation of an air-cooled chiller model. Eleventh International IBPSA Conference. Glasgow, Scotland: IBPSA; 2009.

[72] Li K, Su H. Forecasting building energy consumption with hybrid genetic algorithm-hierarchical adaptive network-based fuzzy inference system. Energy and Buildings. 2010;42(11):2070-6.

[73] Li K, Su H, Chu J. Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: A comparative study. Energy and Buildings. 2011;43(10):2893-9.

[74] Swan LG, Ugursal VI, Beausoleil-Morrison I. Occupant related household energy consumption in Canada: Estimation using a bottom-up neural-network technique. Energy and Buildings. 2011;43(2–3):326-37.

[75] Nie H, Liu G, Liu X, Wang Y. Hybrid of ARIMA and SVMs for Short-Term Load Forecasting. Energy Procedia. 2012;16(Part C):1455-60.

[76] Kavousi-Fard A, Kavousi-Fard F. A new hybrid correction method for short-term load forecasting based on ARIMA, SVR and CSA. Journal of Experimental & Theoretical Artificial Intelligence. 2013;25(4):559-74.

[77] Hong WC. Hybrid evolutionary algorithms in a SVR-based electric load forecasting model. International Journal of Electrical Power & Energy Systems. 2009;31(7):409-17.

[78] Wu Q. A hybrid-forecasting model based on Gaussian support vector machine and chaotic particle swarm optimization. Expert Systems with Applications. 2010;37(3):2388-94.

[79] Le Cam M, Zmeureanu R, Daoud A. Comparison of inverse models used for the forecast of the electric demand of chillers. Proceedings of Building Simulation 2013. Chambery, France: IBPSA; 2013. p. 2044-51.

[80] Agarwal V, Bougaev A, Tsoukalas L, Kollias S, Stafylopatis A, Duch W, et al. Kernel Regression Based Short-Term Load Forecasting. In: 2006 ANNI, editor. Lecture Notes in Computer Science: Springer Berlin / Heidelberg; 2006. p. 701-8.

[81] Asber D, Lefebvre S, Asber J, Saad M, Desbiens C. Non-parametric short-term load forecasting. International Journal of Electrical Power & Energy Systems. 2007;29(8):630-5.

[82] Amaral LF, Souza RC, Stevenson M. A smooth transition periodic autoregressive (STPAR) model for short-term load forecasting. International Journal of Forecasting. 2008;24(4):603-15.

[83] Espinoza M, Joye C, Belmans R, DeMoor B. Short-term load forecasting, profile identification, and customer segmentation: a methodology based on periodic time series. IEEE Transactions on Power Systems. 2005;20(3):1622-30.

[84] Soares LJ, Souza LR. Forecasting electricity demand using generalized long memory. International Journal of Forecasting. 2006;22(1):17-28.

[85] Al-Hamadi H, Soliman S. Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model. Electric Power Systems Research. 2004;68(1):47-59.

[86] Monfet D, Zmeureanu R. Ongoing commissioning of water-cooled electric chillers using benchmarking models. Applied Energy. 2012;92:99-108.

[87] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011;2(3):27.

[88] Jurado S, Nebot À, Mugica F, Avellana N. Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques. Energy. 2015;86:276–91.

[89] Powell KM, Sriprasad A, Cole WJ, Edgar TF. Heating, cooling, and electrical load forecasting for a large-scale district energy system. Energy. 2014;74:877-85.

[90] Ko C-N, Lee C-M. Short-term load forecasting using SVR (support vector regression)-based radial basis function neural network with dual extended Kalman filter. Energy. 2013;49:413-22.

[91] Che J, Wang J, Wang G. An adaptive fuzzy combination model based on self-organizing map and support vector regression for electric load forecasting. Energy. 2012;37(1):657-64.

[92] Ghofrani M, Ghayekhloo M, Arabali A, Ghayekhloo A. A hybrid short-term load forecasting with a new input selection framework. Energy. 2015;81:777-86.

[93] Touretzky CR, Patil R. Building-level power demand forecasting framework using building specific inputs: Development and applications. Applied Energy. 2015;147:466-77.

[94] Ninagawa C, Kondo S, Isozumi S, Yoshida H. Fine-time-granularity fast demand control of building HVAC facilities for future smart grid. Innovative Smart Grid Technologies (ISGT Europe), 2012 3rd IEEE PES International Conference and Exhibition on. IEEE, 2012, p. 1-6.

[95] Callaway DS. Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy. Energy Conversion and Management. 2009;50(5):1389-400.

[96] Sinitsyn NA, Kundu S, Backhaus S. Safe protocols for generating power pulses with heterogeneous populations of thermostatically controlled loads. Energy Conversion and Management. 2013;67:297-308.

[97] Goddard G, Klose J, Backhaus S. Model development and identification for fast demand response in commercial HVAC systems. Smart Grid, IEEE Transactions on. 2014;5(4):2084-92.

[98] Olivieri SJ, Henze GP, Corbin CD, Brandemuehl MJ. Evaluation of commercial building demand response potential using optimal short-term curtailment of heating, ventilation, and air-conditioning loads. Journal of Building Performance Simulation. 2014;7(2):100-18.

[99] Grant J, Eltoukhy M, Asfour S. Short-Term Electrical Peak Demand Forecasting in a Large Government Building Using Artificial Neural Networks. Energies. 2014;7(4):1935-53.

[100] Leduc M-A, Lavigne Jr K, Daoud A, Poulin A. Demand Reponse in Commercial Buildings: a Cold Climate Field Study. ASHRAE Transactions. 2015;121(2):1-8.

[101] Taieb SB, Sorjamaa A, Bontempi G. Multiple-output modeling for multi-step-ahead time series forecasting. Neurocomputing. 2010;73(10):1950-7.

[102] Reddy TA. Applied data analysis and modeling for energy engineers and scientists. New York, NY: Springer Science & Business Media, 2011.

[103] Mihai A. Calibration of a Building Energy Model Using Measured Data for a Research Center. Montreal, QC: Concordia University, 2014.

[104] Zibin N. A Bottom-Up Method to Calibrate Building Energy Models Using Building Automation System (BAS) Trend Data. Montreal, QC: Concordia University, 2014.

[105] Monfet D. New ongoing commissioning approach of central plants: methodology and case study. Montreal, QC: Concordia University, 2011.

[106] Han J, Kamber M. Data mining: concepts and techniques: Morgan Kaufmann, 2006.

[107] Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research. 2003;3:1157-82.

[108] Fagiani M, Squartini S, Gabrielli L, Spinsante S, Piazza F. A review of datasets and load forecasting techniques for smart natural gas and water grids: Analysis and experiments. Neurocomputing. 2015;170:448-65.

[109] Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression. Neural networks. 2004;17(1):113-26.

[110] Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and computing. 2004;14(3):199-222.

[111] ASHRAE. Standard 90.1-2004, Energy standard for buildings except low rise residential buildings. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. 2004.

[112] EnergyPlus. 6.0 Input/Output Reference: The Encyclopedic Reference to EnergyPlus Input and Output. US Department Of Energy. 2010.

[113] ASHRAE. Guideline 14-2002, Measurement of Energy and Demand Savings. American Society of Heating, Ventilating, and Air Conditioning Engineers, Atlanta, Georgia. 2002.

[114] Tukey JW. Exploratory Data Analysis. Reading, MA: Addison-Wesley Publishing Company, 1977.

[115] Costa A, Keane MM, Torrens JI, Corry E. Building operation and energy performance: Monitoring, analysis and optimisation toolkit. Applied Energy. 2013;101:310-6.

[116] Bezdek JC. Pattern recognition with fuzzy objective function algorithms: Kluwer Academic Publishers, 1981.

[117] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics. 1987;20:53-65.

[118] Chipperfield AJ, Fleming P, Pohlheim H. Genetic Algorithm Toolbox^{TM} User's Guide. University of Sheffield, Department of Automatic Control and Systems Engineering; 1994.

[119] Beale MH, Hagan MT, Demuth HB. Neural Network Toolbox^{TM} User's Guide. Natick, Massachusetts, United States: The MathWorks, Inc; 2014.

[120] Macas M, Moretti F, Fonti A, Giantomassi A, Comodi G, Annunziato M, et al. The role of data sample size and dimensionality in neural network based forecasting of building heating related variables. Energy and Buildings. 2016;111:299-310.

Appendix A. Forecast of the cooling coil load with sequence A

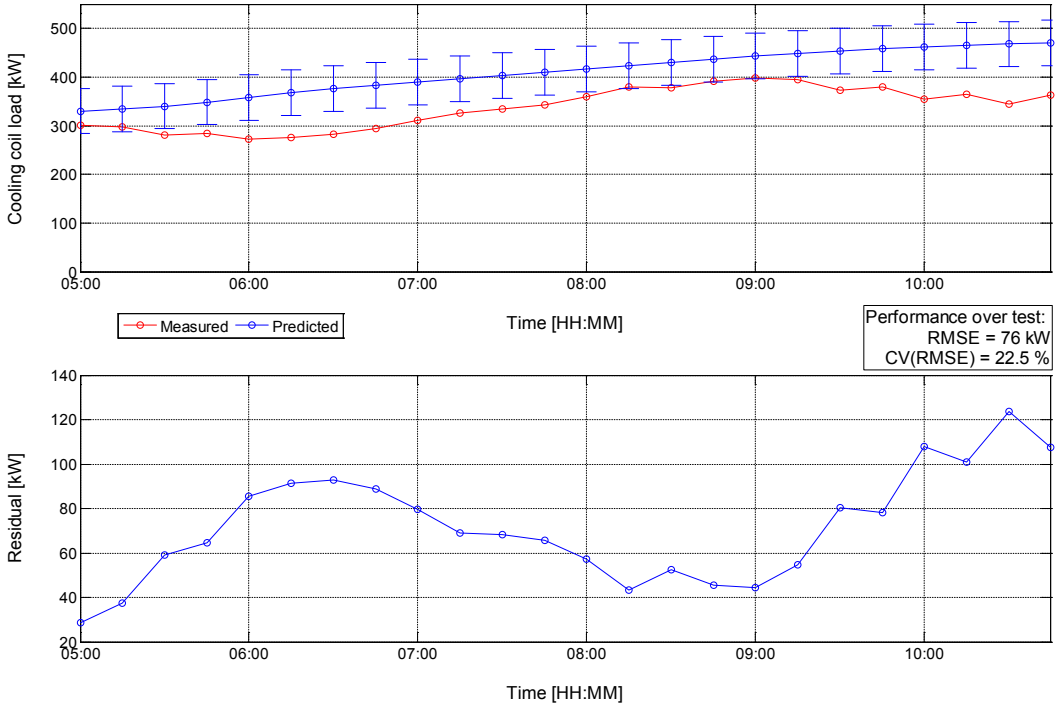

Figure A.1 Forecasted cooling coil load over the six-hour test set on 13th of July, based on cluster #1.

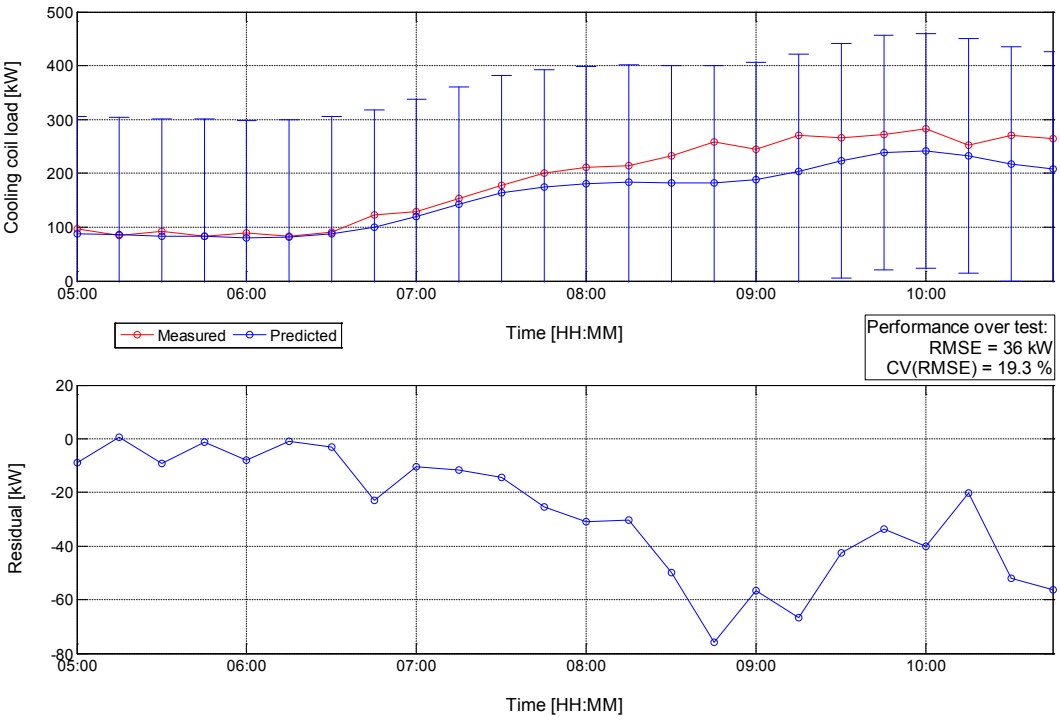

Figure A.2 Forecasted cooling coil load over the six-hour test set on 20th of August, based on cluster #2.

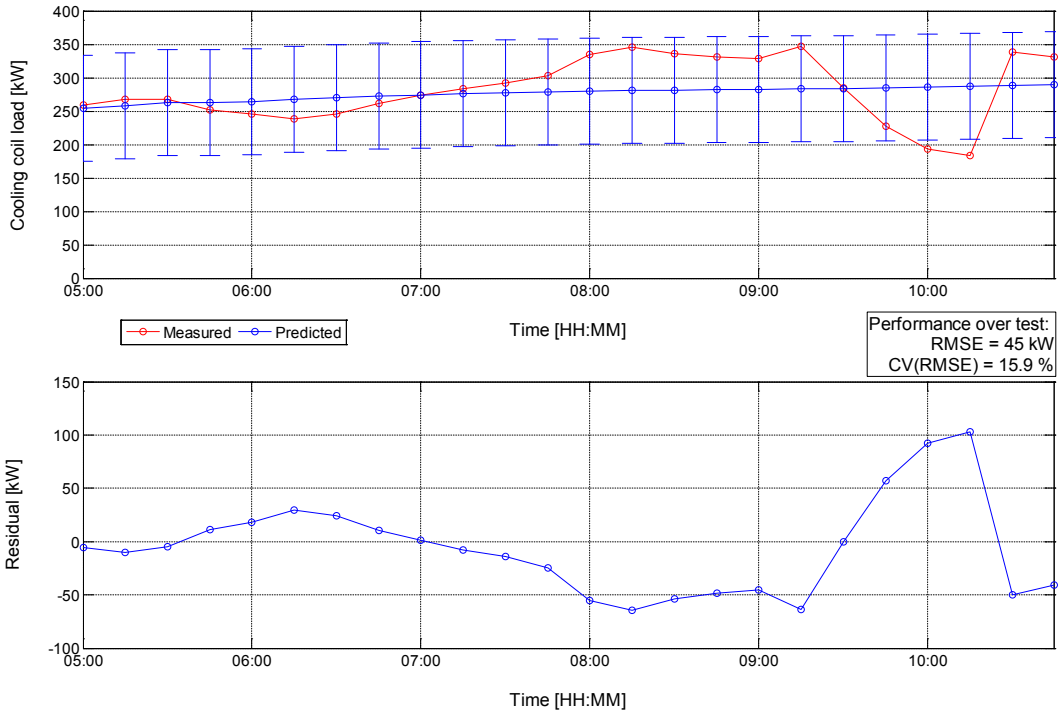

Figure A.3 Forecasted cooling coil load over the six-hour test set on 11th of August, based on cluster #3.

Appendix B. Forecast of the regressors of the cooling coil load with sequence A

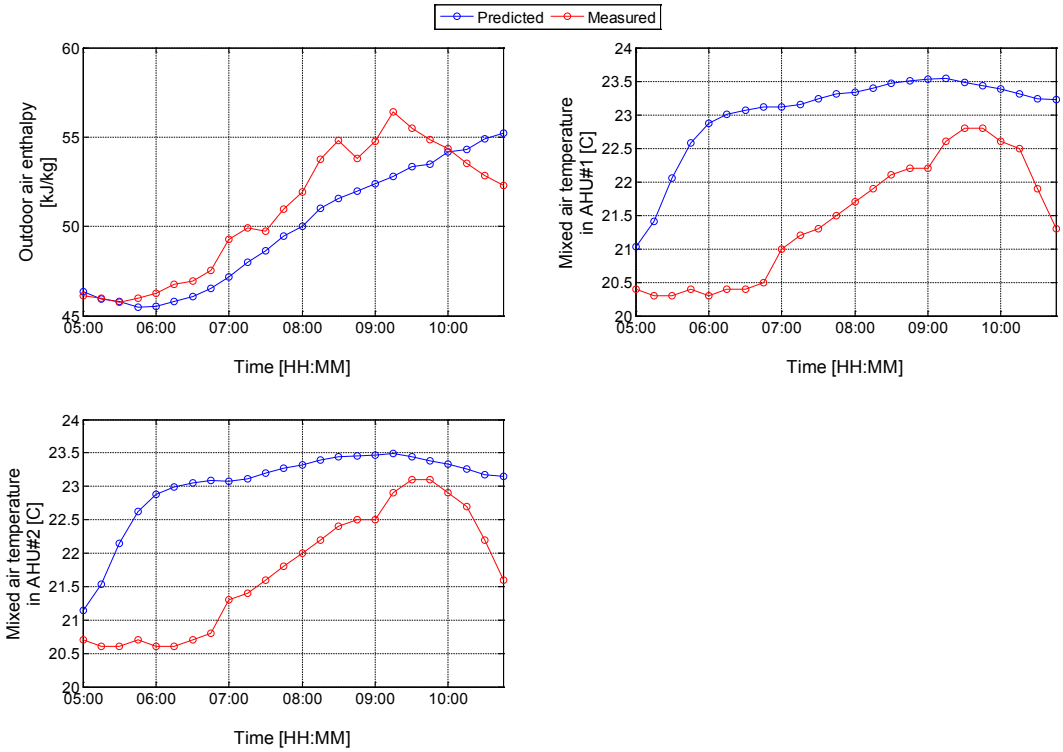

Figure B.1 Forecasted regressors over the six-hour test set on 13th of July, based on cluster #1.

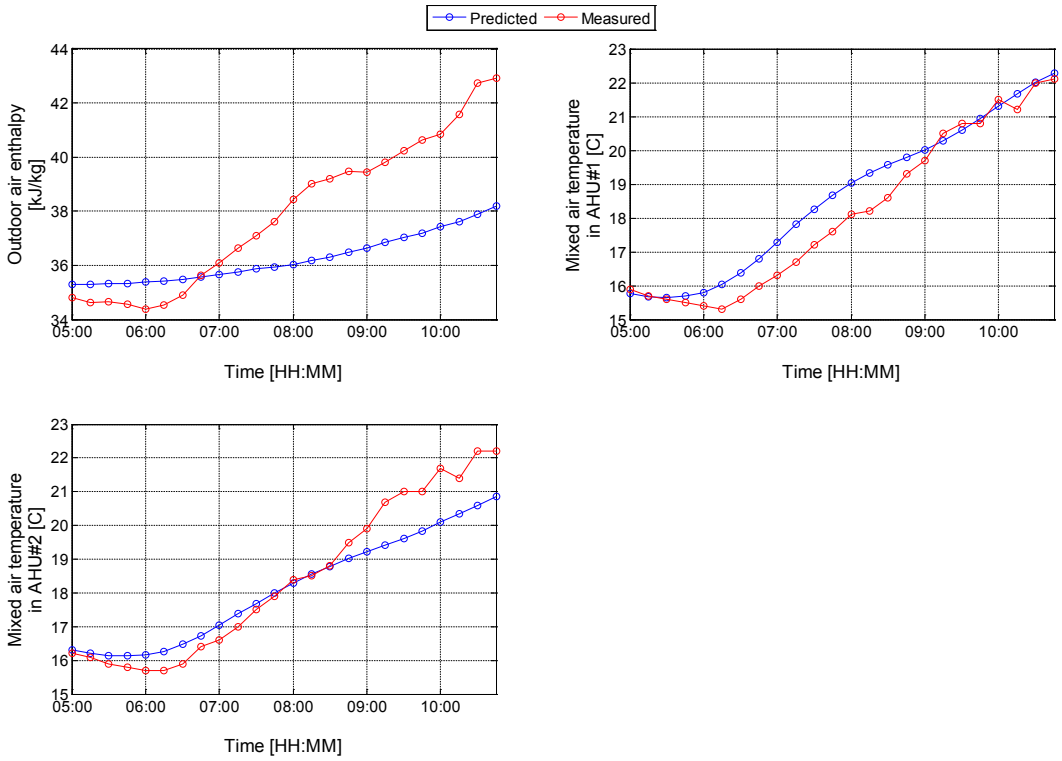

Figure B.2 Forecasted regressors over the six-hour test set on 20th of August, based on cluster #2.

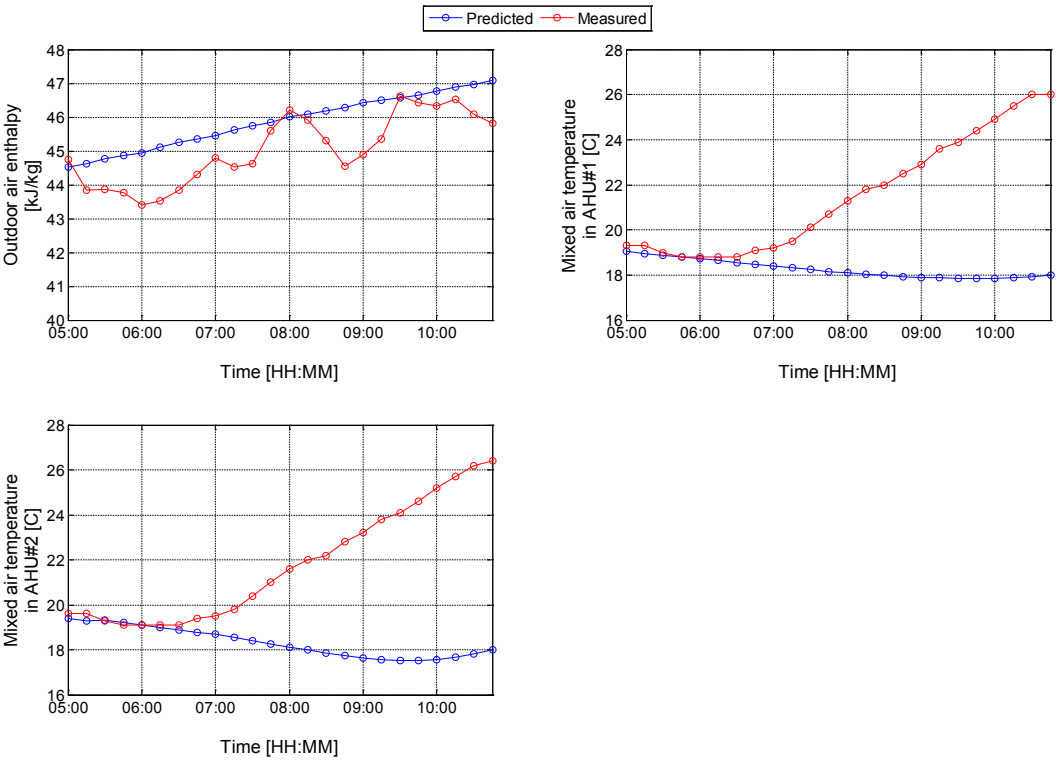

Figure B.3 Forecasted regressors the six-hour test set on 11th of August, based on cluster #3.

Appendix C. Forecast of the building cooling load with sequence A

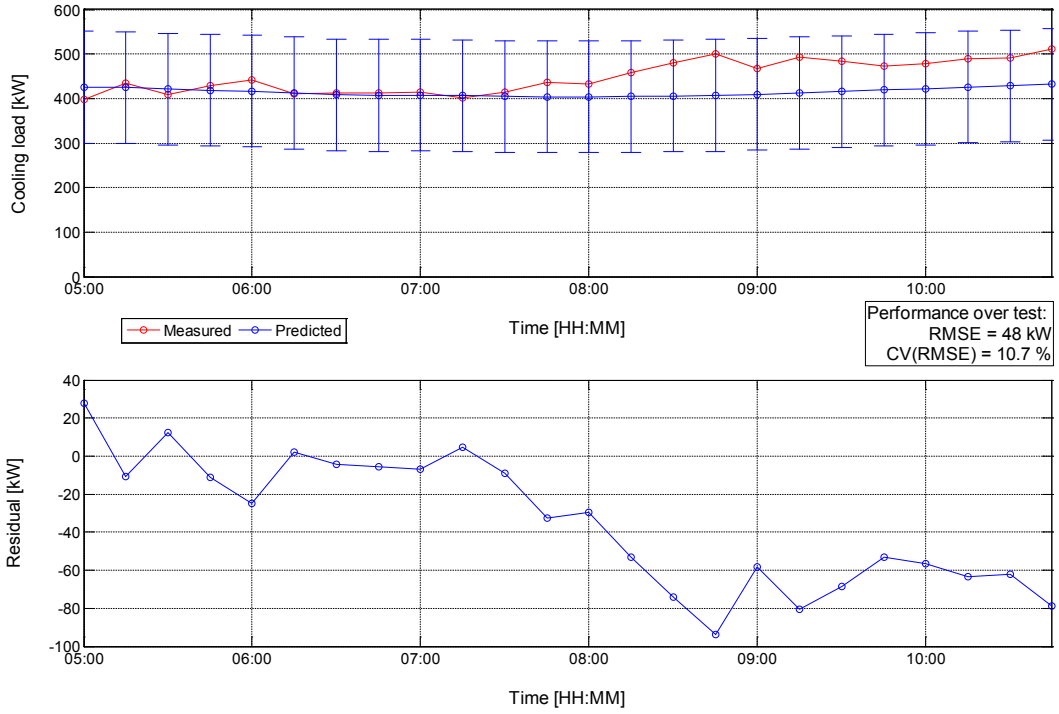

Figure C.1 Forecasted building cooling load over the six-hour test set on 12th of August, based on cluster #2.

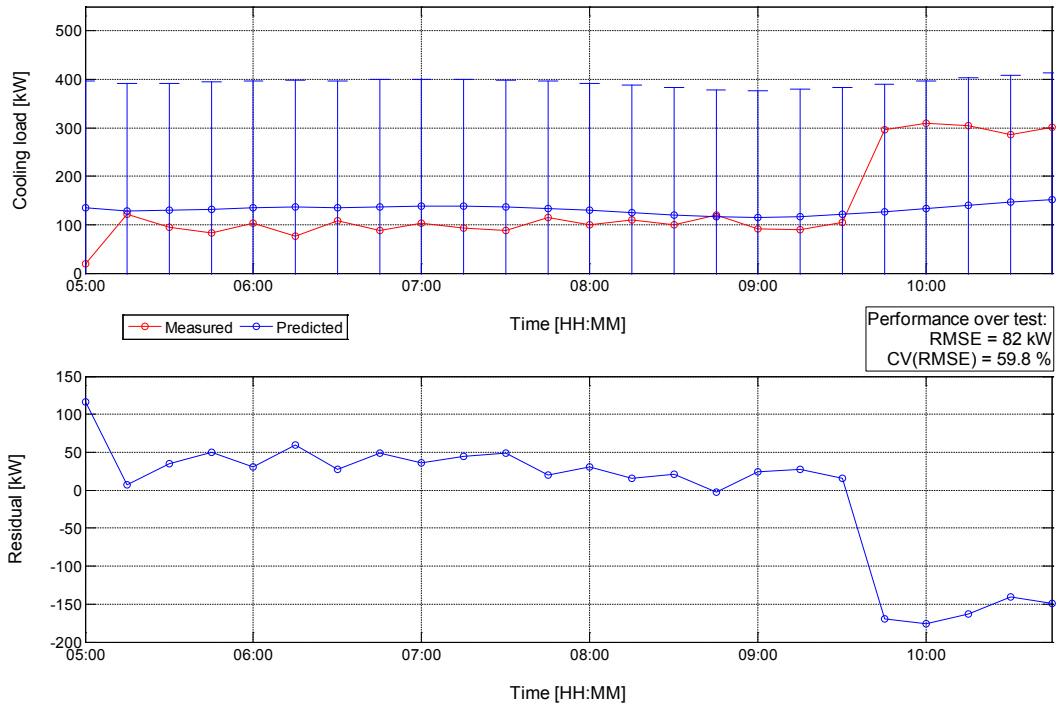

Figure C.2 Forecasted building cooling load over the six-hour test set on 29th of August, based on cluster #3.

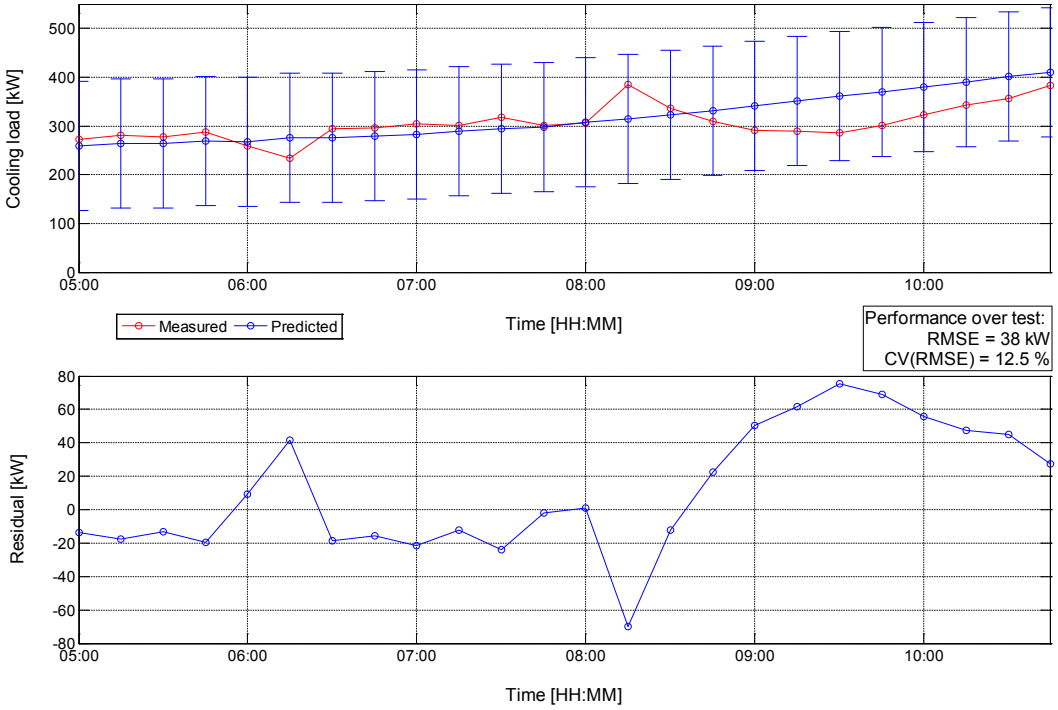

Figure C.3 Forecasted building cooling load over the six-hour test set on 31st of July, based on cluster #4.

Appendix D. Forecast of the regressors of the building cooling load with sequence A

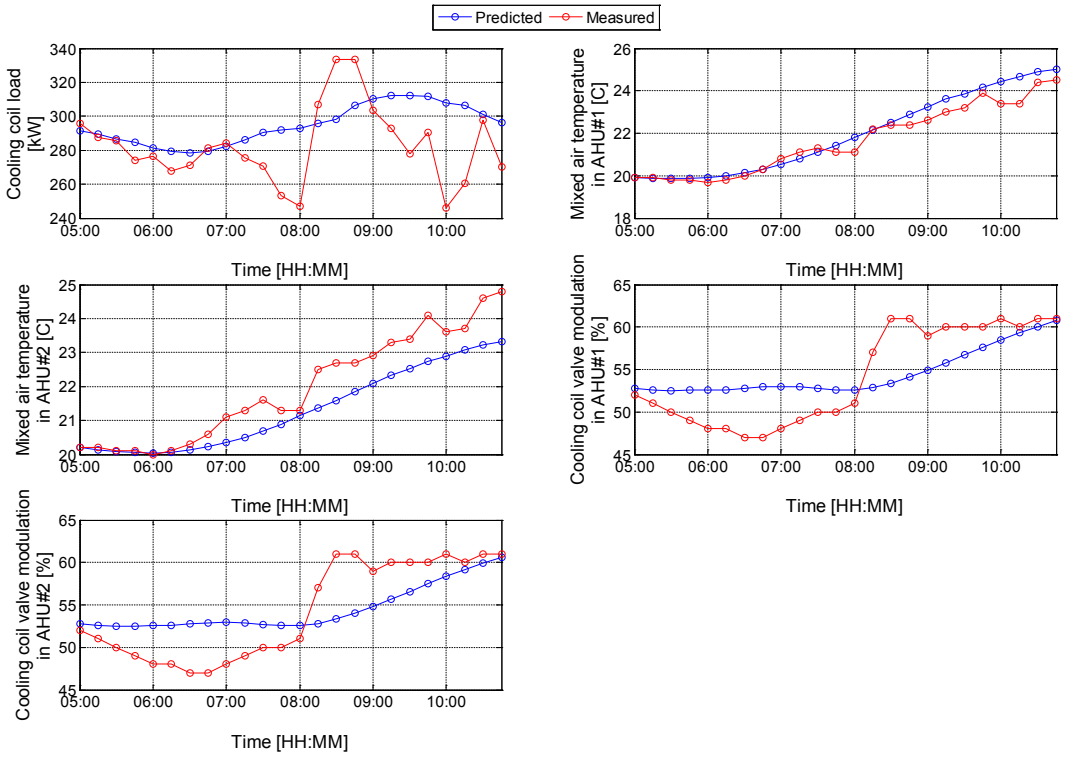

Figure D.1 Forecasted regressors over the six-hour test set on 12th of August, based on cluster #2.

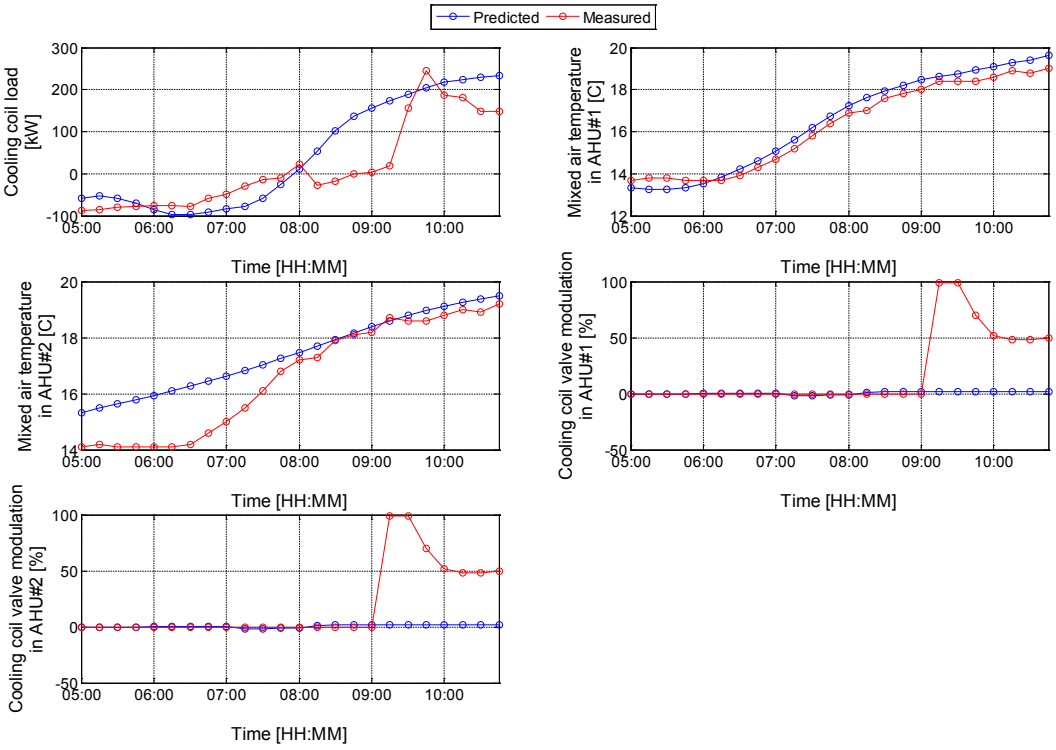

Figure D.2 Forecasted regressors over the six-hour test set on 29th of August, based on cluster #3.

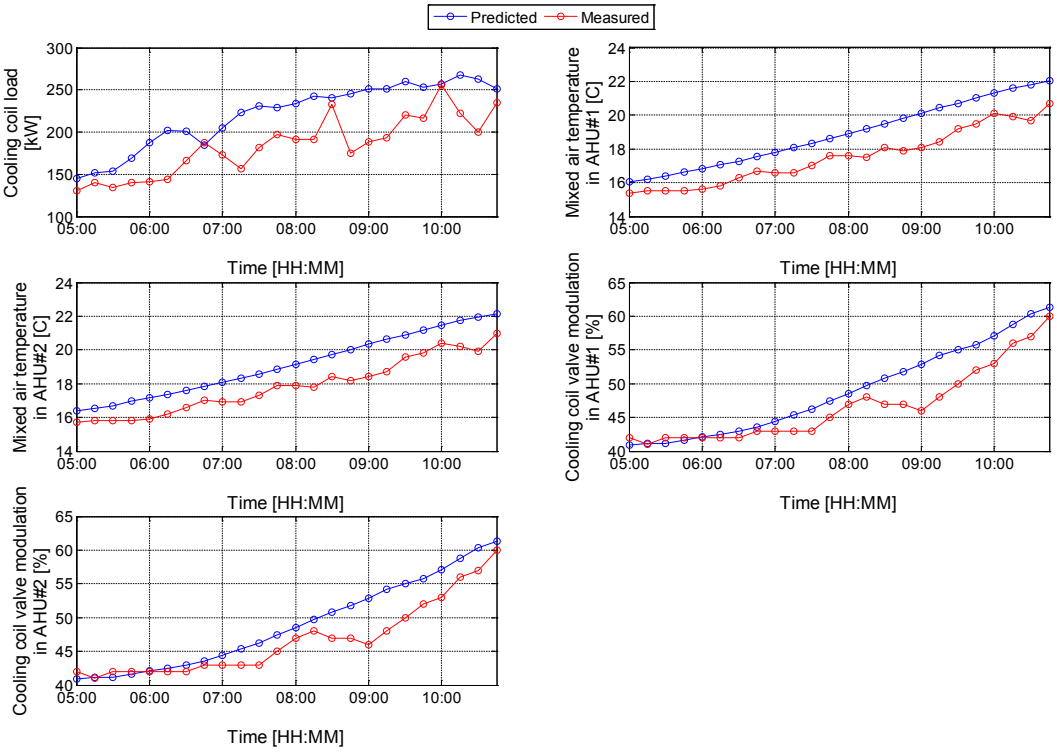

Figure D.3 Forecasted regressors over the six-hour test set on 31st of July, based on cluster #4.

Appendix E. Performance curves of fans in AHU of GE building

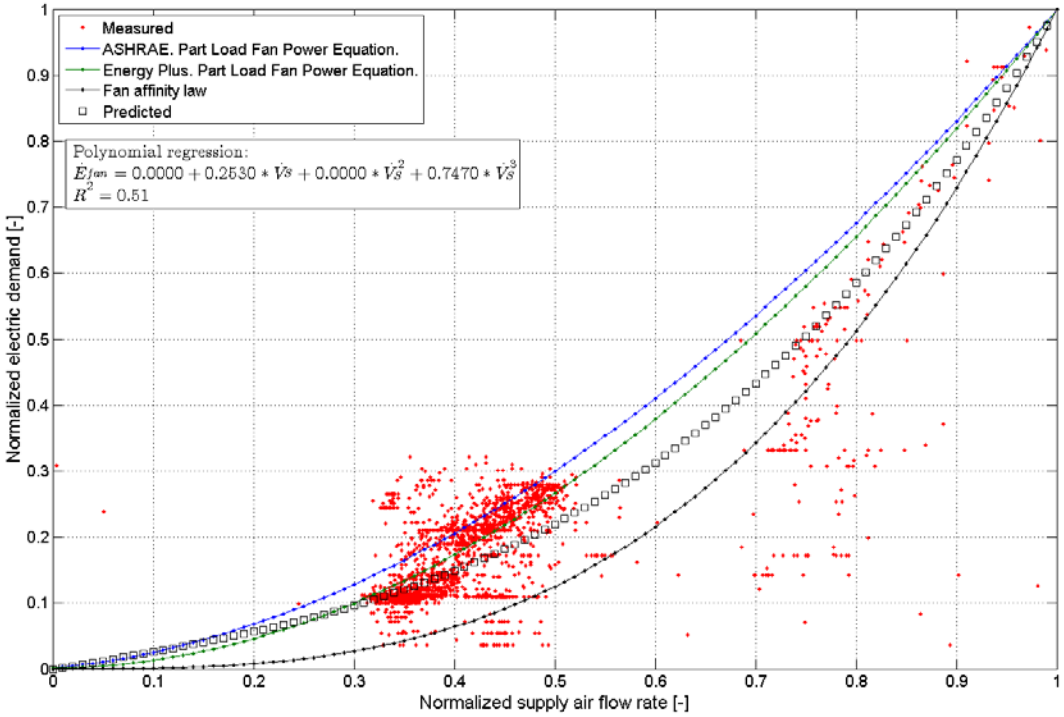

Figure E.1 Performance curve of fan #1

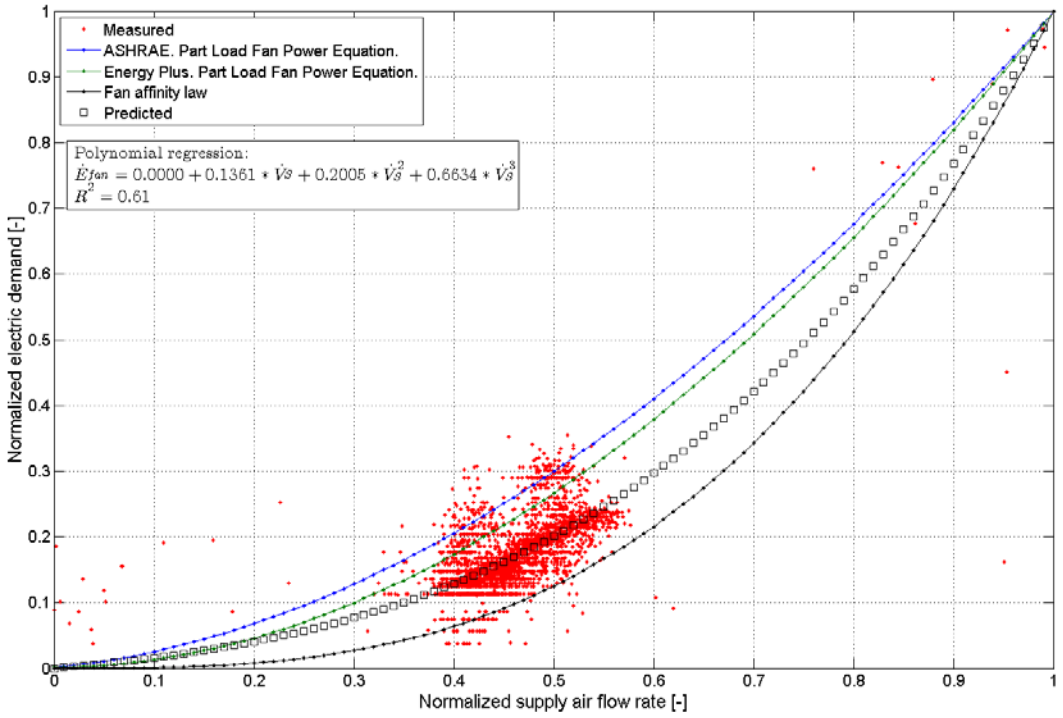

Figure E.2 Performance curve of fan #2

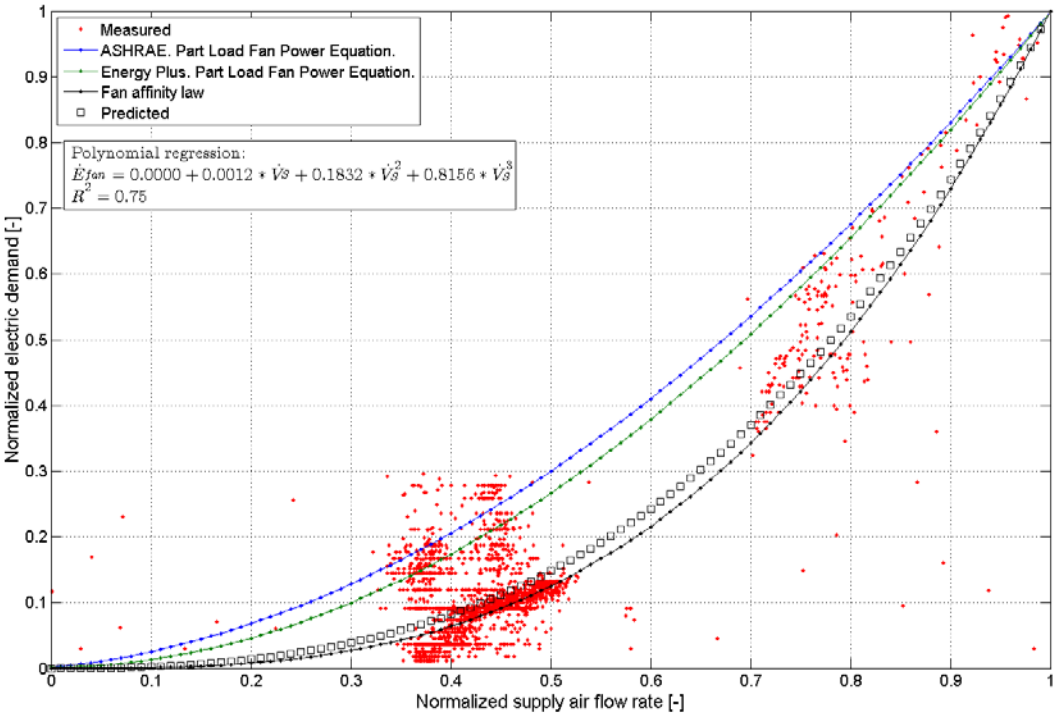

Figure E.3 Performance curve of fan #3

Appendix F. Forecast of the electric demand of the chillers with sequence A

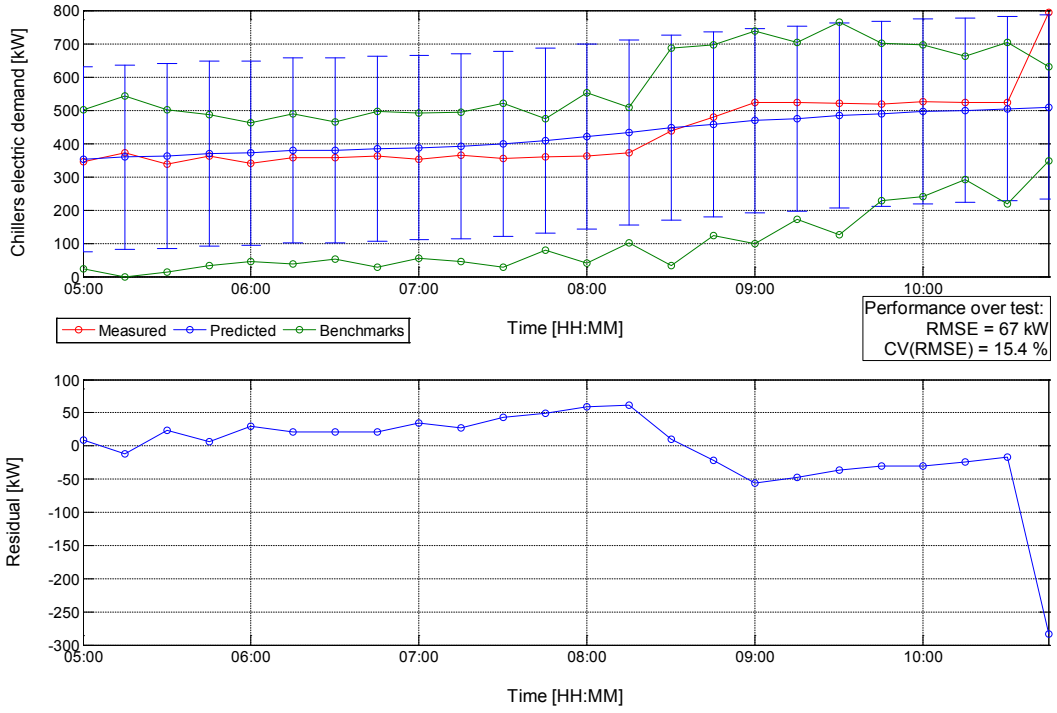

Figure F.1 Forecasted electric demand of the chillers over the six-hour test set on 25th of August, based on cluster #1.

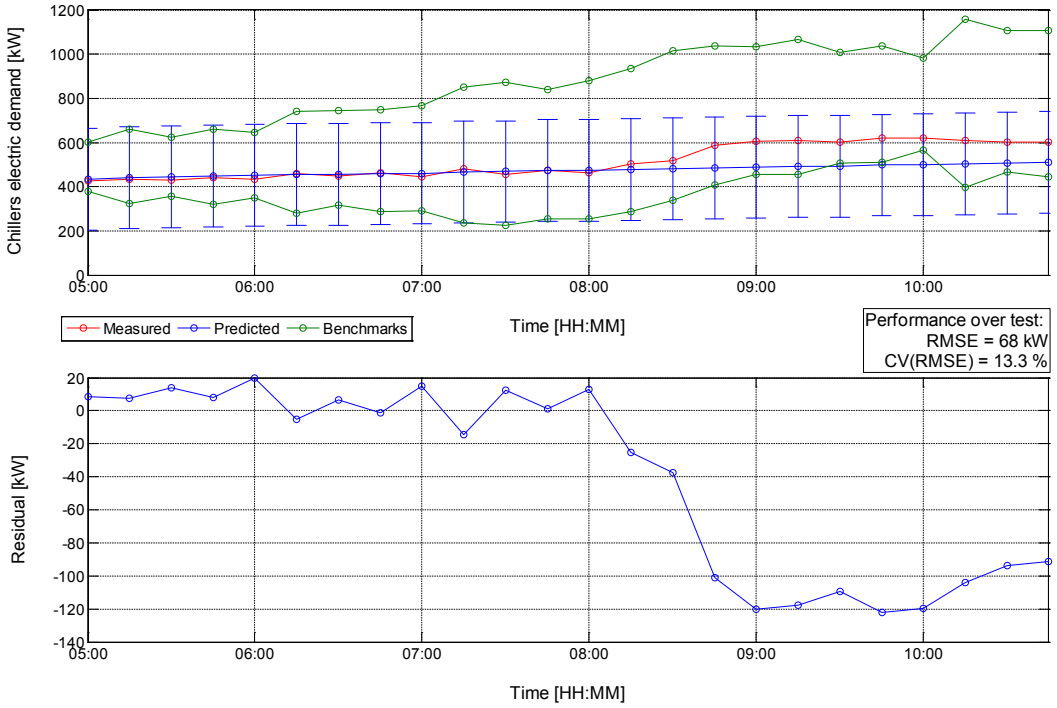

Figure F.2 Forecasted electric demand of the chillers over the six-hour test set on 13th of August, based on cluster #2.

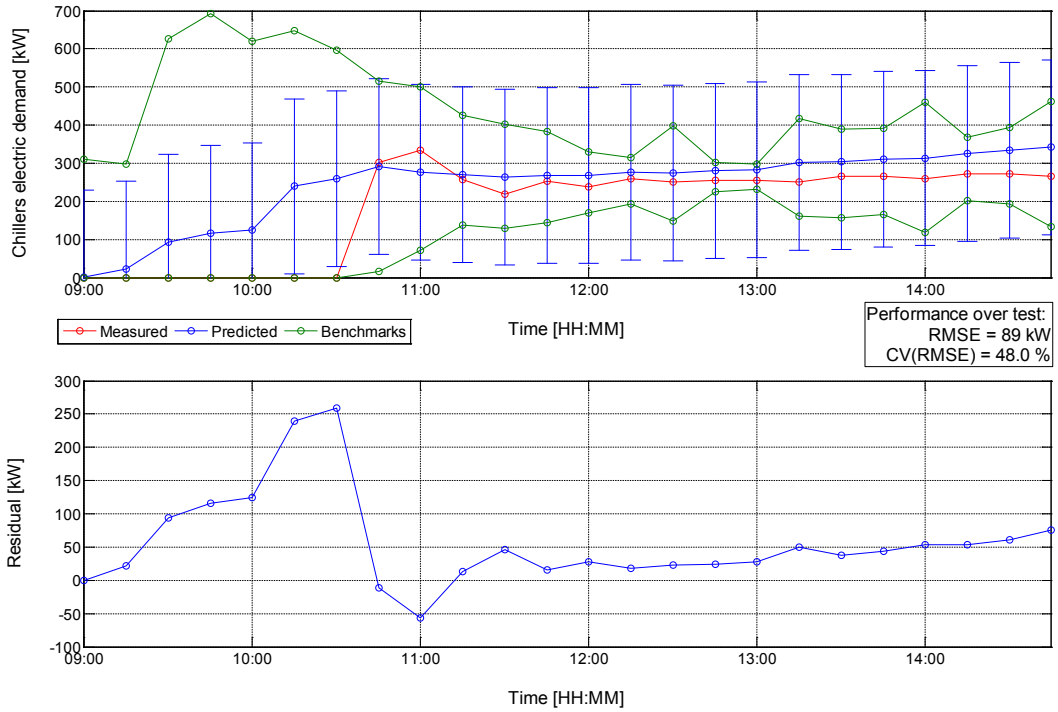

Figure F.3 Forecasted electric demand of the chillers over the six-hour test set on 17th of August, based on cluster #3.

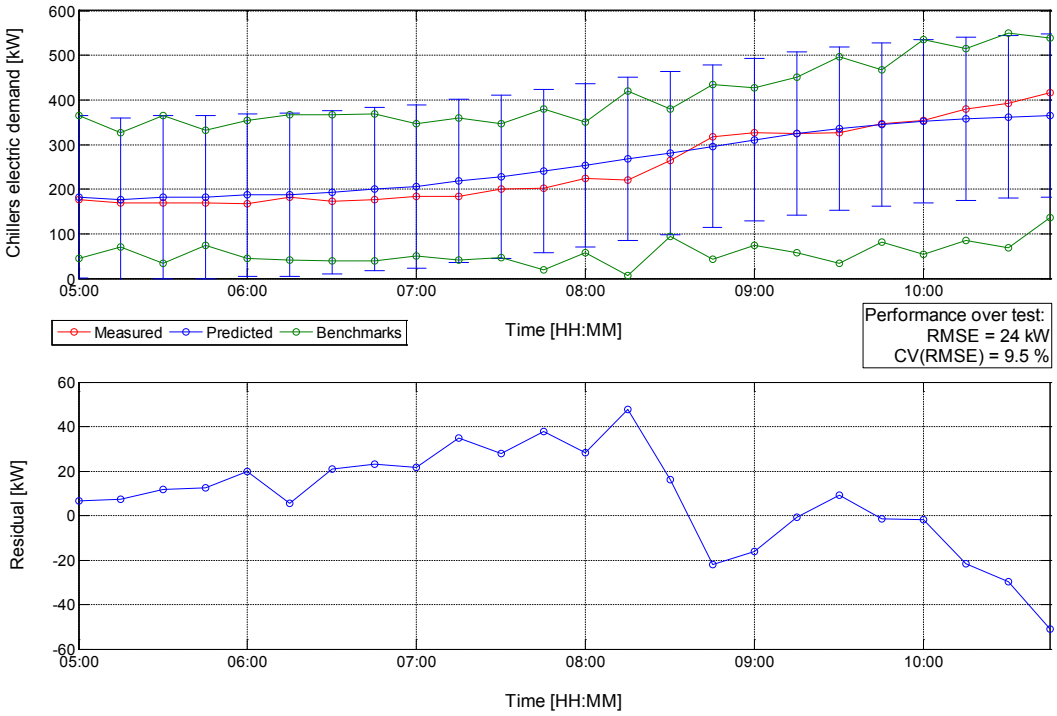

Figure F.4 Forecasted electric demand of the chillers over the six-hour test set on 11th of July, based on cluster #4.

Appendix G. Forecasted regressors of the electric demand of the chillers with sequence A

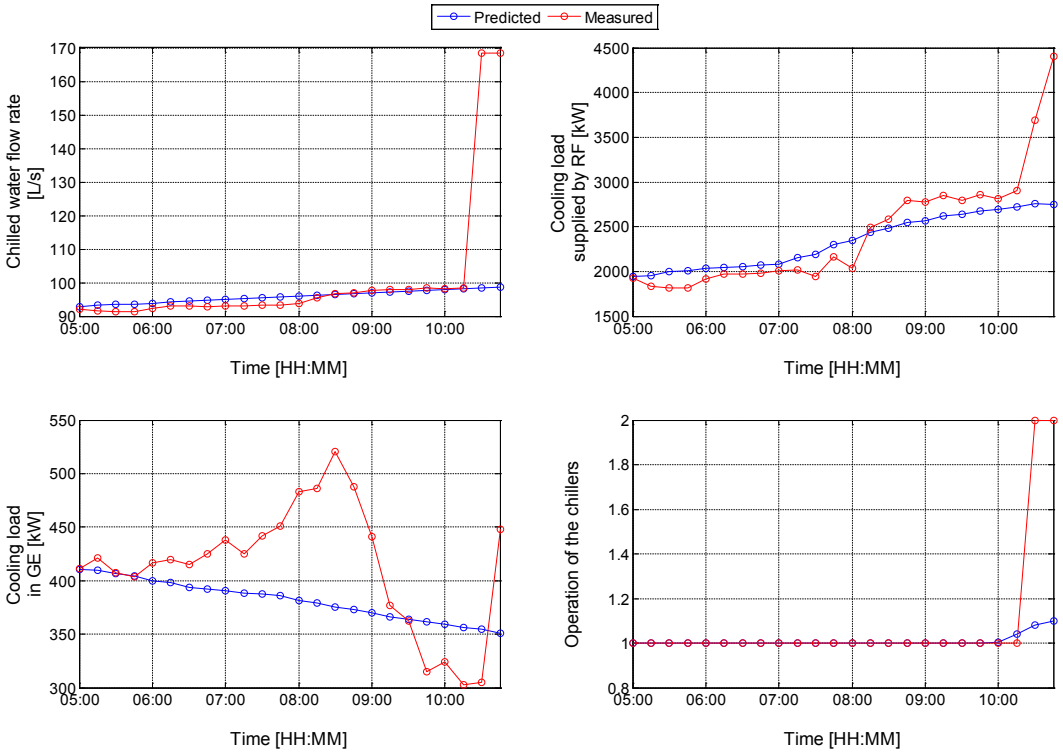

Figure G.1 Forecasted regressors over the six-hour test set on 25th of August, based on cluster #1.

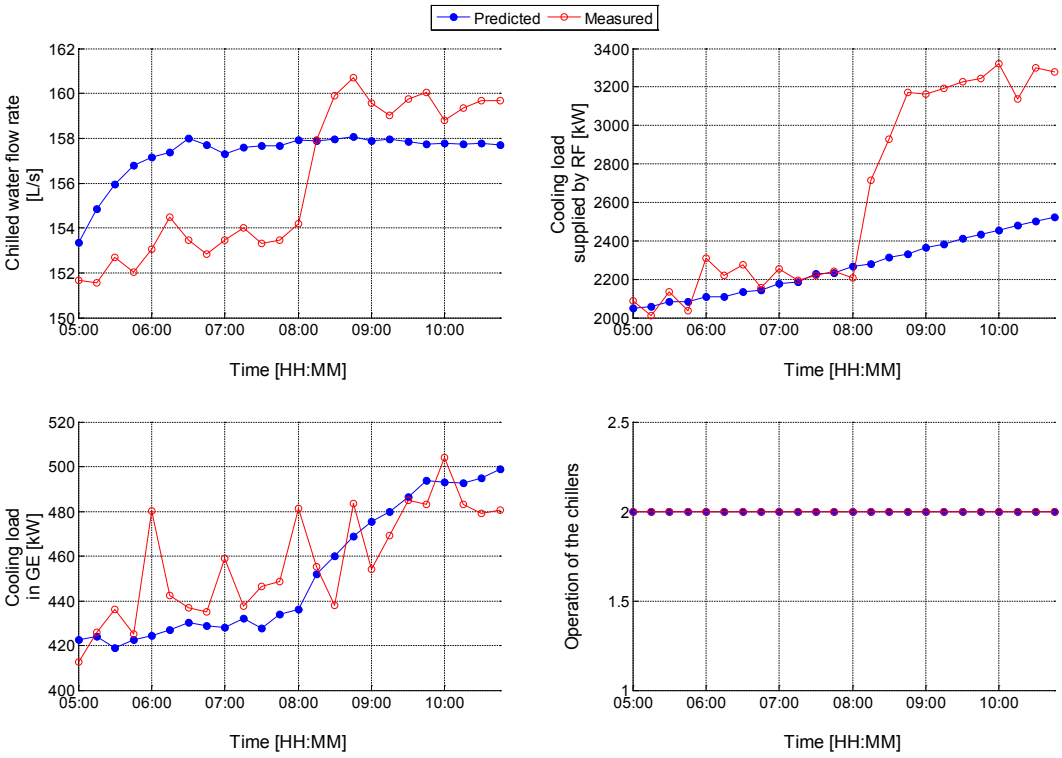

Figure G.2 Forecasted regressors over the six-hour test set on 13th of August, based on cluster #2.

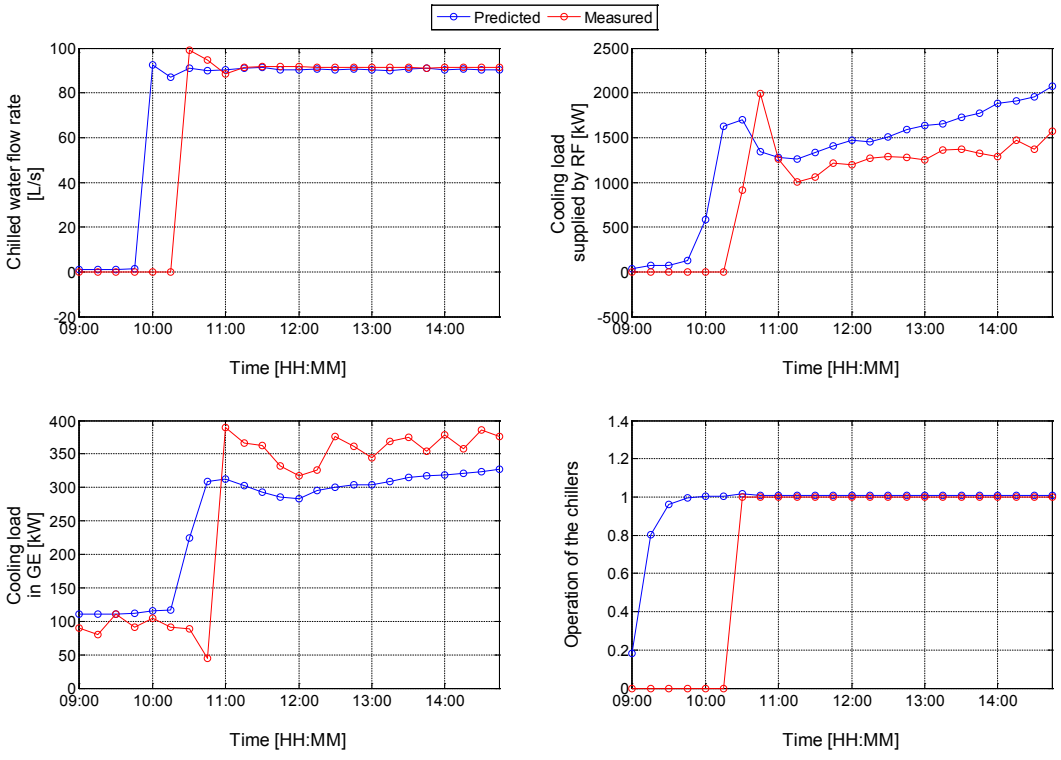

Figure G.3 Forecasted regressors over the six-hour test set on 17th of August, based on cluster #3.

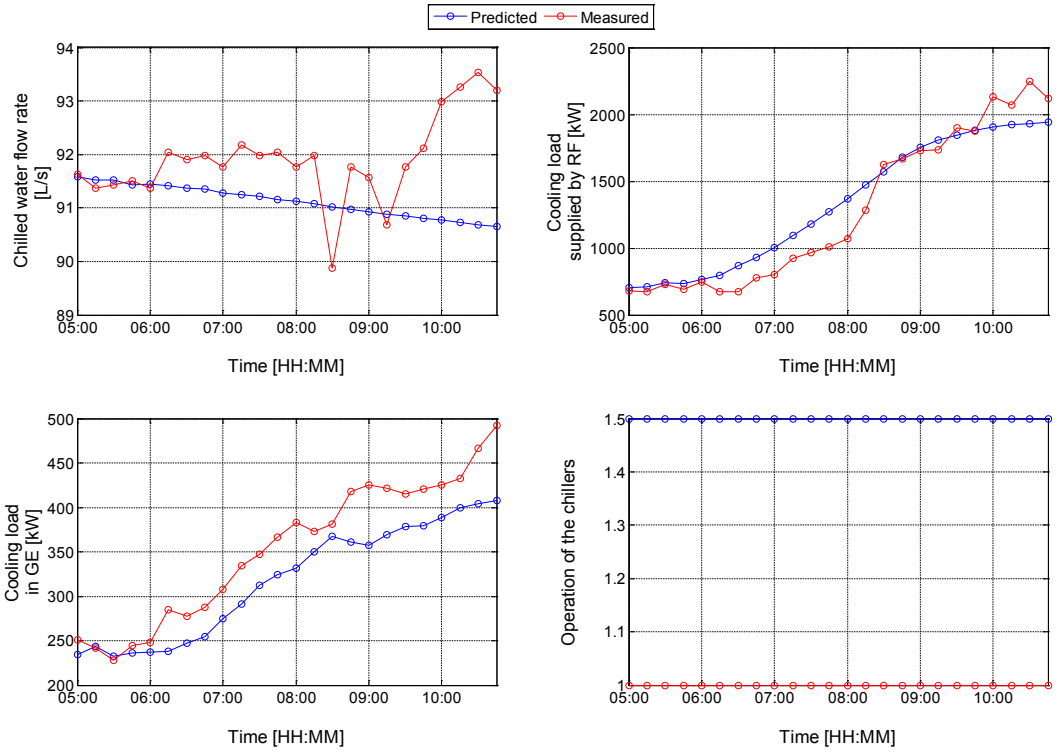

Figure G.4 Forecasted regressors over the six-hour test set on 11th of July, based on cluster #4.

Appendix H. Forecast of the electric demand of the cooling towers with sequence A

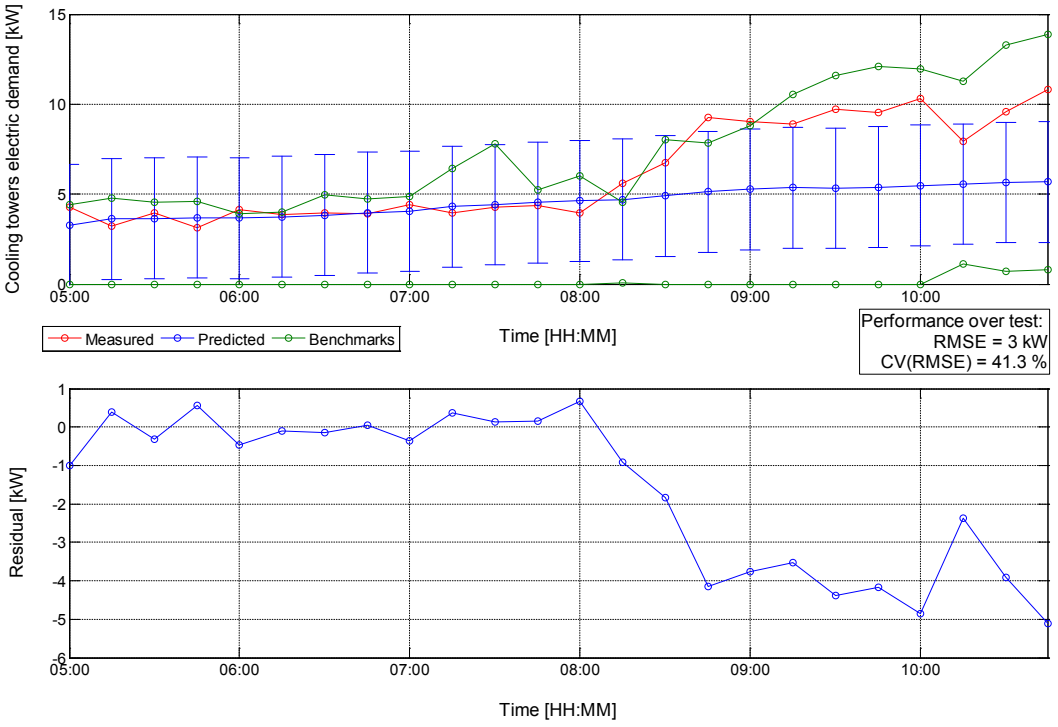

Figure H.1 Forecasted electric demand of the cooling towers over the six-hour test set on 13th of August, based on cluster #1.

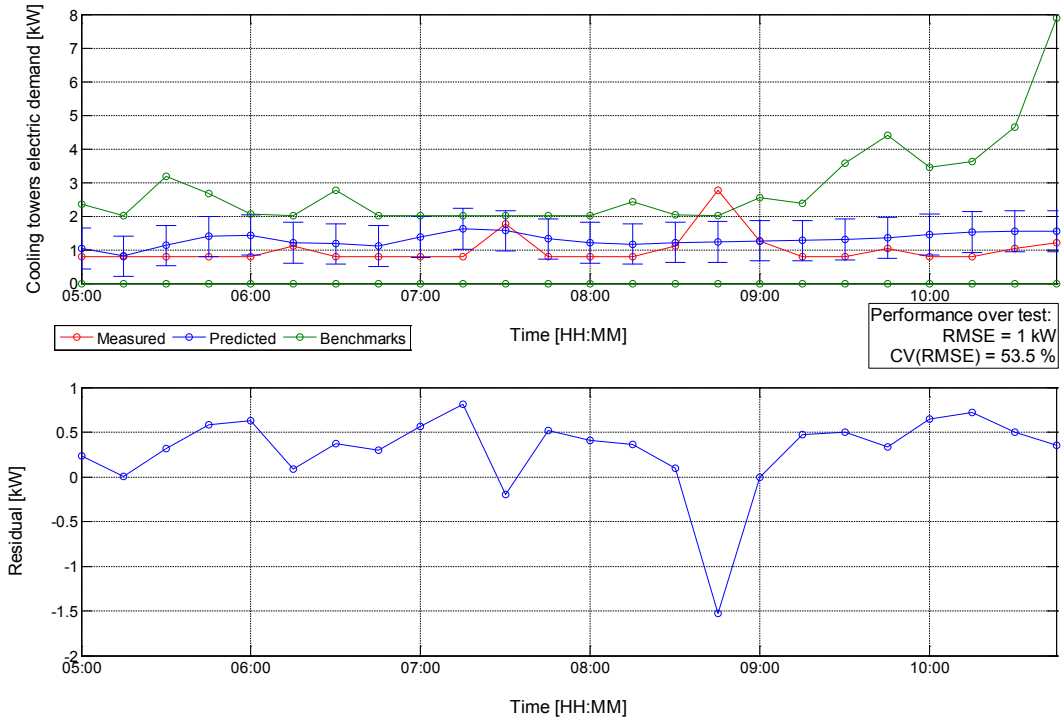

Figure H.2 Forecasted electric demand of the cooling towers over the six-hour test set on 19th of June, based on cluster #2.

Appendix I. Forecast of the regressors of the electric demand of the cooling towers with sequence A

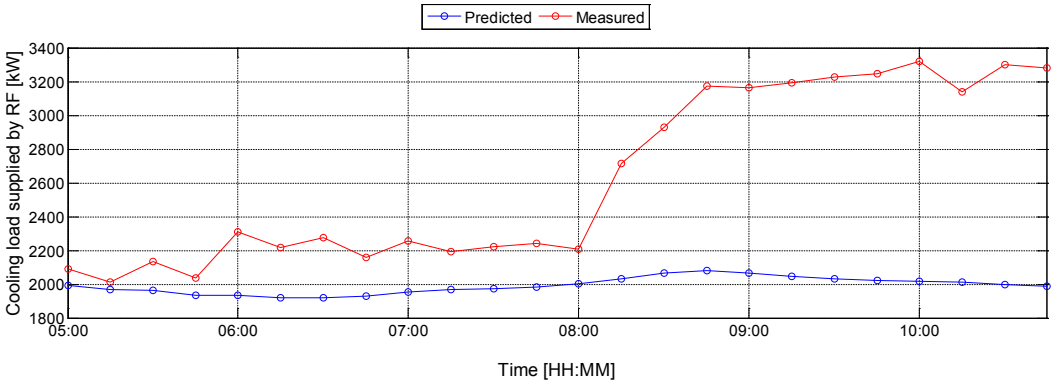

Figure I.1 Forecasted regressor over the six-hour test set on 13th of August, based on cluster #1.

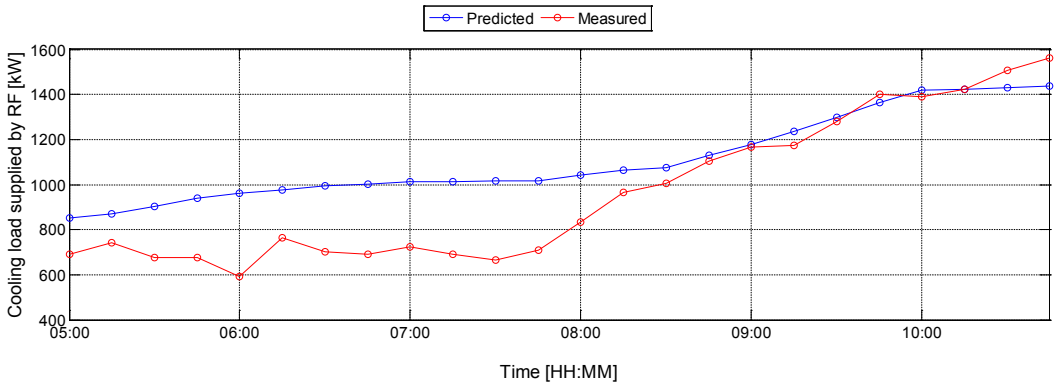

Figure I.2 Forecasted electric demand of the cooling towers over the six-hour test set on 19th of June, based on cluster #2.

Appendix J. Cascade-based forecast with the two other sequences of pre-processing steps

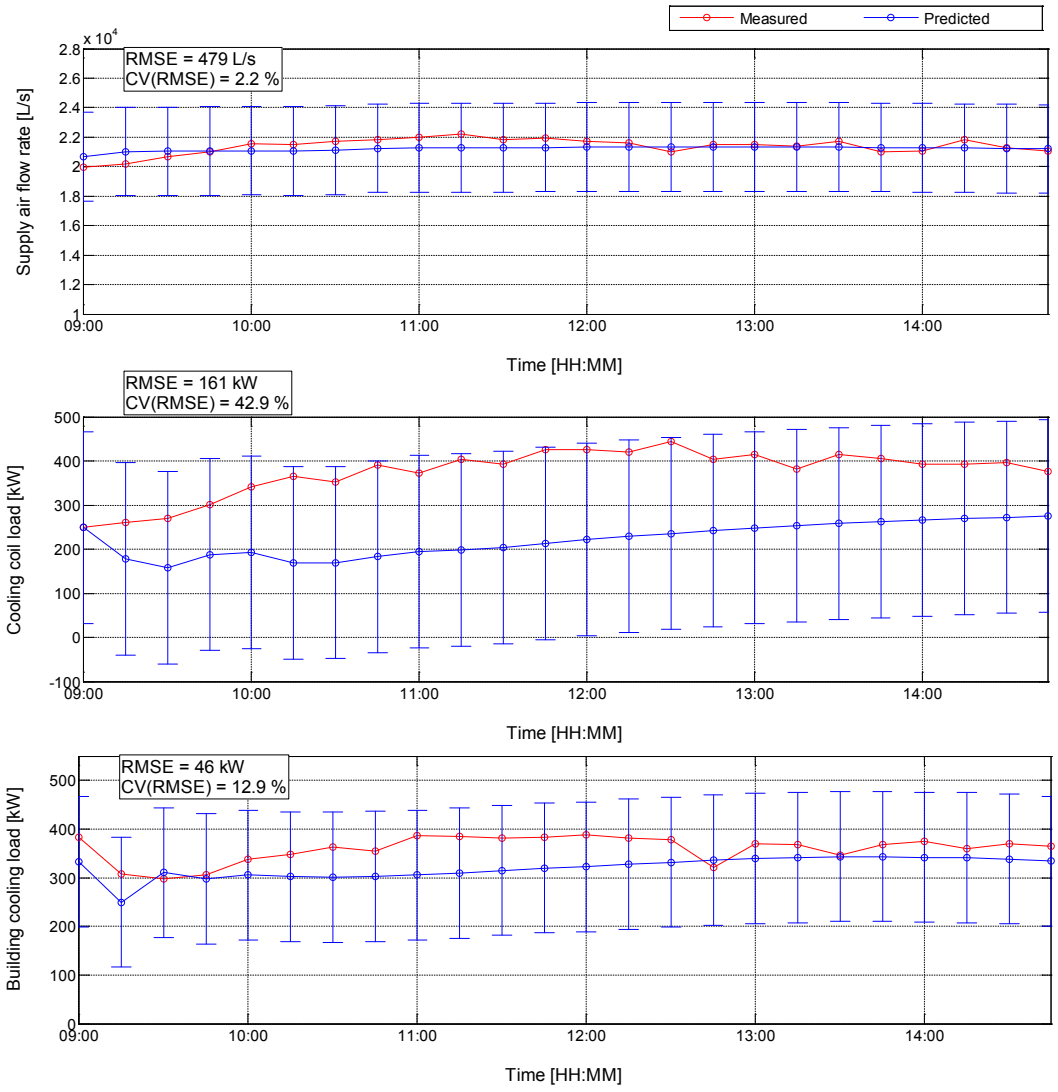

Figure J.1 Forecasting of the first three target variables on Friday, the 30th of July 2014, with sequence B.

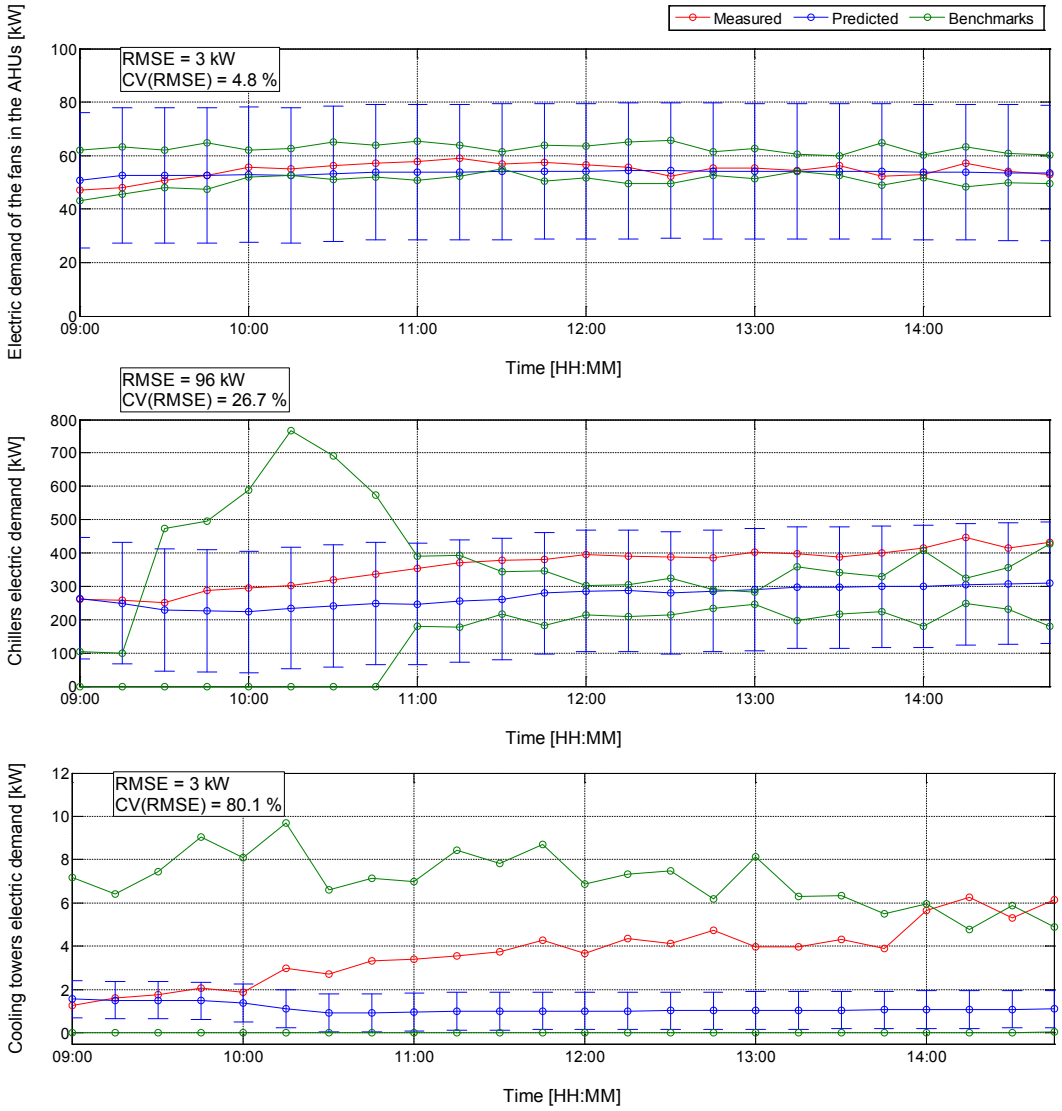

Figure J.2 Forecasting of the electric demand of secondary and primary cooling systems on Friday, the 30th of July 2014, with sequence B.

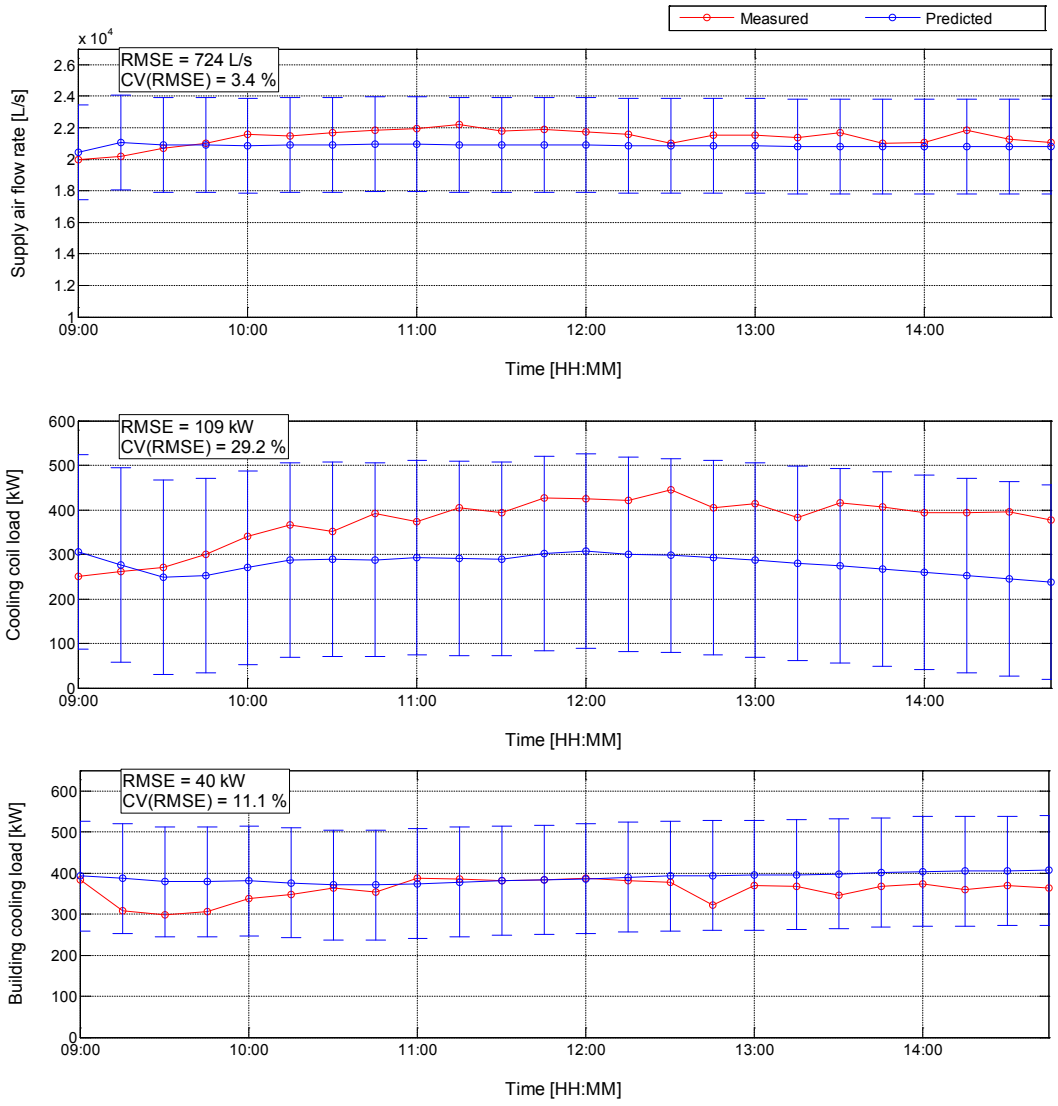

Figure J.3 Forecasting of the first three target variables on Friday, the 30th of July 2014, with sequence C.

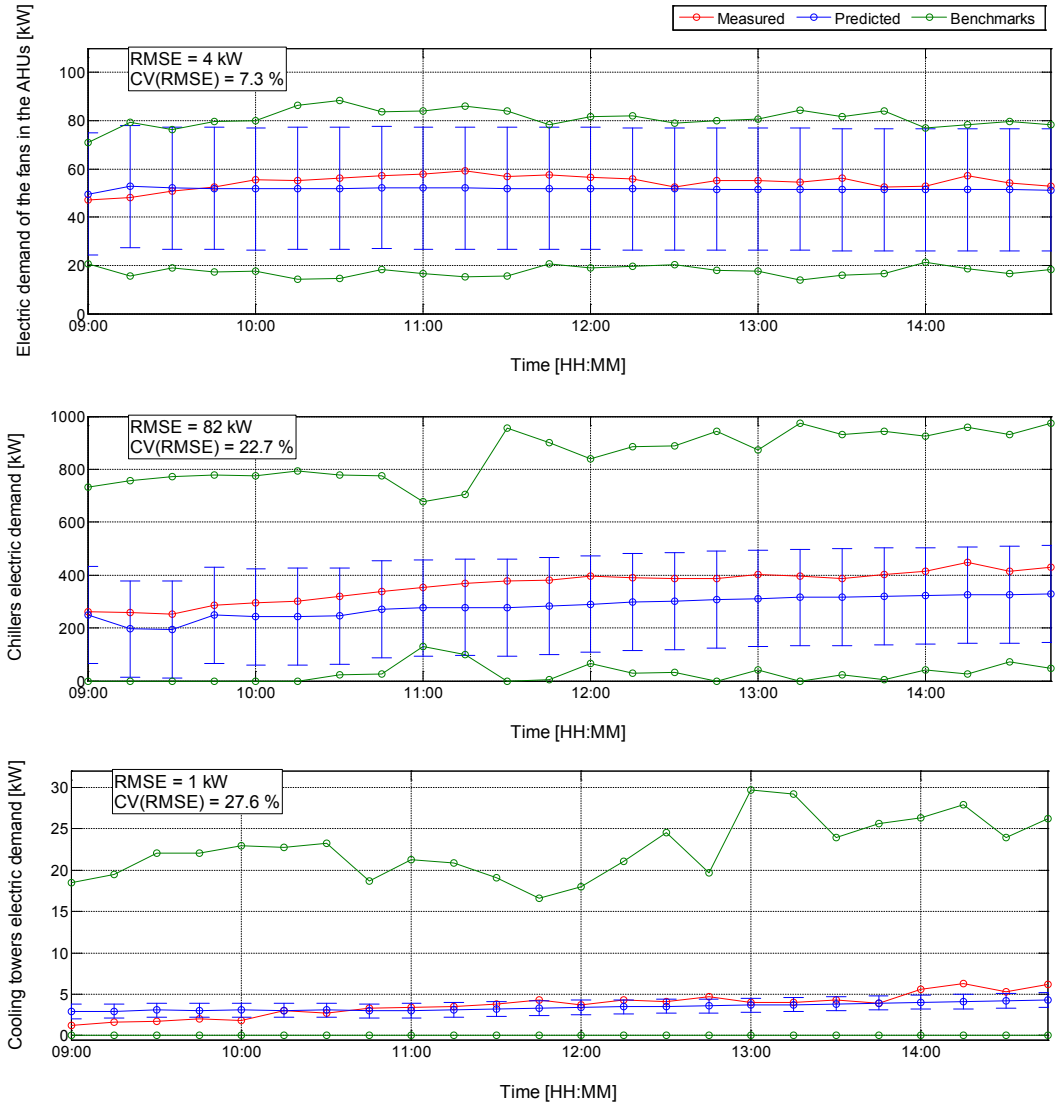

Figure J.4 Forecasting of the electric demand of secondary and primary cooling systems on Friday, the 30th of July 2014, with sequence C.