A MASTER THESIS

INVESTIGATIONS TOWARDS DEVELOPING AUTOMATED ENERGY DIAGNOSTICS AND PREDICTION MODELS, BASED ON BMS SENSOR DATA ANALYSIS.

TUSHAR AGARWAL

TU Delft Delft University of Technology

Investigations Towards Developing Automated Energy Diagnostics and Prediction Models, Based on BMS Sensor Data Analysis.

Master Thesis

by

Tushar Agarwal

In partial fulfillment of the requirements for the degree of

Master of Science

in Sustainable Energy Technologies

Specializing in The Built Environment

at the Delft University of Technology

To be defended publicly on Friday, 25th of August 2017 at 13:30pm.

Student Number	: 4505360	
Research Supervisor	: Dr. Ir. Laure Itard	
Thesis Committee	: Dr. Ir. Laure Itard	TU Delft – OTB Researcher for the Built Environment
	Prof. dr. Ir. Ad van Wijk	TU Delft – The Green Village
	Dr. Tamas Keviczky	TU Delft – Systems and Control
	Mr. Joep van der Velden	External Examiner – Kropman Installatietechniek

This is the final thesis report as of 12th August 2017

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft Delft University of Technology

Author's Note

Lights will guide you home.

Acknowledgements

My time with the research faculty of the OTB at Technical University of Delft, was one where I gained experience and knowledge as a researcher in the built environment. At foremost I would express my gratitude towards Dr. Laure Itard. She has been an utmost pillar of support and confidence for me. Always imparting her wisdom and inspiration, she has become a close figure towards the onset of my future career. My research work, and this thesis is guided by her constant motivation, the weekly discussions and editorial skills. Through the courses which I have pursued under Ma'am Itard, I have come to learn a great deal about energy optimization in buildings, for which I shall remain grateful.

I would like to thank Prof. Dr. Ad Van Wijk, for being an integral part of my thesis committee. With his experience in the field of green buildings and an important sustainable innovator, I feel honored to have my work examined by him.

I would also like to thank Dr. Tamas Keviczky, from the department of systems and control TU Delft, for reviewing my thesis and being a member of the committee.

I express my immense gratitude towards Mr. Joep van der Velden, an important member from the industrial sector working on building automations at Kropman, a Technical Installation Company. I am deeply humbled by his involvement in the thesis committee.

I would like to express my extreme gratitude towards Mr. Arie Taal, and Dr. Jan Dirk Scahgen for their continuous support from The Hague University of Applied Sciences, Delft. I thank them for their help with the data from the University and their guidance.

I would like to aknowledge the guidance and support provided by Dr, Henk J. Visscher, during my tenure at the OTB, TU Delft.

I would also thank the academic counsellor of Masters in Sustainable Energy Technology, Ms. Leonie Boortman, for her continuous support with regards to the academic as well as non-academic issues during my tenure at this master's program.

I would also like to thank all the PHd candidates from the OTB department of TU Delft, Miss Paula, Miss Faidra, Mr. Tasos and Mr. Aarash for their guidance whilst I was completing my thesis.

Finally, I would like to portray my sincere gratitude towards my parents, Mrs. Sujata Agarwal and Mr. Suresh Agarwala, and my sister Mrs. Saloni Agarwal, for their unconditional love, and support without which I would not have been able to reach this stage. They have been the reason for all that I have achieved today.

Tushar Agarwal

Delft, August 2017

Summary

An investigation into energy optimization techniques for buildings was initiated that led to the development of a Toolbox with several functions for analysis, optimization and prediction techniques for thermal energy demand of a school building. The HHS, or The Hague University of Applied Sciences in Delft, was the most sustainable building for the years 2011-2012. Naturally, the building also incorporated features and capabilities which can help an engineer to study methods of making a smart building, better.

Using sensor driven data from stored databases in the building, optimization and analysis tools have been developed for the building, at the room level. These analyses are automated into the toolbox for any given room of the building, with minor changes. The goal is to help an expert analyze the room in a quick and efficient manner.

Using the indoor/outdoor climate data, occupancy related profiles, and internal heat loads, the model can also generate predictive patterns and determine the explanatory power of each of these variables on the thermal energy demand of a room. To do this, the Toolbox is designed with two predictive modeling techniques, unique in their own ways. The first being a Multivariate Linear Regression model, that allows for estimation of thermal demean based on a linear thermal balance equation of the room. This is followed by the use of Artificial Neural Networks, to dive deep into the intricacies of the complex data of a room, especially in the case of a highly controlled indoor climate of a room. The goal here was to understand the predictive capacity of these techniques over a) real time data, and b) over the data belonging to a room and not the entire building.

Finally, looking outwards to optimizing energy demands of buildings, this Toolbox, aims at estimating quick wins that can be gathered from a smart building, to reduce energy demand further and tend the building towards nearly zero energy in the future.

Table of Contents

Table of Figure

Appendix

List of Tables

Appendix

1.Introduction

1.1 The Current Scenario.

The rapid increase in human population over the last decade, has led to some major discoveries, most of which were aimed at making human life a better, fulfilled life – however, this was done at the cost of the health of our ecosystem. There has been an exponential rise in energy productions and consumption for almost every single attribute, associated with mankind, depleting the resources of this world, plundering it to intolerable levels. Large amounts of fossils used for the production of typically, energy, has led to an increase in Greenhouse Gases (GHGs). Buildings, both residential and commercial (Offices and Universities alike) are currently one of the largest energy consuming sectors, accounting for over one-third of the total global energy consumption, and are equally responsible for heavy rates of CO_2 emissions in the world [1] .

In 2015 the Paris Climate Conference, or also called the COP 21, was the official event started from RIO in 1992, aiming at bringing together over 190 countries to achieve a legally binding and universal agreement on climate. This agreement aimed at keeping global warming below 2°C [2]. This achievement would require an estimated 77% reduction in total CO_2 emissions in the buildings sector by 2050 compared to today's level [1].

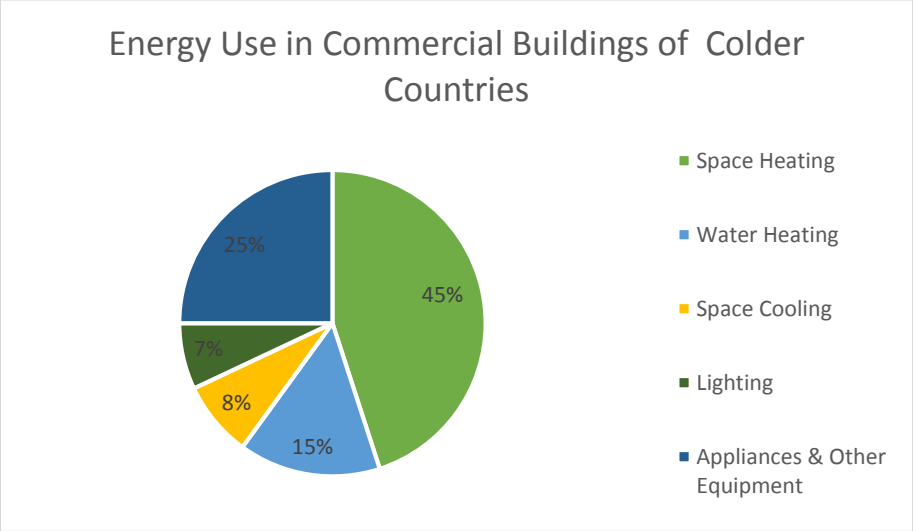


Figure 1a – Distribution of consumption levels of energy in different sections in buildings as of 2015 (IEA, 2015).

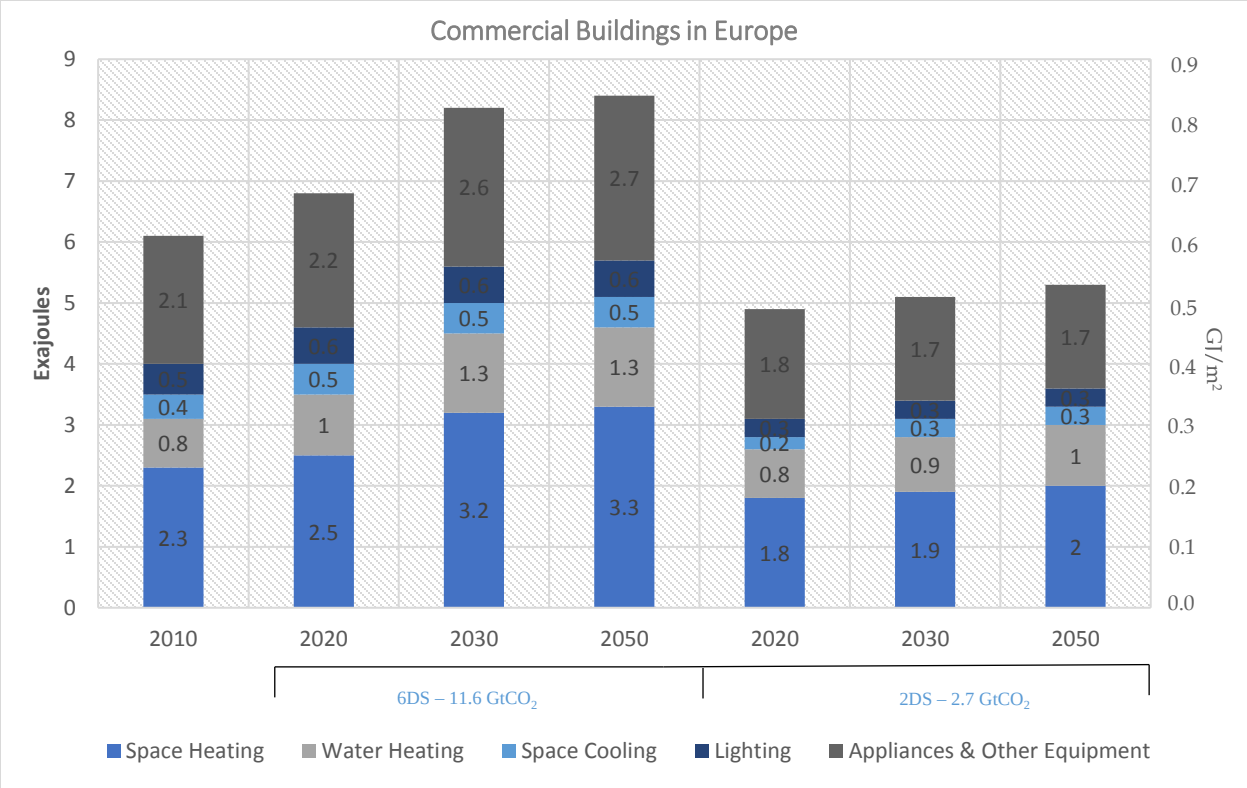


Figure 1b – A graph of the results obtained on improving different sections of energy use in buildings to reduce the overall temperature warming to 2ºC (2DS in the graph), by 2050. (IEA, 2015).

From the figure above (Figure 2), is obtained from the IEA statistics of 2015, which states that an improvement in consumption of energy related to HVAC of buildings, that is, space heating, cooling and ventilation, along with lighting and appliances itself, could be a major factor in reducing CO_2 consumption. The desired pattern of energy demand is highlighted in the trend shown above. This could also help achieve the 2DS (2 Degree Warming) margin. It is abundantly clear that despite the production of high volumes of energy from renewable sources, there needs to be a conscious decrease in the consumption through efficient methods of energy use and monitoring, especially of space and water heating and cooling. This research aims at the demand side reduction and efficient use of energy within the built environment.

1.2 Sustaining the Built Environment

Urban regions of the world are rapidly increasing in size, and structure. Whilst they are consuming a large share of the produced energy, they also provide the concentrated opportunities to save energy [3]. The urbanization over the past decade, has led to a massive rise in structures, for housing, commercial and industrial purposes. The European initiative towards energy efficiency in the built environment has been chiefly monitored and propagated by the EPBD or the European guideline energy performance of buildings[4].

European countries have always prioritized the thermal comfort and living conditions of the indoor environment. In the Netherlands, the Energy Performance Coefficient (EPC) of buildings has become a mandatory policy towards maintaining a standard for newly constructed commercial buildings. It has been sharpened over the last five years from 0.8 to 0.6 for office and other commercial buildings [5]. This has led to massive improvements in the quality of equipment for installations and thermally efficient materials for construction. Buildings with sensor driven database appear to enhance the HVAC and lighting automation, making the building sustainable and smart. However, with the implementation of such measures, the need for monitoring and analysis is vital for the continued service and development of much smarter systems. The main objective of such monitoring is to understand the true energy efficiency of smart buildings, give possibilities to improve this energy efficiency level during operation, and to estimate the overall usage and Demand Response (DR) to offset non-essential peak energy use [3].

Energy Research Centre of the Netherlands (ECN) presented a report in 2015 that strongly correlates space heating with total energy consumption of a building [6]. It shows that improvements in energy efficiency of space heating are mainly responsible for improvements in overall energy efficiency. Figure 2 shows the consumption of primary energy and specific energy for space heating, appliances and building systems (mainly ventilation, water heating and lighting) in both residential and nonresidential buildings of Netherlands [6] .The specific energy consumption for space heating is much higher for nonresidential buildings than residential.

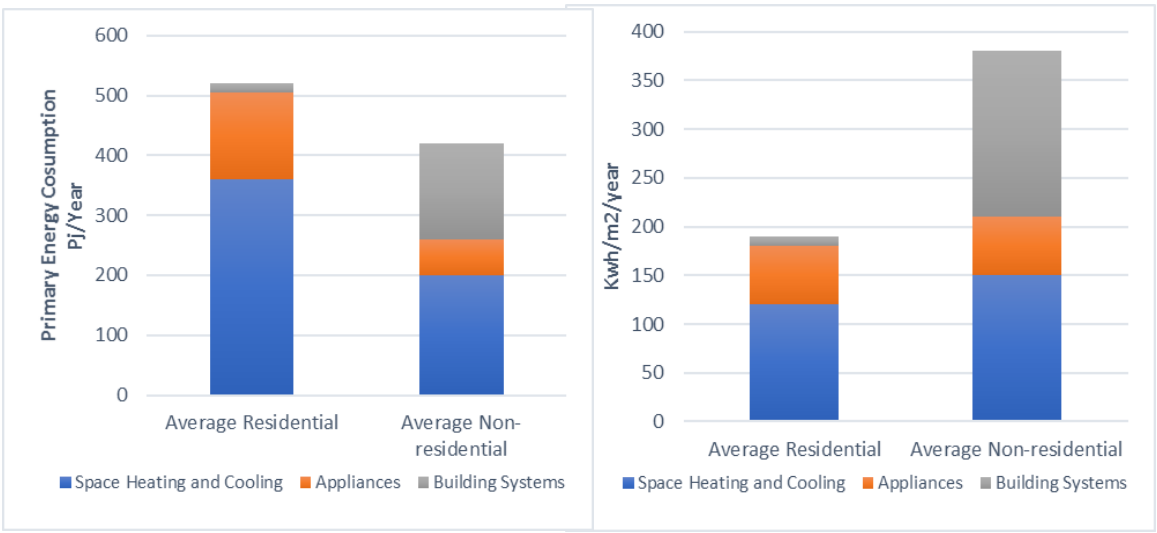


Figure 2 Consumption levels of primary energy and specific electrical energy for space heating, appliances and building systems in both residential and nonresidential buildings of Netherlands (2010) [6].

Considering the current scenario in The Netherlands, great importance must be given to improvements in demand of energy for space heating and cooling of existing buildings. This thesis deals

with office and in particular University/School buildings. Each building tends to exhibit unique energy signatures from each room, based on its specific use or function. There have been several studies at the entire building level, however, this thesis also looks into indoor climate and comfort control. This comfort control may be better studied at the room level. For example, in the case of a school building each classroom, or staff room, differs in occupancy levels, lighting levels, energy demand levels, etc., and thus have different set points for a comfortable indoor climate. Moreover, these rooms are individually powered, and monitored in today's smarter buildings via the Building Management Systems (BMS) using sensors. However, a more refined and coarse methodology is needed to understand the actual patterns exhibited by each room, so as to find ways of reducing energy demand especially during peak hours. This can be achieved through understanding the demand and supply of every room in terms of energy, especially heating and cooling energies. Working at the room level calls for a atomized tool, so as to analyze multiple rooms of an entire building with swiftness and accuracy.

What follows is an overview of how the current scenario of building-side energy demand stemmed the motivation for this research. This is followed by a brief about the Installaties2020 project, under which part of this research was conducted, along with the aim and objectives that have been dealt with this research.

1.3 Background & Motivation

1.3.1 The Installaties2020 Project

The motivation for this research results from the need for automated tools for determining energy reduction potentials in the built environment, whilst maintaining a high order of indoor climate comfort. Trying to find the balance between high comfort and lowered energy consumption is a complex task. HVAC lighting and sensor installations, need to be highly controlled and fine-tuned to the building characteristics, in order to be efficient. This process is difficult, time consuming, as it is never the same for any two given commercial buildings, let alone two separate rooms. Several initiatives have been brought into light over the past decade – projects from the EU governments, research institutes, etc. A group of corporate companies in the Netherlands, along with certain research institutes, began a project titled the Installaties2020 [1].

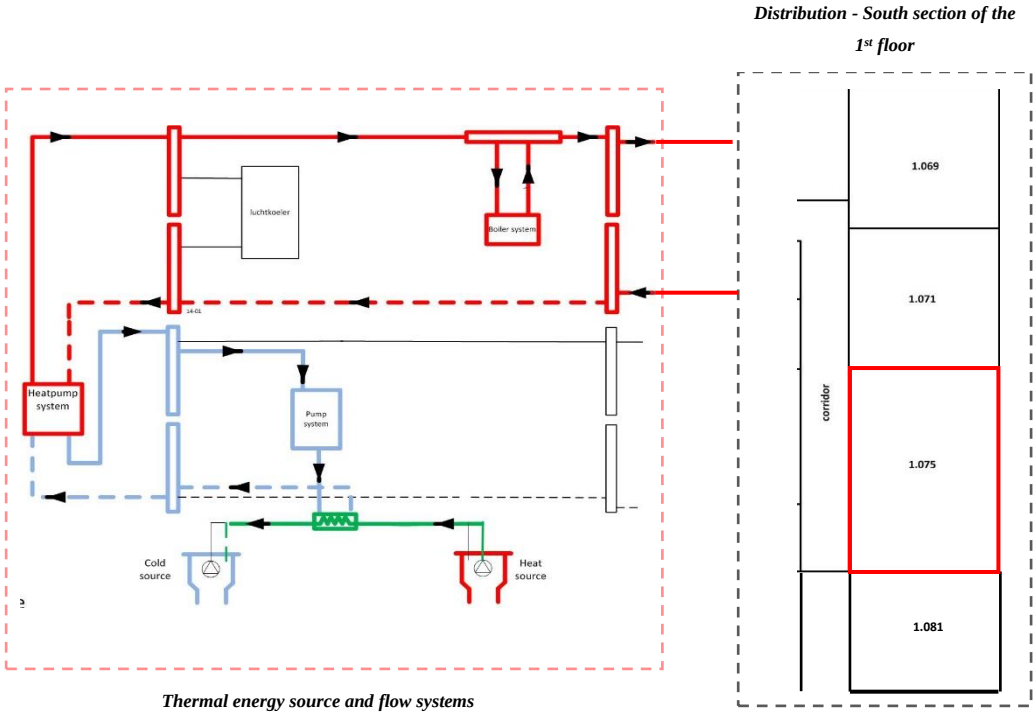
The purpose of this project is to develop diagnostic models and find optimal functioning and controls of building services to achieve targeted goals of energy and CO_2 reductions as well as better indoor comfort level. A part of this project was based out of a building of the HHS, or The Hague University of Applied Sciences, in Delft, a city in south Netherlands.

[1] Refer to the website of "www.installaties2020.weebly.com", for further information.

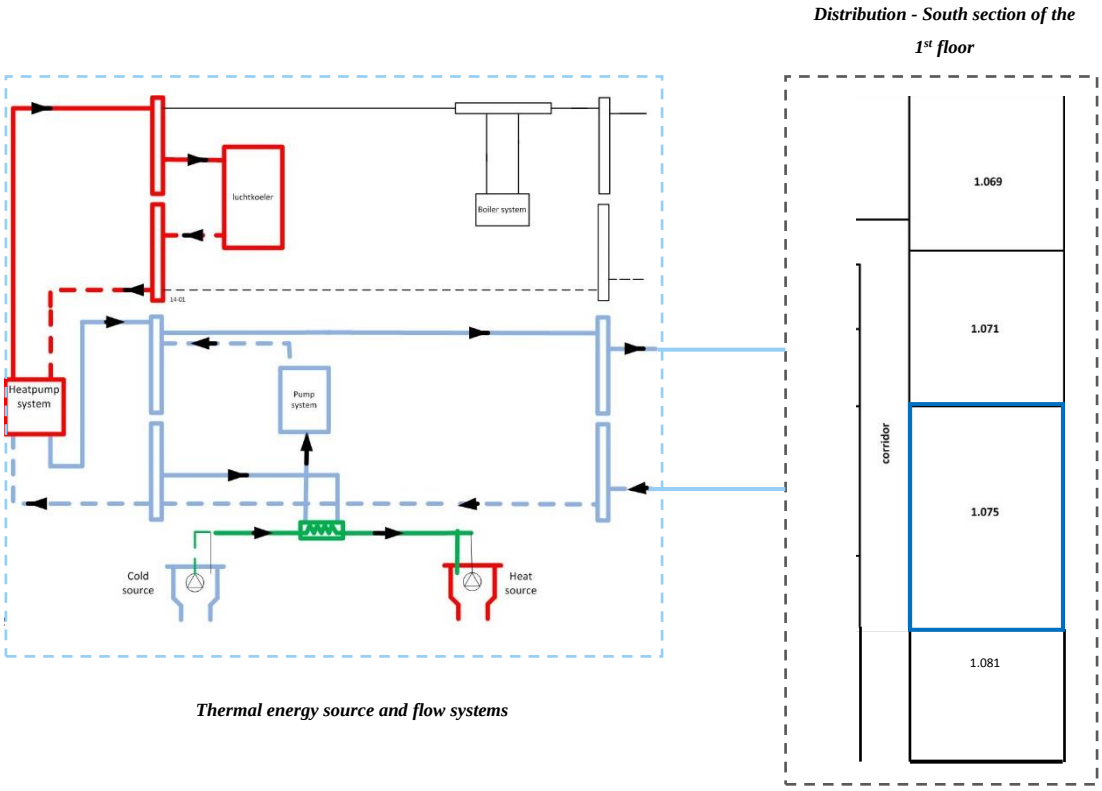
1.3.2 The HHS – Building in Delft

The case-study building is a University building located in Delft. The proposed research framework is built around the data and information gathered from The Hague University of Applied Sciences, as this building is seen as being quite representative of the way future building with controlled HVAC equipment will be designed – highly energy efficiency complex installation with a sensor rich environment. The building is fairly new - it was built in the year 2009, and has been one of the most sustainable buildings of Netherlands (2010-2011). The building has been used for research due to its complex HVAC systems, including an Aquifer Thermal Energy Storage (underground heat and cold storage ATES), and sensor rich rooms with advanced control and data monitoring and storage capability. The rooms of the HHS are heated via a floor heating system and by ceiling panels in which water is circulated. The building consists of a majority of classrooms and a few offices for lecturers.

In terms of sensors, the building has a central Building Management System (BMS) which relays information from sensors placed in each room. These are the PIR (presence) sensors, CO_2, temperature (both air and wall surface) humidity, ventilation, electrical plug and lighting sensors. Alongside these, there are sensors for obtaining valve position data, which gives a good estimate if the floor heating or cooling circuits in the rooms are open or closed (see the appendix A.1.1). It is important to note that the data of this building has been stored in a database and is managed by a software named Octalix and Priva which also helps in monitoring and analyzing data.



Thermal energy source and flow systems



Figures 3 a and b - floor plan of the ATES and the Heat pump system, along with the supply of thermal energy to the rooms facing the south side of the building on the 1st floor – a shows the heating demand whereas b shows the cooling demand and the corresponding flows of water.

The floor heating and cooling system at the HHS is provided using thin water pipes divulging out of a main pipeline which supplies warm or cold water based on the demand of a given section of the building. The ATES system is used to supply most of the heating and cooling demands of the building. However, in the case of high heating demands, extra electrical heat pump is available. An important aspect of this system is that there must be a balance in the thermal energy stored in the ATES through the year to allow for smooth functioning of the ATES system. A detail on the functioning of the floor heating system is provided in the Appendix A1.1.

Certain images of the room 1075, and the placement of sensors have been placed in appendix A1.2.

1.3.3 Energy Improvements in Smart Buildings

Smart buildings with a highly optimized Building Energy Management Systems (BEMSs) are developing all across Netherlands, and other parts of the world. BEMS are integrated with the active systems of the building such as the HVAC, lighting, and operational times [7]. A majority of these systems are based upon highly advanced computational and information technology.

With the availability of intrinsic and widespread data, from sensor rich environments within the buildings, research into the possibilities of strengthening BMESs and developing more automated tools for a user can become the first steps towards developing true smart buildings. Analysis and deep learning can help develop techniques for increasing the robustness and reliability of such system. By doing so, more paths of energy reduction methods can be drawn up within existing, smart buildings. The idea behind this research is to investigate the means by which automated tools could help understand the functioning of a building in coherence with its BMES and find points where improvements can be made.

1.3.4 Thesis Layout

What follows after this introduction is the research outline and objectives of this study. Chapter 3 describes the literature and the state-of-the-art of existing methods and techniques, related to this research. A literature study on the most necessary topics is undertaken, both theoretical and practical in relation to the built environment and energy simulations of buildings. This chapter helps the reader to understand the knowledge gap present in lieu with the objectives of this thesis.
Based on the findings of chapter 3, the most apt methodologies needed to develop this research are described in chapter 4. This chapter explains in brief the types of models which shall be developed, in order to find answers related to the research questions.

Chapter 5 describes the data being used in this research, and the processes involved in order to mine and organize this data. This is important since the project is based on real-time data recorded over an entire year at the HHS rooms.

Following this are the ‘step-toolboxes’ developed for the automated Toolbox of this thesis spread across individual chapters. Chapter 6 regarding the graphical analysis step-toolbox, chapter 7 based on the correlations and its importance in building energy estimations, chapter 8 on the Multivariate Linear Regression and Prediction models, and chapter 9 on the Artificial Neural Networks models.

Finally, chapter 10 concludes this research and its findings, by answering certain important research questions based on the findings of the previous chapters. The recommendations arising from this research are placed in chapter 11.

2. Research Outline

2.1 Problem Definition

The use of smart systems such as those incorporated via the Building Energy Management Systems (BEMS) in buildings alongside complex HVAC installations demands for more automation and increase in accuracy of demand and supply of energy predictions. As mentioned above, the HHS building is a sensor rich building. The control responses of such buildings are fine tuned to the sensors placed at the room level. However, even today there are several causes of concern regarding the complete automation of building operations due to the complexity involved in such systems. Some of these have been labelled below:

1. *Challenging and Highly complex data*

   The data obtained for this research and other buildings included, are highly complex to organize, monitor and analyze in relation to the complex HVAC and lighting system in place, especially when it comes to high frequency data (hours and minutes).

2. *Errors and inaccurate data*

   The reliability on the data obtained from BEMS systems is a cause for concern. According to Arie Taal [8], assessing the reliability of data regarding energy systems is quite challenging. There needs to be a simple and fast method to monitor and assess the data, and make sure it is comparable to the data from other sources either through normalization, etc.

3. *Need for understanding influential parameters for different rooms*

   The complexity of models being used today lack simple information such as the effects of parameters on the energy demand.

4. *Disparity amongst the energy demand prediction and actual values.*

   Even with the increase in control methodologies, there seems to be differences in the predicted vs. actual energy consumption.

5. *Present of faults in the system*

   It is often understood that there needs to be automated monitoring systems for better understanding of errors and detection of faults within sensors and HVAC systems [8].

6. *Need for continuous commissioning*

   In the field of BMES the need for continuous optimization of the control systems is an important task to keep improvising on the performance of the building. Therefore, continuous commissioning and repairing of degraded HVAC systems and sensors is needed.

7. *Expert Analysts Needed*

   In order to program, and operate a BEMS automated system, for energy monitoring there needs to be expertise involved. This model aims towards introducing a more user-friendly tool.

2.2 Research Objectives

This research aims at running parallel to the visions and goals of the Installaties2020 project and develop fast automated and statistical data-driven models for individual room levels. This research investigates various possibilities of using the BMS data for energy performance optimizations and deals with the following.

a) Develop a model inclusive of all steps, from identification and cleaning of data from a sensor rich environment, up to the detailed analysis and prediction control of thermal energy in a room.
b) Establishing a generic automated model to provide detailed graphics with regression analyses of the functioning of the room and the building.
c) Providing a methodology involving the use of correlation coefficients to understand the most probable and effective parameters influencing the heating and cooling demands at room level. This helps determine the most prominent losses and gains of thermal energy in classrooms.
d) Utilization of the established parameters to perform multivariate regression analyses. These could be used in turn to train models for prediction of thermal energy demand.
e) Utilization of ANN techniques to improvise on the prediction demand of energy.

3. LITERATURE

This chapter aims at giving a directive towards the design and mathematical models which have been studied and implemented in this research so as to satisfy the explained research objectives.

Firstly, a brief introduction into the principles of heat transfer is presented. This is important as the major losses and gains of thermal energy in a room is dependent on several factors. A basic understanding of the mechanisms of heat transfer and the direction of flow of thermal energy is thus explained, along with the implications on its overall balance at the building and/or room level.

Secondly, upon understanding the fundamentals of heat transfer in buildings, and the factors through which the balance of thermal energy of a room is disturbed, a literature study is described relating to the methodology of correlation coefficients developed during this thesis. This is done in order to understand the degree of effect each variable has on a room's thermal demand. It gives an immediate insight of the physical and operational characteristic of the room and the systems in play to maintain thermal comfort within the room.

Lastly, the state-of-the-art of different prediction models used for thermal energy data, are studied, to help the reader gain a brief insight into the comparative analysis of the complexity and accuracy of each model with regards to room-level thermal energy prediction.

3.1 Heat Transfer in Buildings.

Heat transfer within buildings is determined by the interaction of heat flows through the three main modes of conduction, convection and radiation, between a building and its surroundings. This section will explain the basics of heat transfer within buildings and the relevant mathematical formulae which define the phenomena numerically.

Figure 4 shown below is a diagrammatic representation of the basic heat transport components in a room with floor heating/cooling water controlled system. The difference in temperatures between the indoor and the outdoor air temperatures is the main driving force behind heat transfer in buildings. with four different temperature points; the ambient temperature (T_e), the indoor air temperature (T_a), the wall surface temperature (T_w) and the floor surface temperature (T_f). The floor is heated or cooled based on the set point temperature of the indoor air (T_a). The indoor temperature balance of a room is constantly affected by several parameters such as the side room temperatures (T_{a1} and T_{a2}), causing conduction of heat through walls. Occupancy, lighting and appliances lead to increased internal heat gains $Q_{internal}$. The sources of heat

energy supplied to the room are Q_{solar} (solar radiation through either 1. walls or 2. windows) Q_{floor} (radiation from the floor heating system). Q_{floor} (can also be negative indicating cooling mode), $Q_{ventilation}$ is the mechanical heat or cooling supplied via air through the HVAC systems and $Q_{internal}$, the heat produced by lighting, appliances and people.For detailed explanation see section 8.1.

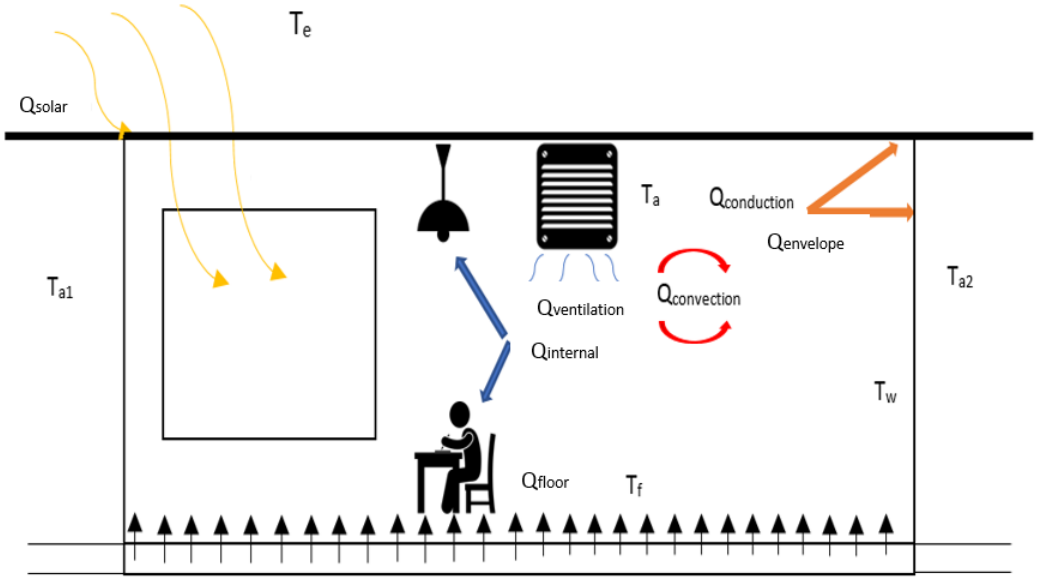


Figure 4 Description of the distinct modes of heat transfer in a room with a floor heating/cooling system.

The distinct variables have a direct or indirect effect on the total energy demand of each room to maintain a constant indoor temperature. In order to carry forward with a predictive model for thermal energy demand, it is important to recognize the magnitude of effect of these variables on the thermal energy demand of a room[9]. Not just for predictive control, but such studies also help in understanding the characteristics of the room and the systems installed.

3.2 Correlation Coefficient

In this research, there are several parameters which have been recorded and used as input data towards understanding thermal energy demands at the room level. Each parameter is,

1. *affected by another parameter individually and*
2. *together in correlation they affect the thermal energy demand*

Some important mathematical terms used in order to explain Correlation are defined as follows;

1. ***Covariance*** – is the descriptive measure of a linear relationship between two variables (Eq. 3.1). It does not provide the strength of this relationship, but the direction. A positive direction means directly proportional, and a negative direction equals inversely proportional variables.

   For two variables x and y and N samples in total, the population covariance is given by,

$$\sigma_{xy} = \frac{\Sigma^{i}(x_i - \overline{X_N})(y_i - \overline{Y_N})}{N} \qquad \text{Eq-3.1}$$

   Where $\overline{X_N}$ and $\overline{Y_N}$ are the population means of the data x and y.

   The population covariance is used since covariance is being performed on the entire data set and not a sample size alone [10] [11].

2. ***Simple Correlation*** - Since covariance best answers the direction of a linear relationship, correlation values estimate the direction and strength (Eq.3.2). Whilst covariance has no upper or lower boundary, correlations (r) are scaled from -1 to +1, that is, these values are standardized.

$$r_{xy} = \frac{cov(x, y)}{s_x s_y} \qquad \text{Eq-3.2}$$

   Where, s_x and s_y are the standard deviations of the data x and y [10].

   However, with more than two variables, the simple correlation is replaced by multiple correlations explained ahead.

3.2.1 Multiple Correlation - An overview

As pointed out previously, before considering modelling techniques for predictive control of thermal energy, there is a necessity to understand the degree to which factors/disturbances affect this thermal demand. The indoor temperature varies with the energy flows interacting with the building, e.g. heat gain from solar radiation, occupants, heating system, along with the thermal properties of the building envelope[12]. This section will give a descriptive understanding of multiple correlation coefficient between diverse factors, based on literature.

Researchers have been developing models based on factor analysis methods to understand the effects of parameters on energy consumption. Li Yuana et. al.[13], utilized Pearson's correlation in order to estimate the correlation between influencing factors effecting energy demand and the total energy consumption. This was performed for large public buildings and involved the estimation of electrical energy. Alongside this, an *F-Statistics* testing was also performed for determining the significance of the correlation coefficient.

The p-value is the measure of the significance of the null hypothesis [14] [15]. The significant level generally chosen for the null hypothesis is 5%, or (p-value < 0.05). This means that there is a confidence level of 95% that the coefficient is a correct value for the correlation between the dependent and the independent variable. See appendix A.2 for more details regarding statistical significance.

Other researchers such as D. Bing et. al. [16], studied the effects of outdoor climate on the energy consumption of commercial buildings by estimating the goodness of fit. This was done by using the Pearson's correlation coefficient as described above. The outcome of this research was that the outdoor variables, typically the dry bulb temperature, was heavily correlated to the energy demand for the building. Similar research by Xiaoquing Wei et.al.[17] was performed on occupancy and energy use of a building, showing heavy correlations between the two. Detailed information regarding Multiple correlations has been jot down in the appendix section A2, and chapter 7.

This research deals with combining the correlation effects of outdoor parameters, building characteristic HVAC parameters, and occupancy related parameters to the thermal energy demand. There is not much work done on the degree of influence these parameters have on heating demand of floor heating systems, at room levels [18].

3.2.2 Partial Correlation Coefficients Matrix for Building Energy Models

The underlying principle of Partial Correlation Coefficients Matrix (PCCM) method is to be able to identify the direct relationship between variables, by eliminating the causal effect from indirect pathways [19]. This helps to eliminate any interdependency that may exist between two or more multiple variables, effecting the overall correlation. For example, the outdoor variable such as solar radiation may have a certain effect on the outdoor temperature. Jointly these two have an indirect effect and a direct effect on the building energy demand. The goal of PCCM is to capture only the direct effects, and ignore the indirect correlations which may exist. Thus, partial correlations help in establishing a degree of sensitivity of a independent variable towards the dependent variable [19].

The use of partial correlations has been missing from the literature pointed out above [12 13 14]. This research shall thus adopt Partial correlations as a secondary correlation methodology to find the individual effects of parameters on the net thermal demand of a classroom, and make a comparative study of the use of Multiple vs Partial correlations.

3.3 Predictive Control Modelling - An Overview

3.3.1 Model-based approach

Model based calculations help in answering several questions about the building, in this case, a given room's actual performance and efficiency. Energy models for buildings started with analytical models in the 1940s by Bruckmayer[18] which dealt with conduction through a single element of the building. Several researchers have used a model based approach for developing predictive models for building related energy demand; refer to [9 15 18 20 21]. According to Giorgio Mustafaraj et. al. [22], the estimation of energy can

be tackled using two different classes of modelling techniques. Physical or white box modelling, and the black box modelling. Based on the level of available information and prediction accuracy the researcher can choose his/her prediction model [21]. In between, grey box models are found, which are a combination of physical and data-driven models. Almost all model techniques are categorized into these three main types.

- ***White Box Models*** (also called law-driven models): are the transparent models, based on mathematical equations along with a high dependency on building's characteristic data [21]. Several models which already exist include those based on the ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) standards, such as Energy+, TRNSYS, ECOTECT, DOE-2, etc. which are efficient detailed energy simulation techniques. Their disadvantage is the difficulties in calibration procedure, and certainly time consuming, or each building to be modelled due to extensive user based inputs [12] [23].
- ***Black Box Models*** – are best suited for detailed energy simulations of a specific building due to their convenience and quick modelling. These do not require any theoretical knowledge of the building, but simply rely on statistical methods of relating input to output parameters. Typical modelling techniques involved are Artificial Neural Networks (ANN), Genetic Algorithms Multiple Linear Regression [12] [18] [15]. These networks do not need physical knowledge about a building or room. This is advantageous in the case of existing buildings, with smart meters, and sensor rich environments. Analysis by using such models and can give a lot of information on load and demand of energy, but no data regarding the building itself is needed [12].
- ***Grey Box Models*** – are a combination of both white and black box models. These models use a mix of building parametrical data and statistical methodologies. Basic examples of this models are the RC network models [15] [21] [23], that can be used in order to recover building characteristic data using inverse modelling techniques as performed by Parab [23] (section 3.3.2.2 ahead).

Furthermore, building energy models can be classified into two categories, static or dynamic [23]. A steady state model is used for long intervals (weekly, monthly, yearly) measurement data and cannot be thus used for short term transient states in indoor temperature or building properties. A dynamic model on the other hand, is able to capture transient measurements and are mostly used on hourly, or sub hourly levels. Dynamic state models are best fit for solving mathematical and statistical models.

What follows is a descriptive comparison of the models used by several authors till date, to perform prediction and training of datasets. This would allow the reader to understand the basic necessities that must be taken into consideration to choose a suitable model for prediction.

3.3.2 A comparative study of Predictive Models in Use

The categorization of models in white, grey and black box models do not all incorporate features of predictive modeling. In view of the research objective, the literature survey shall focus on comparing the models used by several authors till date and thereby choosing an appropriate building energy predictive model.

3.3.2.1 White-box Models

White box models are state-of-the-art detailed physics based models. These are effective in designing whole-buildings and are highly dependent on the input data of the building parameters and energy systems. Thus, the accuracy of such a system depends very much on the knowledge and prowess possessed by the user. In existing buildings, it becomes even more difficult to arrange for information about the physical building structure, its RC values and this leads to decrease in accuracy of the building simulations. Although one may understand that the results from such a tool are quite good, and to the point, it comes at a great cost of time consumption as it increases with increase in the complexity of the designed model.

3.3.2.2 Grey-box Models

For smart buildings, *Model Predictive Control (MPC)* has gained a lot of attention, and is a method that can be used to enhance the functioning of Building Management Systems [24]. Jan Sirosky [20] in his research explain that MPC is a set of control strategies used in order to minimize the objective function – the energy use in buildings. This method incorporates the use of inputs and operating conditions tuned towards the overall reduction in objective function [20] [24].

Inverse modelling is another concept, which is used to calibrate simplified models. Here, the model parameters can be determined by matching the output of the model as close as possible to measurement data [12] [23].These simplified models are less complex, with a lower number of variables to optimize. This reduces the computation time considerably, however need a higher number of initial parameter estimations. For example, the inverse modelling technique could be used to determine building parameters such as RC values, ventilation rate, etc.

V. Parab in her research thesis [23], describes a mathematical model using lumped capacitance models, to retrieve building characteristic informations (overall RC values) by studying the inputs, (outdoor environmental conditions, indoor temperature and presence estimates), and the output (energy demands) using inverse modelling based on the Maximum Likelihood Method. This was however, mainly pertaining to neither office/school buildings, nor room level data, but the entire house. A main disadvantage of this method however is that it lacks a clear relation to the physical parameters (the overall RC value includes RC values of the walls, floor, roof and specific heat loses due to the ventilation flow rates), thus leading to

a difficulty in identifying how correct the determined values are [21]. The work of Parab referred to other works, some of which are cited here [25] [26] [27].

Rowan de Nijs [28] in his research utilized the inverse modelling technique coupled with RC networks as well, to find the optimum resolution of the RC network needed for thermal energy demand forecasting. His conclusions dealt with an extensive RC model, including 9 parameters with 4 capacitances and 4 resistances in the network and 1 variables to describe the window area. This grey-box model although built on office buildings with floor heating systems, (close to this research case building at hand) it was not built on sensor driven data, but on an emulator, wherein the data was created using a white box model, for which a lot of physical information about the building itself. The work of Rowan referred to other works, some of which are cited here [18] [25] [29] .

3.3.2.3 Black-box models

Several researchers have also developed Black box predictive models. *Statistical methods*, using linear and nonlinear methods have been adopted to find the effect of several parameters on building energy, thus being able to predict the building energy demand. Kristopher et al [9]. Lopez [15], and many others (See references 8-14) use *Linear Regression* (LR) and *Multivariate Linear Regression* (MLR) models to estimate the thermal energy demand. These models are dependent on data related to outdoor factors (weather related), building parameters (indoor air and surface temperatures and ventilation rates) and consist of one independent variable. Lopez made use of a function, namely 'stepwise fit', on MATLAB 2017 [30] which is a MLR model. The models work on a simple linear formula wherein coefficients for each parameter are estimated. The functional form of this approach is represented as follows:

$$Q_h = constant + \sum_{i=1}^{n} C_i \,.\, X_i \qquad \text{(Eq. 3.3)}$$

Wherein, Q_h is the hourly thermal energy demand in Watt-hour (Wh), C_i is the coefficient estimated for the i^{th} parameter X_i. It should be noted here that the model developed by Lopez belongs more to the category of grey-box models than to the category of black-box models, as the Xi parameters were chosen in accordance to a simple model describing the thermal balance of a one-zone building.

The accuracy of these models is measured by *'goodness of fit'* or R^2 value, Mean Squared Error (MSE), which shows how closely data could be trained, thus giving an estimation of the goodness of predicted values too. More explanation on the goodness of fit can be found in chapter 8 section 3 (8.3).

Mustafaraj et al. [22] in his research pointed out several *advantages* of MLR.

- *These models are relatively simple with a lower number of model parameters.*

- *They have a set of simple equations and the results have a higher physical meaning and are easier to deal with than results obtained by other machine learning techniques (see below)*
- *The model can incorporate real time data with ease.*

One disadvantage of this methodology is that the effective non-linear parameters are not accounted for. This leads to a reduction in the goodness of fit on utilizing "real-time" data [18] [31].

Researchers have shown that the use of non-linear MPC models can enhance the functioning of HVAC systems by anywhere around 7% and reduce energy and cost consumptions by half [14] [19]. Some of these models are listed below;

1. *Artificial Neural Networks* – Seginer et al., in their research showed that these are non-linear models incorporating forward or backward learning algorithms with highly non-linear solutions [14] [29]. They are simple to use and take less time for calculations. ANN-MPC have been developed for school/University, Office, Airports and other commercial buildings including residential buildings[32]. However, a room level analysis of Machine learning has been missing, which is one of the main objective of this research. Alongside this, neural networks can also be adaptive and self-learning [22]. ANN is a data-driven modelling technique and serves as a huge potential in case of unknown building parameters. (see references 12,17-18 & 23-24). However, their efficiencies are not the best of all non-linear black box models [28].
2. *Artificial Neuro-Fuzzy Inference System (ANFIS)* – these models are highly complex and make the use of both fuzzy logic and ANN networks [28].
3. *Genetic Algorithms* – These models can be used conjunctively with ANN or with ANFIS and have proven to be of higher efficiency that a standalone ANN network. However, they demand for a more complicated set of networks, making it difficult to automate the models, a major objective of this research [28] [30]. For more details regarding these models see references 28 and 30.

According to Trcka and Hensen, each method has a certain modeling complexity after which the predictive uncertainty will start to increase, leading to a reduction in model accuracy. "There is no sense in going beyond this complexity, as the overall error in the model uncertainty will not decrease" [27].

The choice of using a black, grey or white box model stems from the availability of certain major attributes namely use, difficulty level, training data requirement, calculations time period and accuracy [23] [21]. A combined summary was established based on the summarization by Zinghwei [19] and Parab [21]. Parab summarizes the ASHRAE handbook of 2013, and these help in analyzing the most important models useful for this research.

Table 1 Comparative benchmark by Zinghwei and V.Parab (Adopted from ASHRAE handbook 2013) with a few commonly known modelling techniques for building energy systems [21] [23].

Method	Use[2]	Time Period[3]	Difficulty Level	Calibration effort	Calculation Time	Accuracy
Simple Linear Regression	ES	S,H,D	Simple	Low	Very Fast	Low
Multiple Linear Regression	D,ES,C	S,H,D	Simple	Low	Fast	Medium
BIN method (ASHRAE)	ES	H	Moderate	Low	Fast	Medium
Bayesian Belief Network	D,ES,C	S,H,,D	Moderate	Medium	Medium	High
RC Thermal Network	D,ES,C	S,H	Moderate	Medium	Fast	High
ARMA Models	D,ES,C	S,H,,D	Moderate	Medium	Fast	High
Artificial Neural Networks	D,ES,C	S,H,,D	Complex	Medium	Fast	High
Generic Algorithms	D,ES,C	H,D	Very complex	High	Slow	High
Detailed Energy Simulation	D,ES,DE	S,H	Very Complex	High	Slow	Medium
Computer Simulations	D,ES,DE	H	Very Complex	High	Very Slow	Medium

The decision to utilize Artificial neural networks and MLR becomes increasingly clear by looking at table 1 above. The Use column is checked for modelling methods that align with the research objectives. There are four main uses which are derived out of energy models namely, diagnostics (D), energy saving calculations (ES), design (DE) and control (C). As stated the main objective is to develop thermal energy prediction models for energy saving measures, which classify as, energy saving calculations (ES) and Control (C). We also wish to understand the modes through which the thermal energy balance is disturbed which falls under Diagnostics (D). Thus, we can shortlist the above table based on Use factor of a method, with lower complexity and higher accuracy. Table 2 below describes the chosen approaches.

[2] Uses include diagnostics (D), energy saving calculations (ES), design (DE) and control (C)
[3] Times scales shown are hourly (H), daily (D) and sub-hourly (S)

Table 2 Arrangement of shortlisted data-driven models which can better suit the objectives of this research.

Method	Use	Time Period	Difficulty Level	Calibration effort	Calculation Time	Accuracy
Multiple Linear Regression	D,ES,C	S,H,D	Simple	Low	Fast	Medium
Bayesian Belief Network	D,ES,C	S,H,D	Moderate	Medium	Medium	High
RC Thermal Network	D,ES,C	S,H	Moderate	Medium	Fast	High
ARMA Models	D,ES,C	S,H,D	Moderate	Medium	Fast	High
Artificial Neural Networks	D,ES,C	S,H,D	Moderate	Medium	Fast	High

It should be noted now, that these tables are adapted by authors from different sources and has been cumulatively arranged under this research. Thus, although this may have certain discrepancies, it has been chosen as a basis to form first hand assumptions for selecting the modelling procedure. Upon analyzing the time period and accuracy of the methods above, it was concluded that the model should have high accuracy, and bare calculations which involve hourly, sub hourly, and day-wise timescales of data. Because of a high speed of performance of calculations, The Artificial Neural Networks and Multiple Linear Regression models were chosen to be studied further. This would help in establishing a comparative analysis between Machine learning and Linear Equations, for room level thermal energy prediction of floor heating systems.

The following chapter shall introduce the major research questions supported by the research objectives of this thesis. These research questions aim at filling the knowledge gap found within literature above. Following this, the chapter will describe a method adopted in this research so as to perform the most appropriate models for carrying forward the objectives of this research.

4. Research Methodology - A Novel Combination

4.1 The Research Questions

Following from the research objectives stated in chapter 2, this research investigates various possibilities of using BMS data for energy prediction optimization. The following main research question is formulated -

Can an automated generic toolbox be setup for real-time data analysis of office and school rooms, to predict their thermal energy demand with a small number of input variables, high accuracy and a physical meaning in order to optimize energy performances?

This research question is structured based on the main objective of this research.
Objective a) Develop a model inclusive of all steps, from identification and cleaning of data from a sensor rich environment, up to the detailed analysis and prediction control of thermal energy in a room.
Following this, the research question is further divided into some very important sub-questions, also based on the objectives, which shall help structure the entire research into different phases

Objective b) Establishing a generic automated model to provide detailed graphics with regression analyses of the functioning of the room and the building.

1. *What type of mathematical tools or other statistical methodologies are needed to arrange, clean and organize big data sets for building energy simulations?*
2. *What are the major type of graphical analyses that help determine the functioning of a room in a school building?*
3. *What types of sub-datasets are most useful to analyze the functioning of an office or school room and what type of analyses can be extracted from each?*

Objective c) Providing a methodology involving the use of correlation coefficients to understand the most probable and effective parameters influencing the heating and cooling demands at room level. This helps determine the most prominent losses and gains of thermal energy in classrooms.

1. *Can automated methods quantify the affect different parameters have on the overall thermal demand of a room?*
2. *How useful are mathematical & statistical tools such as correlation and partial correlations in defining thermal energy balance of a commercial room?*

Objective d) Utilization of the established parameters to perform multivariate regression analyses. These could be used in turn to train models for prediction of thermal energy demand.

1. *What are the major difference between backwards (FITLM) and forward(STEPWISEFIT) propagation algorithms in Multilinear Regression Models?*
2. *How well can MLR models train and predict the energy consumption of the HVAC systems, and what generic validation techniques can be used for the same?*
3. *What is the significance of using various timesteps of data, and what are the issues if any, faced by training models with both heating and cooling demand?*
4. *What are the major challenges faced by MLR models for training and predicting data?*
5. *Can past year input vector data replace the present year predictor variables, for predicting thermal energy demand?*

Objective e) Utilization of ANN techniques to improvise on the prediction demand of energy.

1. *Can Artificial Neural Networks be developed for room level thermal energy prediction?*
2. *How well can these ANN networks answer for the non-linear component present in the input parameters?*
3. *What are the advantages and disadvantages of the ANN network vs and MLR network*

What follows is the unique adapted methodology for this research, incorporating models of different types, based on the functional objectives listed out before.

4.1 Adapted Modelling Approach

The selection of the most appropriate model seems to be a rather cumbersome task. Each model has its own pros and cons. The HHS building as already mentioned is a *smart building* with a sensor rich environment focused towards controlled energy use. This includes controlled heating and cooling, ventilation and lighting demand controlled by occupancy rate, etc. The Control Systems also called the Building Energy

Management Systems (BEMSs) acts as a through way for the mechanical and electrical equipment of the building to interact with these sensors. Alongside this, since the major research is concerned with prediction models for thermal energy on an hourly level, the model must be dynamic.

A methodology will be developed based on a funnel topology (from rough to fine) of methods from simple graphical and analytical solutions to statistical models which are based on either simple linear and multivariate regression or ANN. The model will be divided into sections, running parallel to the literature survey done before. This would be a unique blend of different modelling techniques aimed at achieving the aforementioned objectives.

4.1.1 Energy Profiling and Monitoring

A chief technique in evaluating the performance of a building, is to study its energy consumption profiles, against the backdrop of certain parameters. This helps in explaining the functionality and effectiveness of the building systems. The Building EQ report of 2010 [33] claims that a monitoring and evaluative system for building energy demands, can help in performance enhancements for faulty systems, and render energy savings possible. Benchmarking building performances against that of other buildings helps in a comparative analysis.

Following the trends used in the Building EQ report, energy profiling and monitoring will be the first step of this thesis. It will use the data recorded from the sensors in the rooms, and deliver visual representations. This box will also be able to showcase seasonal, and occupancy related (Working and non-working hours) energy use.

4.1.2 Attributes Effecting the Thermal Balance

Once a grasp is developed on the energy patterns and HVAC functioning patterns of the room(s), a clear guideline can be established towards understanding the key parameters which affect the thermal balance of the specific room. This is done using multiple and partial correlation plots. All recorded variables selected by the user shall be analyzed to find dependencies and interdependencies amongst them and the thermal energy. A research towards this would deal with answers relating to major causes for energy demands in a specific building, quick fixes to reduce these demands, faults, goodness of the thermal building structure, etc.

4.1.3 Prediction Models – a Comparative Study

The last stage of this research toolbox will be to incorporate the attributes found above to apply them for predictive models. The use of MLR and Machine learning (Artificial Neural Networks) were the

two important models chosen for this purpose making the adapted approach as a 'data-driven approach with a dynamic black-box model'. A research towards this would result in answers related to possibilities of prediction and its accuracy at the room level of office/school buildings. Using two models would also help to throw light on comparative answers related to the effectiveness and disadvantages of both. Further explanation shall be provided in the chapters 8-9.

4.1.4 Flow Scheme

Below (figure 5) a descriptive flow chart has been portrayed, which is a description of the entire flow-process adopted in this research.

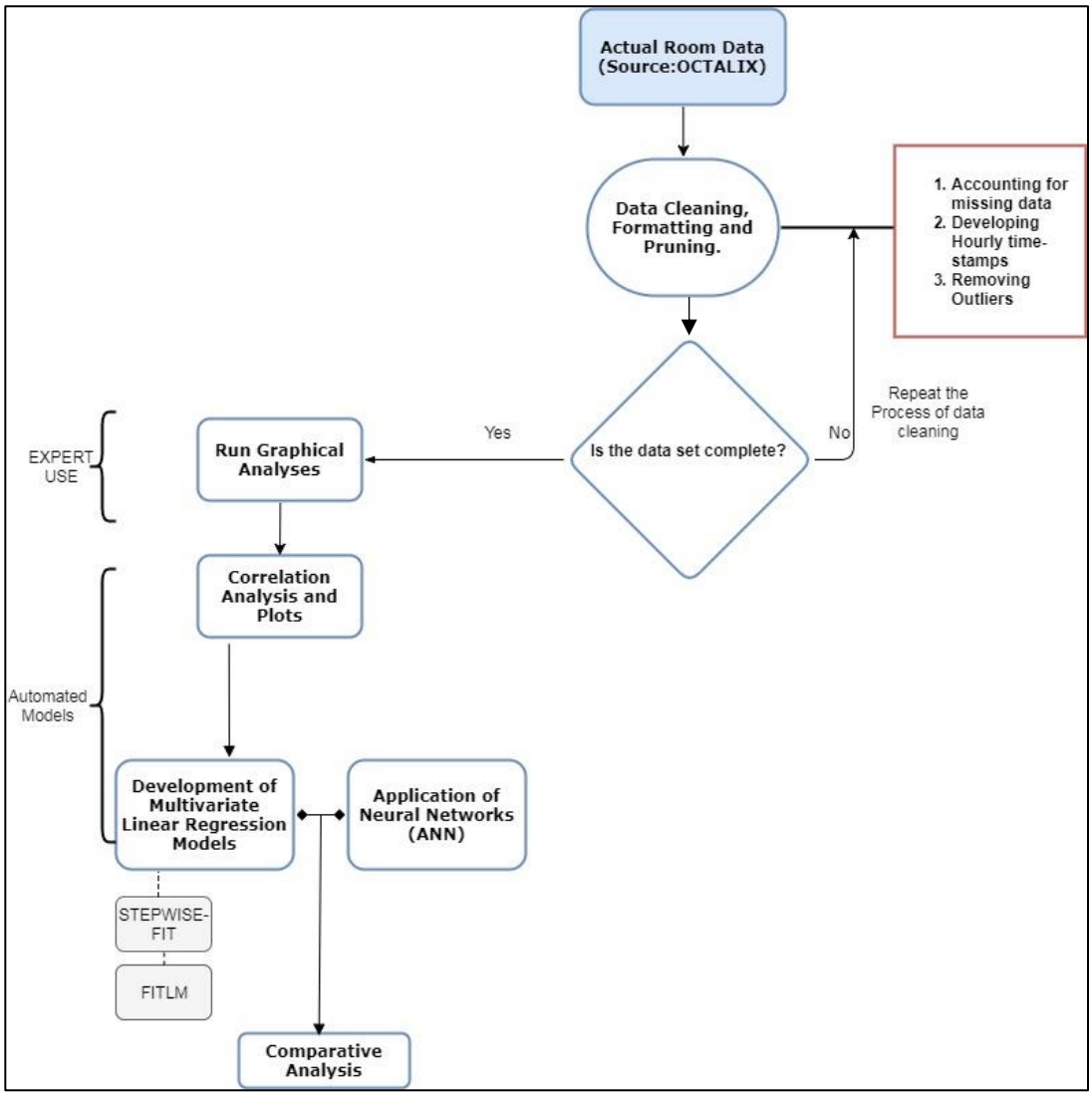


Figure 5 Flow-chart of the Toolbox structure developed during this research.

5.Data Description and Preparation

This section explains the methodology in which data is retrieved, cleaned, organized and formatted to be utilized in the three experimental models. The information provided in this section is vital in terms of explaining to the reader the most important data-types and methods of pruning the data to best fit the set of models developed in this research.

5.1 Data Sources

The entire HHS building's HVAC is centrally controlled with the help of a sensor rich environment in each room. The controls and flows to and from the rooms, and the indoor climatic state of the room are all measured, every 6 minutes, by these sensors. This data is stored in the BMS. As stated previously, the basic data set obtained from the Octalix, and PRIVA BMS system of the HHS University building, is carefully pruned and cleaned to be used on analytical models. The data set at the room level, contains the following information, see table 3.

Table 3 obtained sensor data from the room 1075 of the HHS for 2015.

Appliance Electrical Energy (J)	Lighting Electrical Energy (J)	Heating Energy (J)	Cooling Energy (J)	Wall Temperature (^{0}C)	Air Temperature (^{0}C)	Co2 level (PPM)	Air flow (m^3/hr)

Air Flow Temperature (^{0}C)	Floor Water Supply Temperatures (^{0}C)	Floor Water Return Temperatures (^{0}C)	Side Room Temperatures (^{0}C)	Valve Position for Floor water supply [open/closed]

Another important source of data for this research was the continuously monitored hourly data, available for the outdoor climate, from KNMI (Royal Dutch Meteorological Institute) website of corresponding Rotterdam Station [4]– an open source platform by the Dutch government. The outdoor temperature, solar radiation and wind speed are selected for this research. Rotterdam was selected as the nearest weather station for our research building, the HHS, which is in Delft, see table 4.

Table 4 The variables adopted from the KNMI outdoor weather monitoring tool, along with their individual unit.

Average hourly Temperature	**^{0}C**
Average Hourly wind speed	m/s
Average hourly Solar Radiation	W/m^2
Timestamp hourly	[-]

[4] http://projects.knmi.nl/klimatologie/uurgegevens/selectie.cgi

Apart from these two major sources, other sources are the horizontal solar radiation measuring equipment present on the roof of the HHS building.

5.2 Data Cleaning and Organization

MATLAB R-2017 software [30] has been used to automate a generic model with 4 steps to clean and prune , and organize the data for the purpose of this research. This model can be applied to any of the rooms of the building. The data files undergo treatment in the following manner –

1. *First, the missing data cells need to be accounted for* - Like every system, the data retrieved have missing values. It was important to find the *NaNs* (missing values), which may lead to inaccurate calculation or abruptions during the functioning of the modelling period.

Using running average, the data can be estimated, as the method has been fruitful in similar researches by V.Parab [23]. This is done by substituting the missing value with the average of the value above and below the missing value, However, there are instances when the values were not recorded for more than two (up to days) consecutive time stamps. Under such circumstances, the particular day is removed from the data set under consideration. This is done in order to have a uniform number of hours, representative of the number of days in the data sheet. It should be noted that only 5-10 days were removed at the utmost. Thus, for a dataset of 1 year, this is perhaps no more that 2.7%.

2. *Conversion to Hourly Timestamp*- Once the missing values have been removed, the timestamp of the data is brought to a common hourly timestamp. This helps obtaining 1 value for every 10 data points of 6 minutes. It should be noted however, that indoor climatic conditions have a much smaller time constant. One may observer energy patterns changing over a 10 or 15-minute time interval. Thus, a research on 6-minute intervals is suggested to be more advantageous, (see chapters on conclusion and recommendations). The available outdoor climatic data is at an hourly timestamp, thus limiting the scope of this research as well.

The averaging of the data can be done in two ways, either, by achieving hourly data from a subset of the 6-minute interval, i.e. averaging the data over an hour, or by simply picking a value of an hour period (in this case 10 readings of 6-minute intervals), as the average, which is less accurate. The temperatures were all averaged using the general method of subset averaging. Values like the thermal heat or cold demand, or electrical demand, were integrated over the 6-minute intervals to obtain hourly (Wh) energy use.

3. *Conversion to standard units* - The Energy recorded by the sensors is in Joules. So, the value is converted in Watt-hour (Wh). This is done in order to maintain uniformity between different variables.

4. Removing Outliers - The only remaining task for pruning the data is to remove outliers. Outliers are points in a data range which are invalid, and have no significant meaning with respect to other data points of the same category. This is due to auto-resetting of the sensors, or due to a malfunction. An example would be a recorded value of 10 MJ of heating energy in 6 minutes, which is practically not consumed whilst the average value is 1 KJ. These outliers are removed by the simple and common method of Mean Absolute Deviation (MAD). First for each value of a given vector, the median is calculated of a window composed of the value and it's six neighboring values. The standard deviation of each value is then calculated about its window median range. For values above a standard deviation of ± 3 are considered to be outliers, which correspond to not more than 1.2% of the data. This method has been used mainly because there are very few points in the data which form obnoxious values, as they lie extremely far in the distribution range. They were replaced by the median values.

In conclusion, the processing of data (especially real-time data) is an important task towards atomized model development. This chapter has shown and given answers for the sub research question –

"*What type of mathematical tools or other statistical methodologies are needed to arrange, clean and organize big data sets for building energy simulations?"*

- Using mathematical techniques such as Weighted Running average, the missing data can be substituted with appropriate, high probability data.
- Using methods such as Mean Absolute Deviation (MAD), the outliers of a data can be kept in check.
- Statistical tools can help look into to the normality of a dataset, and whether the data is normally distributed or not.

Once the data is organized, pruned and ready to be used, models can be developed to begin with the analysis phase. What follows is a study of the first step-toolbox developed under this research, the graphical analysis.

6.Graphical Analysis

6.1 An overview

This chapter deals with the first applicable model of this toolbox - a graphical analysis of the data recorded from the sensors. The energy demand in a room of an office or a school/University building depends strongly upon (i) Outdoor Climatic conditions, (ii) The Indoor climate (iii) The Occupant-related Energy use [34] [35]. The use of energy for maintaining a comfortable indoor climate has been known to account for more than half the energy consumption in School and Office buildings [35]. To be able to understand the building (room) and its energy performance, it is important to assess these three parameters.

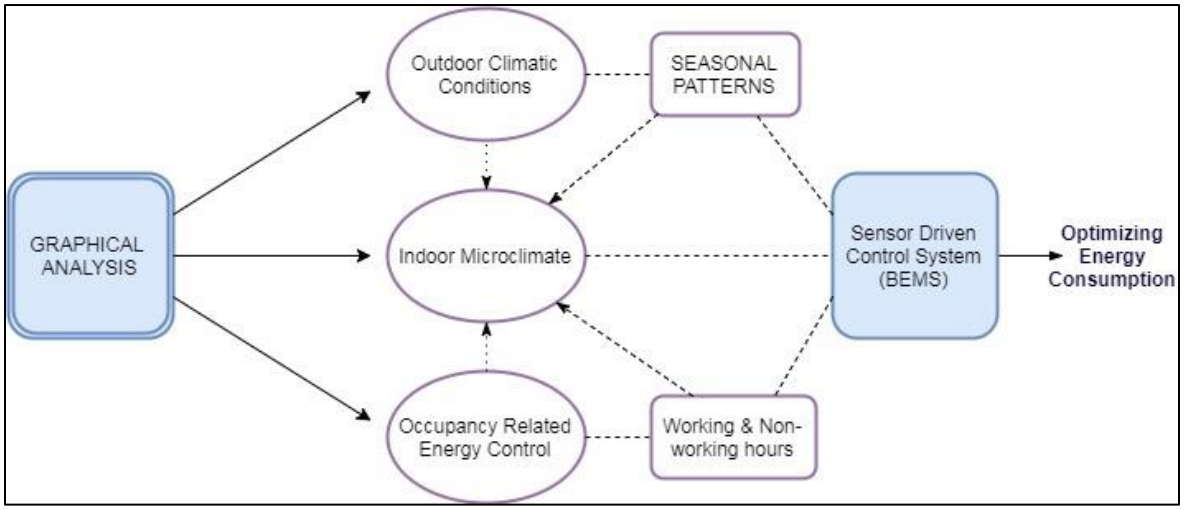


Figure 6 Flow scheme of the graphical analysis step-toolbox.

As mentioned before, in a sensor, rich environment such as the HHS rooms, the conditions of the indoor microclimate and the parameters affecting its stability (outdoor climatic conditions and occupancy) are recorded for analysis. Using this data, the functions and controls of the systems in play at the room and building level can be assessed. Figure 6 shows that the indoor microclimate control is dependent on the outdoor climatic conditions and occupancy related energy use [34] [36]. Monitoring control schemes related to energy use are important especially in commercial buildings, such as the HHS, where the occupancy plays an important role in the overall energy consumption. Today there are several models available for experts, which help detail out analysis of a room's and/or building's energy demand, see for instance the Building EQ reports [33] or the Dutch RVO [37]. The motive of this chapter is to find methods of automating the most important analyses which could help understand the dynamic functioning of the room.

For example,

1. the rooms are not always in use, and this brings in the question of *working vs non-working hours* and the energy patterns associated under these two categories of hours in the room,
2. or even the seasonal pattern of energy use in certain buildings, wherein summers experience a different pattern of ventilation and cooling as compared to winter.

An automated model helps to analyze the most important patterns and parameters affecting the Indoor microclimate and help an expert investigate ways in which the building systems could optimize energy use.

Thus, the *first* most important objective of this model is to help the reader form certain first hand qualitative and quantitative results about the room by a robust visualization of data. This will help the user (an expert) analyze the room's indoor climate, energy use, and thus the effectiveness of the controlled HVAC systems.

The *second* objective of this model is to investigate the most important type of analyses which should be automated, and therefore the most important sensors needed to assess a room's performance.

Thirdly, the maintenance of proper indoor climate, especially in rooms with high occupancy rate, demands for both thermal and electrical energy. However, buildings, including highly controlled buildings, tend to develop faults in the systems, which bleed energy whilst trying to maintain thermal and physical comfort in a room. This study would thus also help in quickly *identifying faults* which have a negative influence on the indoor air quality (IAQ) balance at the room level.

As explained in chapter 1, the research uses a typical classroom, number 1075, of the HHS as a case-study (see to section1.3.2).

Qualitatively, the only information available beforehand regarding this room, is that it;

- *has certain occupant related HVAC settings (sensor-based controlled environment)*
- *has a south facing wall, with large portions of window (refer to images in the appendix A2.1 for approximations)*
- *has three internally facing walls, including the roof and floor*
- *is functioning as a classroom*

6.2 Indoor climate

According to the Rafsanjani [34] and other authors [29, 30] the indoor microclimate is connected to four main parameters which determine the effective health and living conditions of people in buildings.

1. *The Indoor Thermal Environment*
2. *Indoor Air Quality*
3. *Indoor Lighting Levels*
4. *Indoor Noise Levels*

Indoor climate needs to be highly stable in rooms against varying seasonal, climatic and occupancy levels. The analysis of control measures for such operation helps determine the functioning of the room and also shows a strong link of the economy of energy use of the room [36] [38].

6.2.1 Indoor Temperature

Data analyzing has been done using linear regression tools on MATLAB [30] and the graphical results have been placed in this section along with detailed explanations of the findings. The indoor thermal environment is judged by the combination of elements, or factors, such as temperature, humidity, heat radiation and air movement [36]. Temperature of the air inside the room is one of the major factors influencing a comfortable indoor climate, so plotting the indoor temperature against several parameters like outdoor temperature, hour of the day, solar radiation or wind speed will give a first visual idea about how responsive is the indoor climate to occupancy and outdoor parameters.

In the case study of room 1075, looking into the thermal state of the room, we can observe as follows (see figure 7 below) - A constant temperature can be seen being maintained throughout the day, and the wall temperature, measured towards the inside, is only slightly lower than the indoor air temperature.

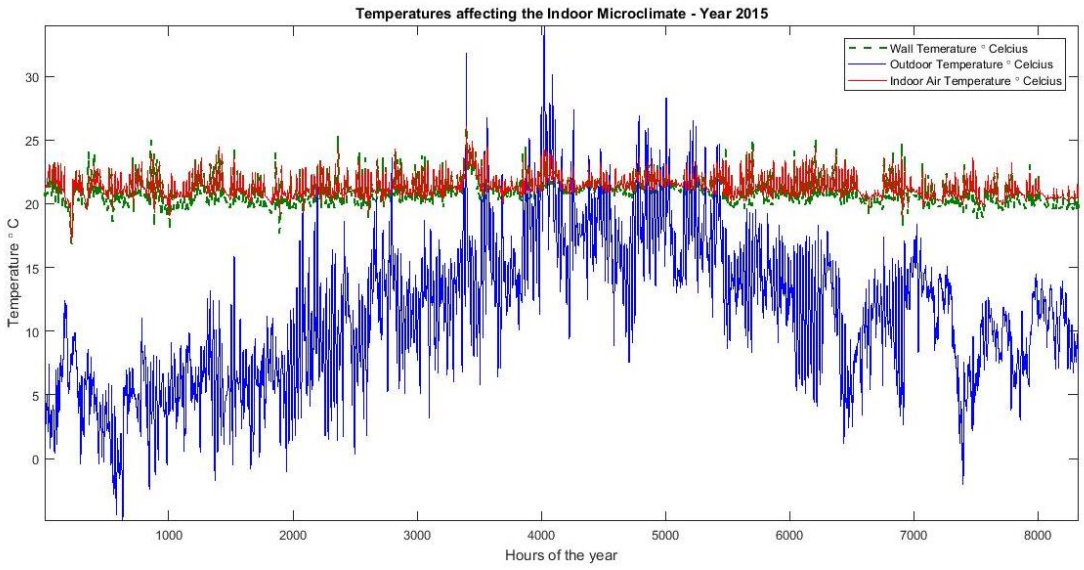


Figure 7 Indoor air, Wall surface and Outdoor air temperatures for the entire year of 2015 for room 1075.

Using the graph plotted above (figure 7) an expert can gather certain first-hand conclusions immediately.

a. The indoor temperature is *highly stable* and maintained so throughout the year. This means that the room has a set point temperature of 21-23 ^{0}C.

b. The indoor temperature is *not altered with regards to* the outdoor temperatures, which have a range from -5 to almost 30 during summer. This also means that the room has a very good thermal envelop, or a very poor thermal envelope with a highly efficient HVAC control system.

Point ‘a’ above can be further verified by the figures 8 and 9 plotted below.

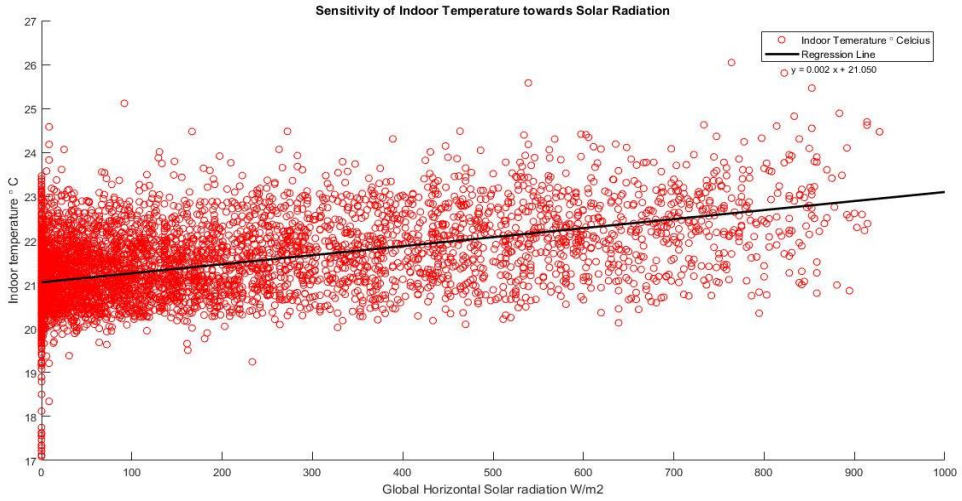


Figure 8 Indoor temperature plotted against outdoor solar radiation. The slope of the linear regression line is almost 0 (m=0.002) with an intercept at 21.05 ^{0}C.

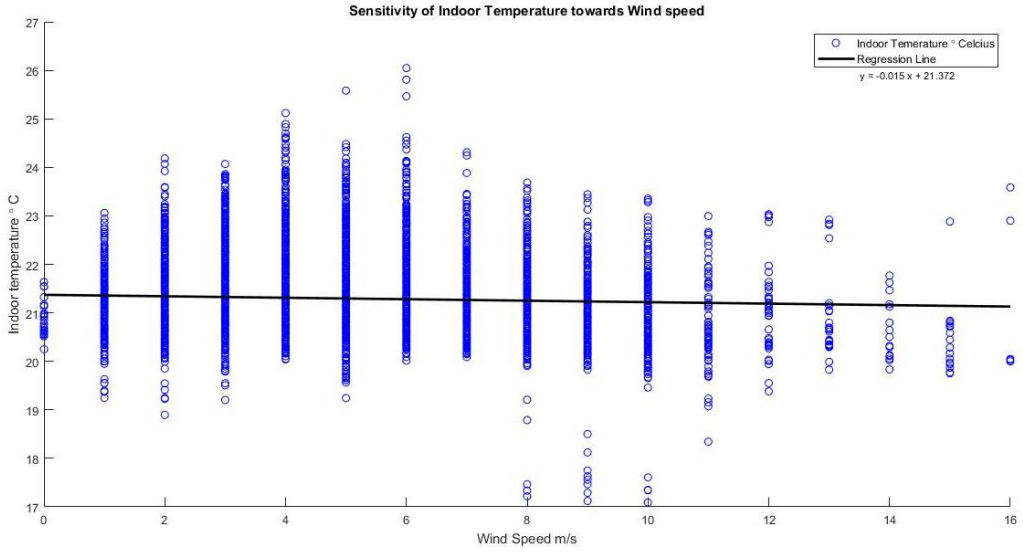


Figure 9 Indoor temperature plotted against outdoor wind-speed. The slope of the linear regression line is almost 0 (m=-0.015) with an intercept at 21.4 ^{0}C.

These two figures depict the sensitivity of the indoor air dry-bulb temperature with that of the outdoor solar radiation and wind speed. These lines were developed by using a simple linear regression analysis on MATLAB [30],using the least square fit function [39].

With regards to solar radiation – There is a 2°C window in indoor air temperature variation with respect to the solar radiation. This room has a wall facing the south side with ~50% windows. With extremely sunny hours (>600W/m^2, see figure 8) an increase of almost 1-2°C suggests that the rooms are provided with shutters/blinds, to avoid massive fluctuations in the indoor thermal balance.

With regards to the wind speed – The wind speeds were measured in bins by the KNMI, and thus these bins can be seen in the plot figure 9 above. The wind speed also seems to have little to no effect on the indoor temperature, suggesting lower infiltration from windows or creeks, and controlled temperatures of ventilation airflows.

This suggests that the building and the room indeed have very good HVAC system and/or insulation.

6.2.2 Indoor Heating and Cooling

The next step in the graphical analysis is to plot the heating and cooling energy against various variables. The thermal energy demand of the rooms is met by water-based floor heating/cooling pipelines. The energy used is estimated by the sensors based on the flow rate and temperature difference of the input and output water flow of the pipes supplying to a room. This is calculated each six minutes in Joules, which is converted to watt over an hour by aggregating the total supple over 10 intervals. The values have been plotted below in figure 10.

As mentioned the ATES, is responsible for almost all the heating and cooling of the building. What is interesting to note from this graphical analysis is that the number of hours of heating and cooling for the entire year for this classroom, is limited to only 486 and 710 hours respectively, out of the 8327 hours of data from the year. This leads towards two possible conclusions about the room and the building in general;

a. *It seems that the thermal insulation of the room is good leading to little heat and cooling hour, see table 5, while the indoor temperature remains stable*
b. *There may be the possibility of a secondary source of thermal energy supply.*

Table 5 - Thermal energy demand from floor heating/cooling system by the room 1075 for the year 2015.

Parameters →	***Total Number of Hours***	***Number of Heating Hours***	***Number of Cooling Hours***	***Number of Hours with no Thermal Energy***
Value	*8327*	*470*	*701*	*7156*
Percentage	*100%*	*5.64%*	*8.41%*	*85.93%*

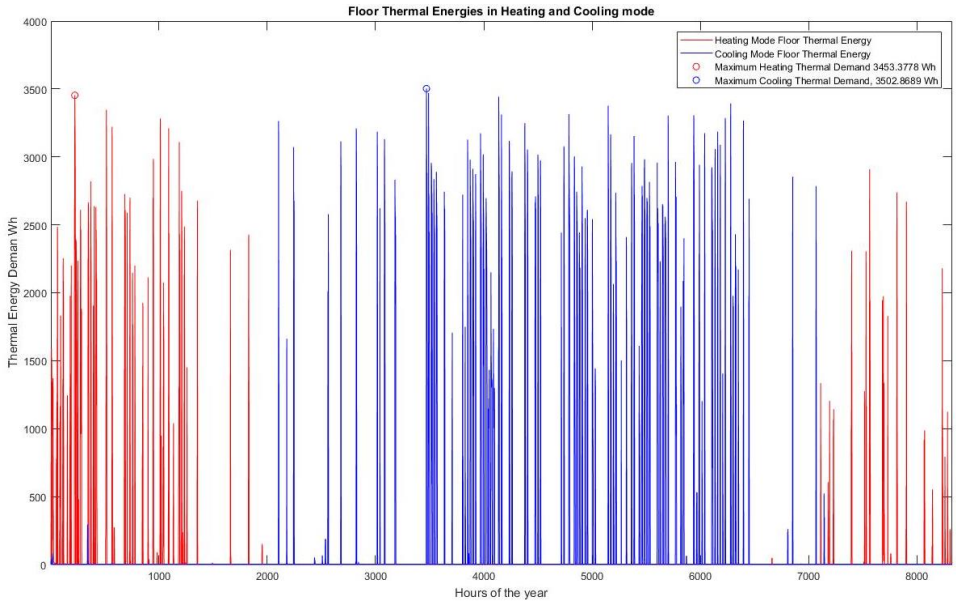


Figure 10 The thermal energy of the floor heating/cooling during the entire year - 8327 hours.

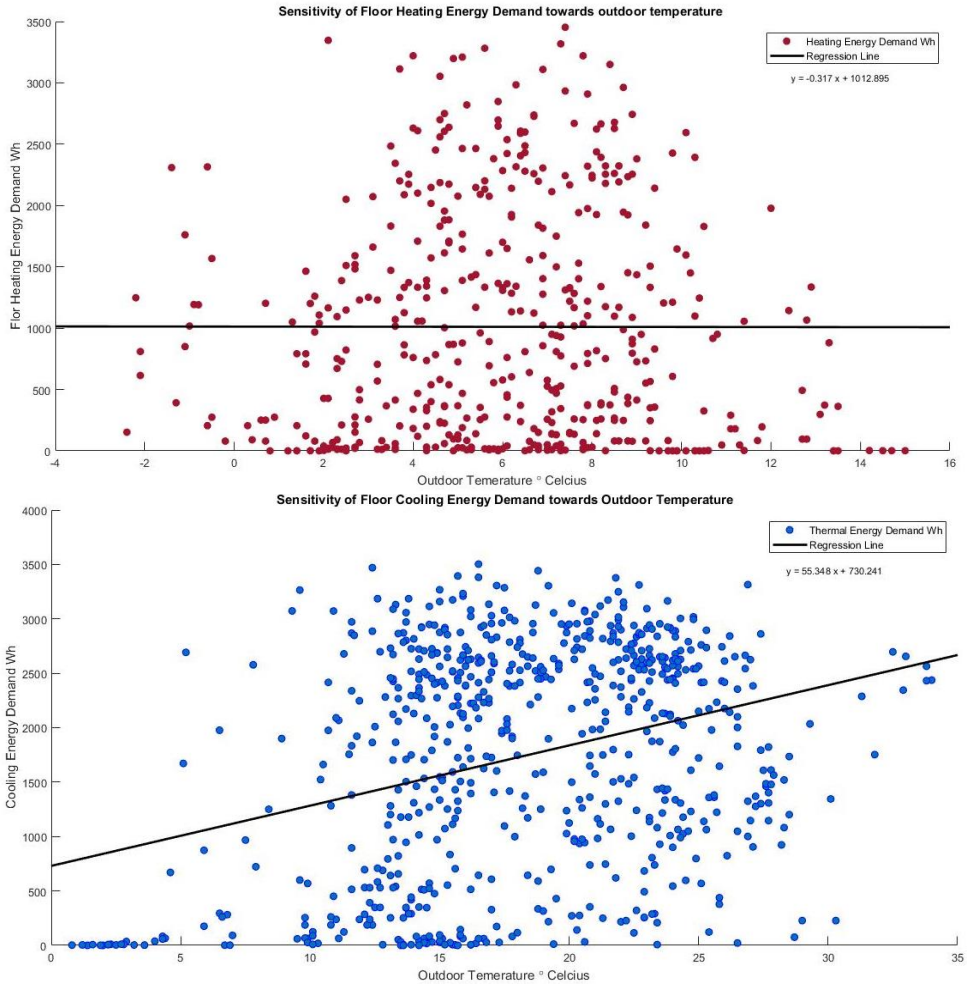


Figure 11 a and b - Sensitivity of Floor heating and cooling demand to the outdoor temperature. The slopes of the linear are -0.37 and 55 respectively.

The heating energy supplied solely by the floor to the room is almost independent of the outdoor temperature variations over the entire period of heating energy demand (figures 11 a and b). The regression lions plotted via MATLAB using the least square fit function, indicate a slope of -0.3. According to Building EQ, this value of almost 0 slope indicates that building has a fairly good thermal insulation with a very good HVAC system.
The cooling energy on the other hand, is quite dependent on the outdoor temperature changes, with a slope of m=55. This indicates a quite linear relationship between the cooling energy demand and the outdoor temperature.

With a good insulation and/or good HVAC control system in place, the indoor climate is less vulnerable to the outdoor climatic conditions [35] [36], as can be seen from the low temperature and thermal energy supply variations with regards to solar and wind speed (figures 8, 9 ,11 a and b). However, the indoor climate under such circumstances (office and school rooms) is most affected by occupants, processes and activities occurring inside the room [40]. A proper HVAC installation is vital for a balance in indoor temperature. The ventilation in buildings with such automated BEMSs are controlled in order to [40];

- *Distribute adequate quantities of air through the room, to satisfy the need of the occupants*
- *Remove odors, and contaminants by flushing out the air within the room through a mechanical exhaust system.*
- *Provide thermal comfort to occupants* – Although this is not the main aim of ventilation in European buildings, air may be heated or cooled in an Air Handling Unit (AHU) before injection into the room to contribute towards the thermal comfort levels.

In the building of the HHS, Delft, rooms are supplied with air from an AHU after it has been treated to maintain thermal comfort within the rooms (see section 6.3). This variable tends to add to the overall thermal energy of a room in the building, when there is ventilation. Moreover, since the data is recorded every six minutes, upon aggregation over an hour, there are instances wherein the room may experience both heating and cooling in the same hour (either from floor systems and/or the conditioned air). Thus, a term, net thermal energy demand is used in this research to represent the overall thermal input to a room in a given hour. A positive value means net heating and negative value would indicate net cooling.

6.2.2.1 Net Thermal Energy at Room Level

One of the more important variables which will be used ahead in the predictive models is the net-thermal energy demand. Two main sources accounted for calculating this net demand are thermal energy;

From ATES and Heat pumps -The heating and cooling energy supplied by these two sources are transmitted via a floor water-based system in this classroom. This is recorded as the primary source of thermal energy as explained above, see figure 10.

From Ventilation supply air -When the BMES spends extra energy in cooling or heating the supply air flow to the room, it adds to the net thermal energy demand.

Therefore;

$$\text{Net thermal energy supplied from HVAC} = Q_{transmission\ floor} + Q_{Ventilation}\ [\text{W}] \quad \text{Eq-6.1}$$

Wherein,

$$Q_{transmission\ floor} = Heating + (-Cooling)[\text{W}] \quad \text{Eq-6.2}$$

$$Q_{Ventilation} = \dot{m} \times \rho_{air} \times C_{p(air)} \times (T_{indoor\ air} - T_{air\ supply})[\text{W}] \quad \text{Eq-6.3}$$

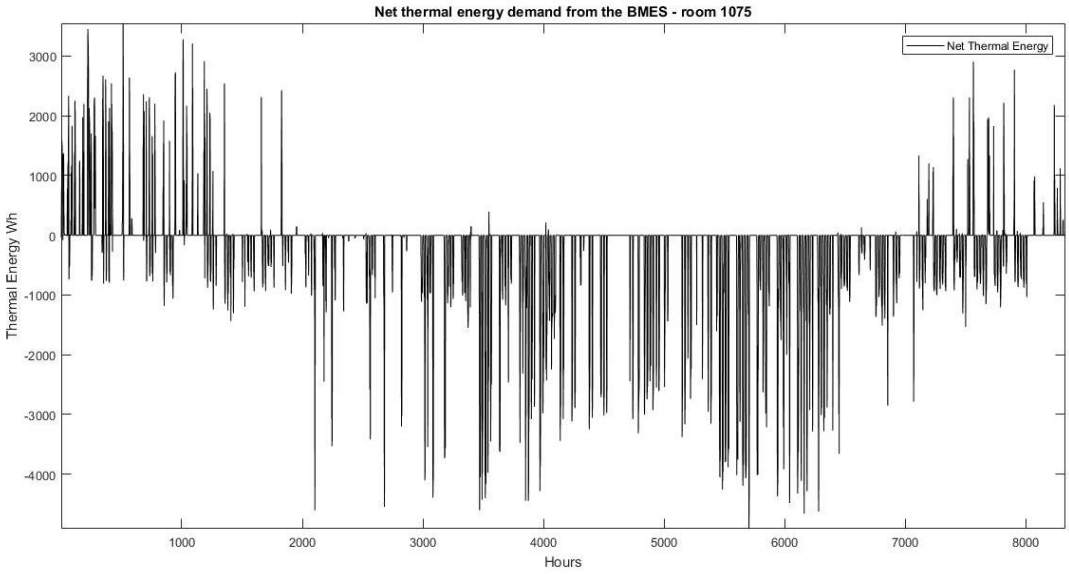


Figure 12 Net thermal energy demand over the entire year at an hourly average for the room.

The use of net thermal energy demand is imperative to this research in the correlation as well as the predictive models. This is the variable which needs to be predicted in order to estimate demands of thermal nature at room levels.

Other yearly patterns generated by this step-toolbox have been placed in the appendix A3 such as

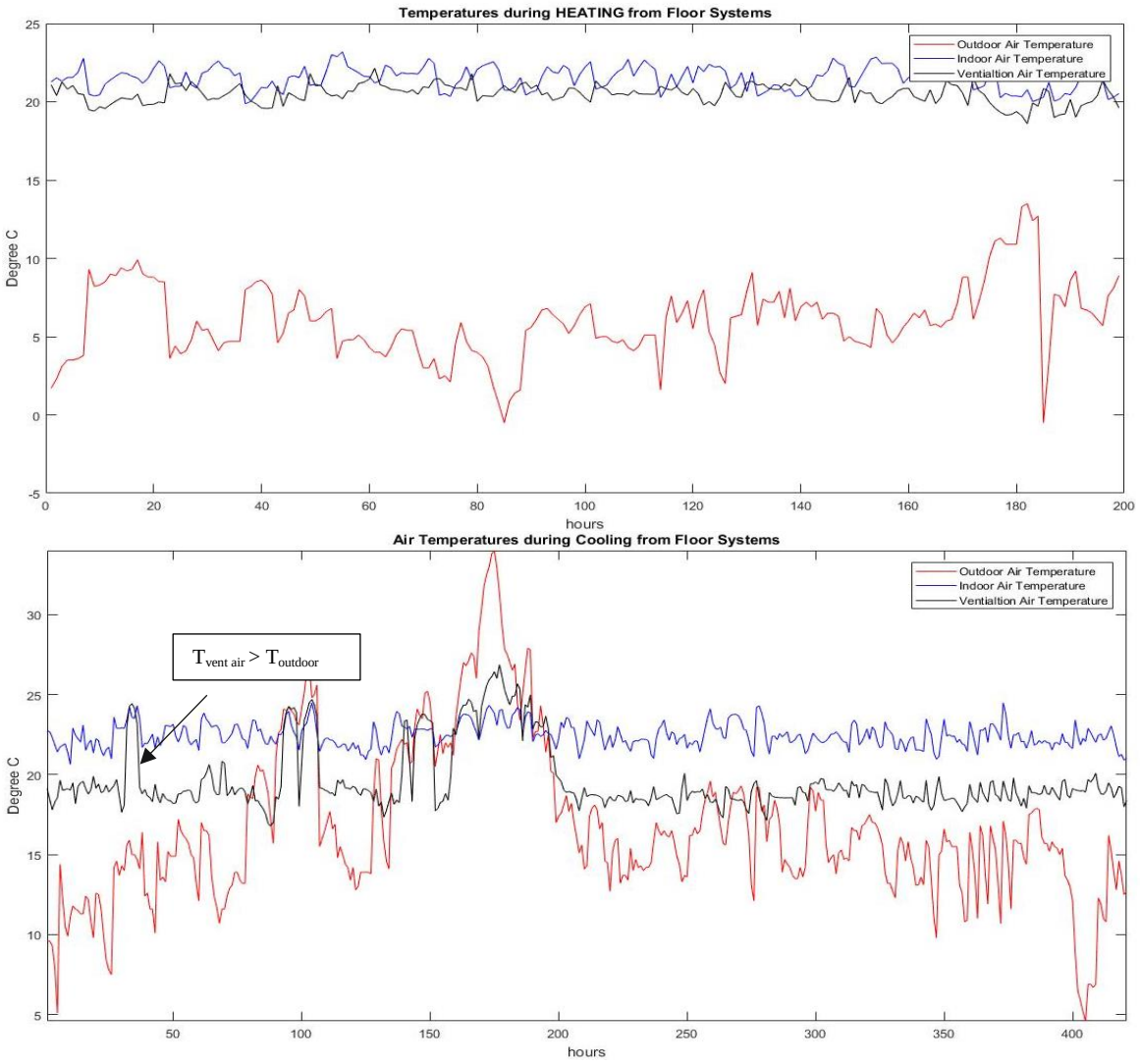
a) The electrical demand through the year (both lighting and appliance)
b) The change in indoor air temperature with altering levels of presence (in terms of CO^2 concentrations).
c) Sensitivity of Thermal Energy demand towards Solar radiation.

6.3 Floor and Supply-Air Heating and Cooling.

6.3.1Supply Air Temperature for Heating And Cooling

The HHS building has two large Air Handling Units (AHUs). This air is conditioned, by filtering, heat recovery, heating or cooling, and humidifying or dehumidifying. Since the HHS rooms are designated classrooms, these interior spaces may need cooling to compensate for the heat generated by occupants, appliances and lighting. From graphical analysis, it is thus interesting to see the variation in supply air temperature against indoor and outdoor air temperature.

Below three images have been generated from the model, which helps show the variation in the supply air temperature during different modes of operation within the room. In the figures below, only hours when the room was occupied have been plotted. The first image (figure 13-a) shows the supply air temperature variations whilst the floor system is on heating mode, figure 13-b showcases the temperature fluctuations during cooling hours and figure 13-c shows the temperature in the absence of any floor heating or cooling.



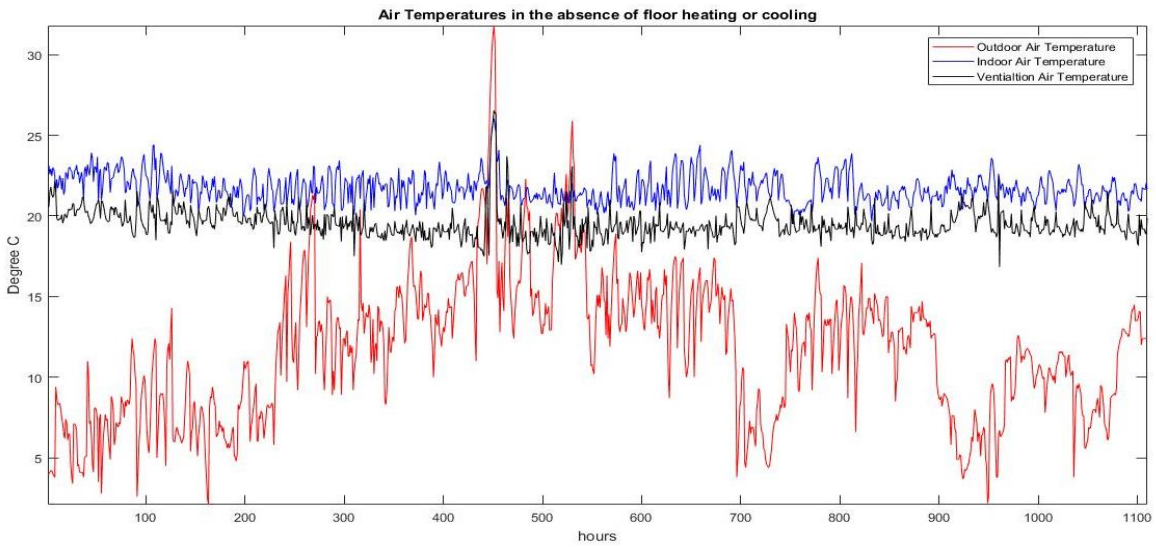


Figure 13 (a-c) Air temperatures of the indoor, outdoor and supply air, during ventilation and (a)floor heating, (b) floor cooling, (c) no floor heating or cooling hours.

It is visible that the supply air temperature is always in between the indoor and outdoor air temperature.

From image 13 a, representing the floor heating mode, it is visible that the supply air temperature is mostly lower than the indoor air temperature. This is the energy from wither exchanging heat (via a heat exchanger) with the fresh air, or heating the air in the AHU. However, there are times when the air is heated up more than the indoor air temperature (marked in green). Under such circumstances, the additional heat is added to the thermal energy of the room, leading to a higher *net thermal energy demand*. Another important analysis in this model is *fault detection*. If in any case, the supply air temperature is lesser than the outdoor temperature, or is quite low, it would suggest a fault with the heat exchanger or the fact that the AHU is not functioning appropriately. Although, this is not seen in the case of this room the model can however, show such faults (if existing) for other rooms.

During the warmer months, but also seen in winter, there is a certain cooling demand. The supply air temperature is lower than the indoor air temperature, see figure 13 b, however almost always greater than the outdoor air temperature.

In any scenario, a basic amount of fresh air is needed to maintain air quality inside the room. Instead of supplying outdoor air directly and leaving the thermal balance entirely up to the floor heating systems, there are two basic options for the AHU;

During cooling *– The air is cooled by the AHU or heat exchanged with the cooler inside air. Thus leading to a lesser floor cooling demand.*

During heating *– The air is heated in the AHU, or heat exchange with the warmer indoor air. Thus, leading to a lesser floor heating demand.*

Under this situation two faults would emerge;

a) *When there is floor cooling, but the air supplied is heated/or heat exchanged with warmer indoor air in the AHU.*
b) *When there is floor heating, but the air supplied is cooled/or heat exchanged with the cooler indoor air in the AHU.*

Table 6 Fault conditions detected due to supply temperature mismatch, in the ventilation units, during the heating and cooling mode of the room.

Mode of Operation	*Fault Conditions*
Cooling mode	$T_{supply} > T_{outdoor}$
Heating mode	$T_{supply} < T_{outdoor}$

Thus, during the cooling mode, there is a major fault within the AHU systems, wherein the outdoor air is being heated to higher temperatures in the AHU. This leads to energy wastage as the supply air is being heated and then cooled inside the room. However, this may fault may not be an anomaly in the control system, but a digression used in order to maintain the balance between the cooling and heating of the ATES.

Lastly, from the diagram c, with hours of no floor heating or cooling but ventilation only, it can be noted that the supply temperature of air always lies between the indoor and outdoor air temperatures. Despite lack of heating or cooling, the indoor temperature is fluctuating within a 2-3ºC gap, 24-21ºC. The input air of approximately 19-21ºC, obtained either by heat exchange, heating or cooling in the AHU. This supplied thermal energy is enough to compensate for the need for heating or cooling at the room level.

6.3.2 Floor temperatures for heating and cooling

The floor heating and cooling system at the HHS is provided using thin water pipes divulging out of a main pipeline which supplies warm or cold water based on the demand of a given section of the building. For example, consider a scenario in which a section has 5 rooms out of 8 demanding for hot water, while the other 3 rooms do not need any thermal energy. The control for these three rooms shut the water supply valves supplying to the room. However, the drawback of this system is that, if incase one of the 3 rooms needs cooling energy due to, let us say, high occupancy then this one room shall not be supplied with cold water, as majority of the rooms demand for heating. Thus, the main pipeline can carry only hot water with it or cold water but not both. This is an important information regarding the functioning of the heat and cold supply system of the building. An explanatory figure has been placed in the appendix A1.1 to further describe this mechanism.

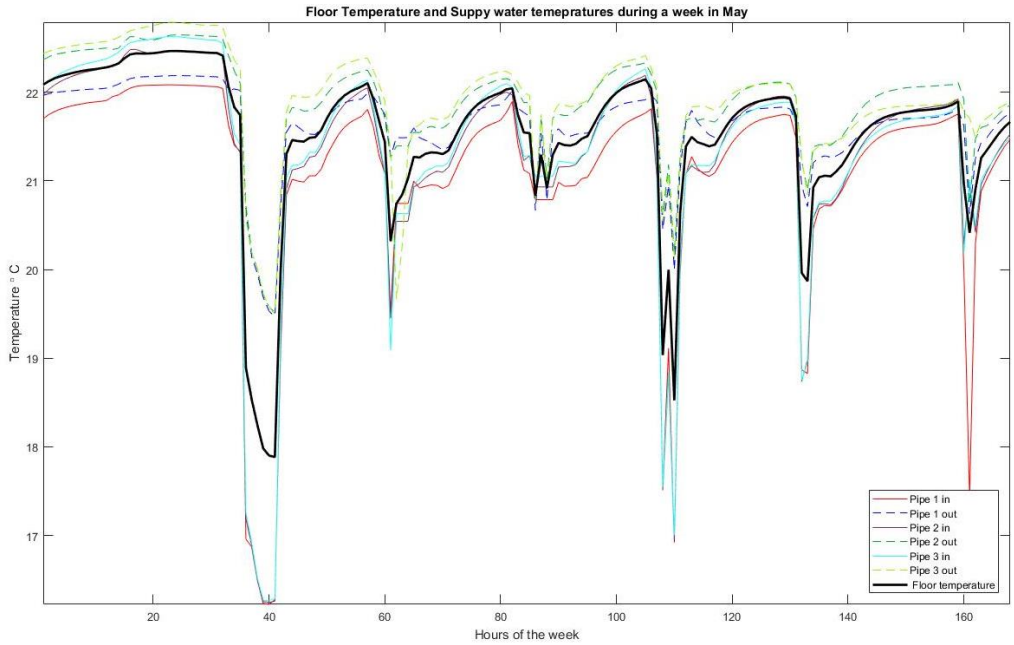


Figure 14 Supply water temperature for the floor heating/cooling system and the calculated floor temperature for the same room, for a given week in May.

In the HHS building, larger rooms are provided with 3-4 pipes depending upon the size of the room. The minimum being two pipes for very small offices. The room 1075 has 3 pipes in the floor, through which water is circulated in a single direction. This information is derived from the temperature sensors present on both ends of the pipe, just before the pipes diverge from or into the main pipeline. These pipes run parallel to each.

The floor temperature is an important parameter that is not measured at the HHS building. It was thus calculated using this graphical analysis step-toolbox by taking the hourly arithmetic mean [5] of the difference in all pipeline's input and exit temperatures. The values have been shown in the graph above over a period of 1 week (figure 14). Also, the losses for the heat transfer between the pipes and the floor are set at an assumed value of 20% for the entire year. This is because there is a slight air gap between the pipes and the floor. Also, the possibility of thermal energy travelling downwards and heating or cooling the ceiling of the level beneath has been taken into consideration, as the concrete has not been insulated to take advantage of the thermal mass. The 20% is an approximation made after consulting with the building designers.

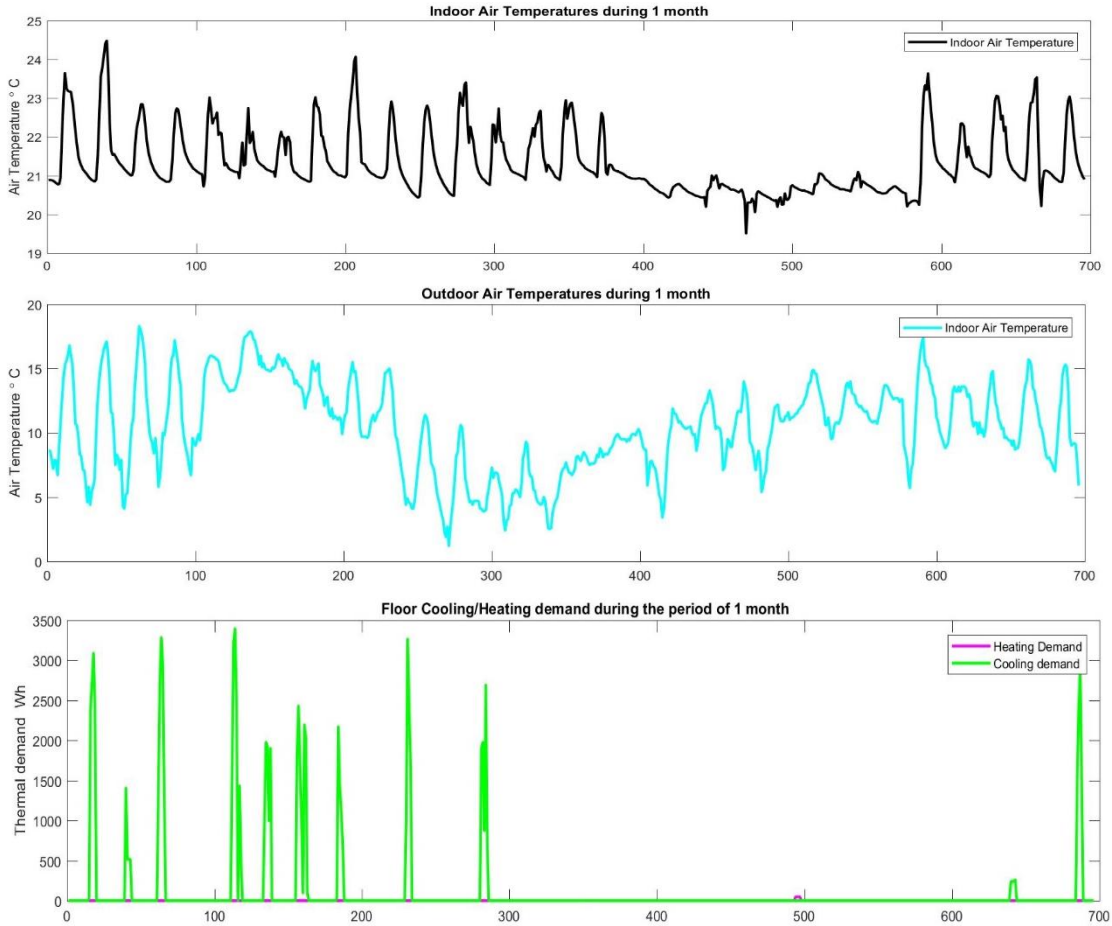
[5] The Arithmetic mean was used instead of the logarithmic mean. The flow in the pipes are unidirectional and the temperatures of one pipe does not affect the temperature in another pipe. They are from the same source and since the temperatures are measured at the entrance and exit of the pipes to the room, the arithmetic mean between these two points is accurate enough represent the temperature of the floor. However, having said this, both means can be used for the purpose.

6.4 Functioning of the room and Operating Characteristics

It is important to realize that by taking different time steps of data, the results of the operation of the building begins to differ. For instance, taking an entire year of hourly averaged data for a room in a University building, yields patterns of energy demands based on the season (heating for colder months, and cooling for the warmer months). However, monthly or weekly patterns show us detailed occupancy and ventilation patterns and their effect on indoor climate.

6.4.1 Monthly Plots

After visualization of the yearly patterns, the graphical step-toolbox has been designed for visualization of data on smaller, zoomed in time intervals as well. This is important to see the exact variation of certain vital parameters which give a picture of the physical and climatic functioning of the room during a period of the user's interest. This section will help the reader get an insight into the functioning of the room with regards to occupancy levels, electrical utilization, ventilation etc., by viewing the data over a period of 1 month of October 2015 (figure 15).



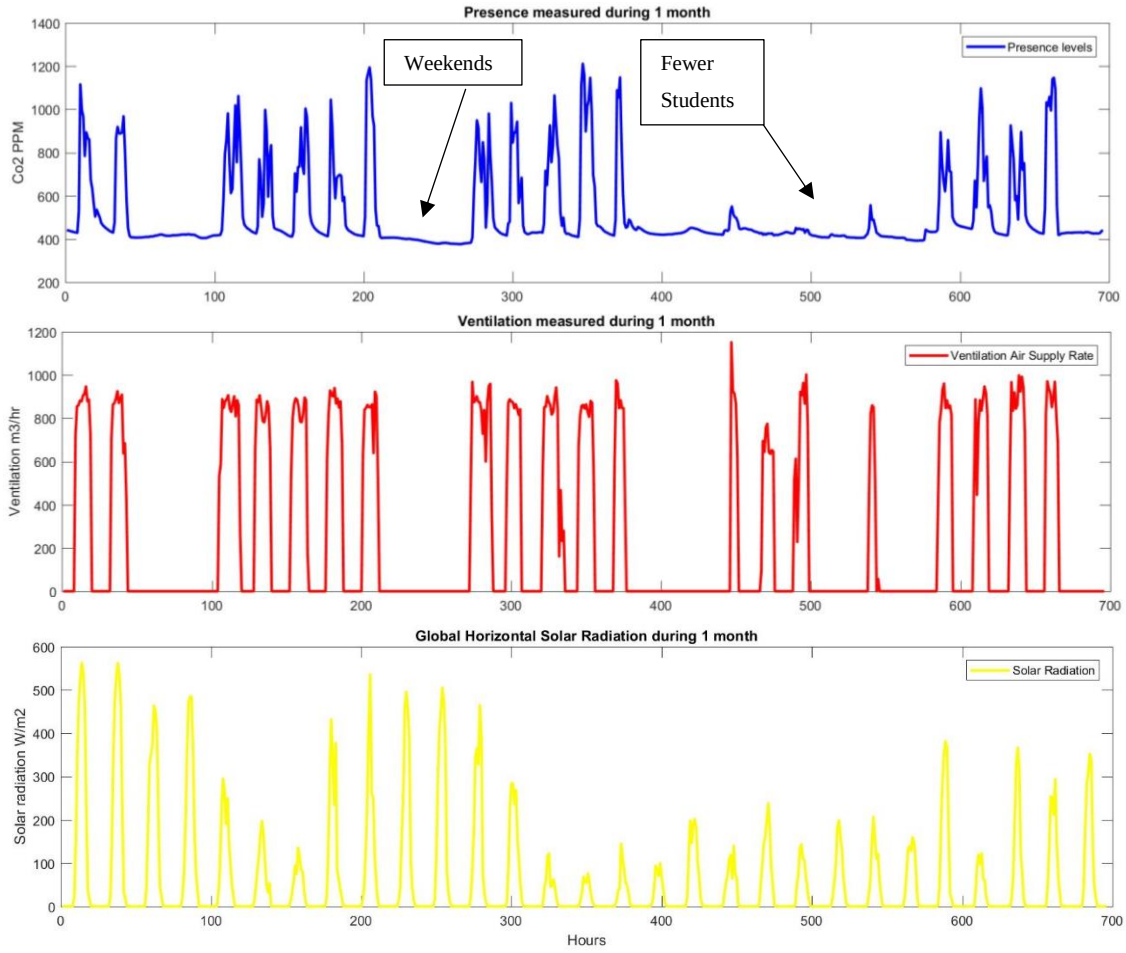


Figure 15 - Higher resolution of the data set by selecting a given month of data (October 2015).

Deductions made from such a plot are as follows;

1. It can be understood from the air temperatures, CO^2 concentrations and the ventilation flow rates that there is a clear division between working hours and non-working hours. There seems to be weekly use (Monday to Friday) of the classrooms.
2. Hours 400-600 of the month showcase something unusual – firstly there is an almost constant indoor temperature. There is lack of cooling during this week, however, at some moments, there is a high ventilation rate, which might have also regulated the indoor temperature. These high ventilation rates correspond to some small CO^2 increases.

3. The lack of CO_2 [6] present in the room during this week, could be representing the Autumn break wherein there were no students. However, they may have been cleaning schedules or small meetings held by faculty members representing a slight rise in the CO^2 concentrations.

6.4.2 Working and Non-Working Conditions

To analyze the *working and non-working* condition of the rooms the graphical analysis step box allows for the user to input the working hours of a day. In the present case, Sundays and hours from 00 – 06 in the morning, and hours from 17:00 – 24:00 in the night were considered as non-working. This was based on questionnaire results from the HHS building. All other hours fall under working category.

It should be noted that the graphics of figure 16 and 17 are based on non-consecutive hours, as they incorporate only working and only non-working hours of the month. The hours represent the functioning hours of the entire building itself. The entire building's hours were taken into account since the BMES functions on these hours and not the hours of each individual room.

[6] Presence has been characterized by the CO_2 concentrations because the presence sensors were not optimally functional. They detect motion only, and thus even with students seated inside the classroom, the presence sensor has the tendency of showing a lack of presence. Thus, concentrations above 480ppm (approximately) have been reported to reflect the presence within rooms.

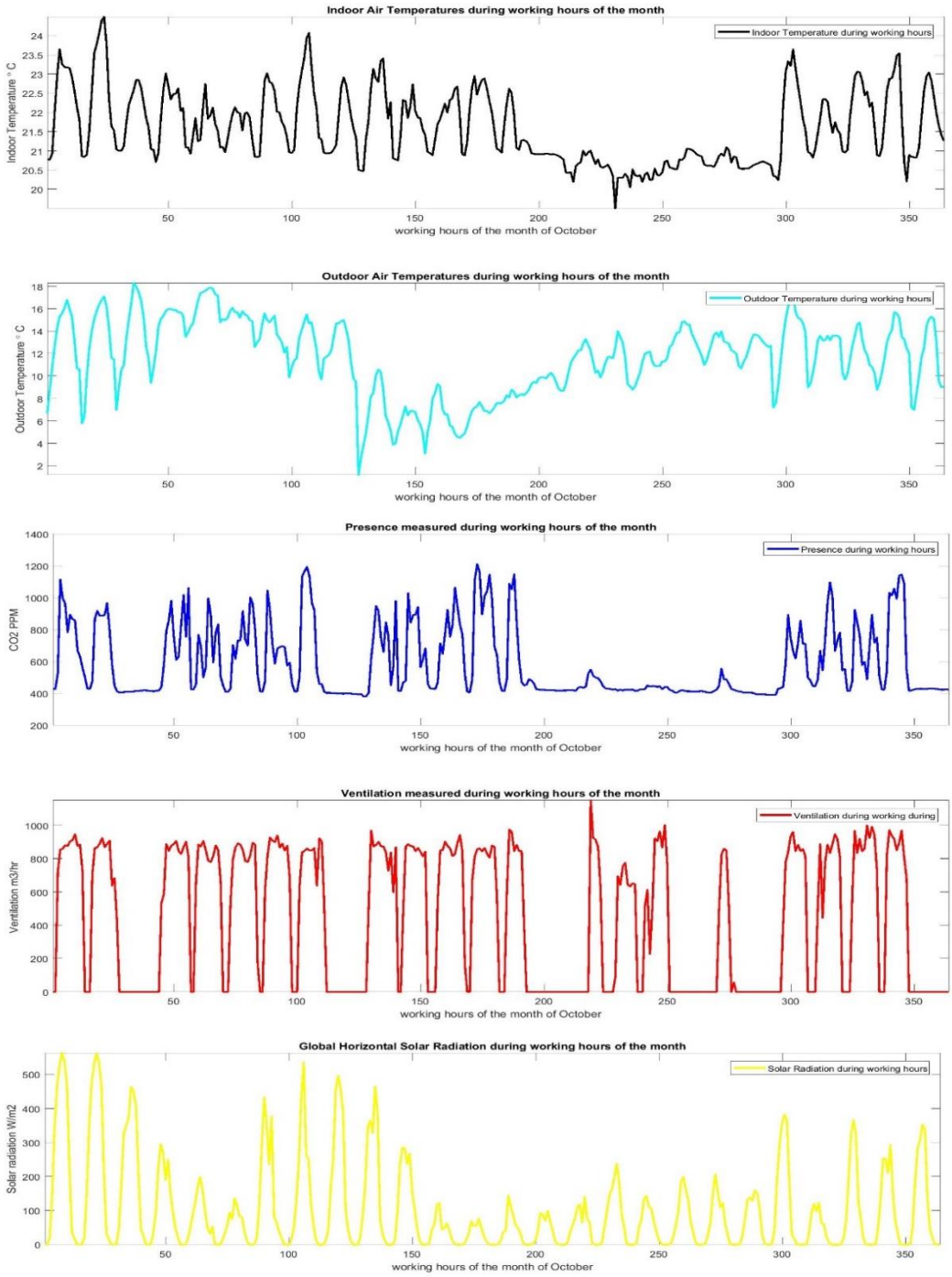


Figure 16 Data recorded only over the working hours of the room for the month of October 2015

Similarly, for the non-working hours, the following graphs were obtained, again for the month of October.

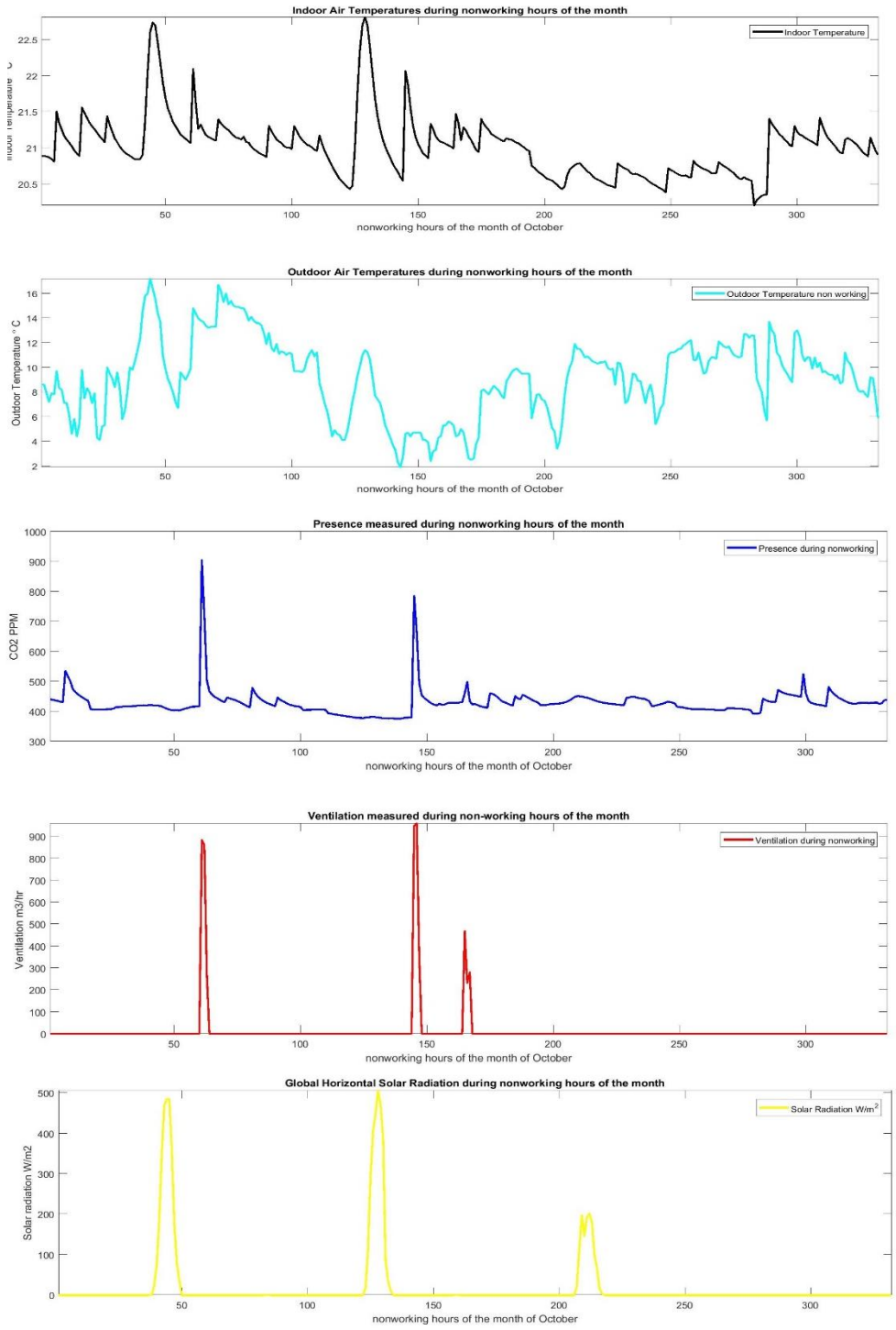


Figure 17 shows the description of several graphs based on sensor measurement, for the non-working hours of the chosen month from the entire dataset (October 2015)

As it can be seen from the two plots (16 and 17), a spike of CO_2 above 480ppm leads to proportionately controlled ventilation demand. One must note that 450-480ppm is the default ppm of CO_2 recorded each 6-minute by the sensors. Despite the controlled ventilation, we can observe that the level of

CO_2 is well above 1000ppm for several hours through the year. Table 7 summarizes the number of hours of presence and ventilation.

Table 7 Total number of hours of presence and total hours of ventilation for an entire year.

Category	*Hours*
Total number of hours	8327
Total number of hours with Presence	1797
Total number of hours with Ventilation	1730
Total number of hours of Floor Heating with Ventilation	199
Total number of hours of Floor Cooling with Ventilation	421

Thus, from the count generated by the analytical model, it can be quite clear, that the ventilation does indeed work coherently with the occupancy level of the room.
During non-working hours, with the elimination of ventilation and occupancy related thermal demand, we can see the response of indoor air temperatures to the solar radiations. This was dealt with in depth further

6.4.3 Day-wise plots

A step further – shorter timestamps, of one day have been also plotted by the analytical model. For this case study, October 10th 2015 was chosen as the day of analysis. See appendix 4, A.4.1 for these plots.

One important analysis was the lag in indoor temperature response to the increasing solar radiation. A normalized plot (-1 to +1) of indoor air temperature, the wall temperature and solar radiation can be plotted (see figure 18).

1. *The solar radiation peek occurs at t-1 hours before the indoor air temperature peaked.*
2. *The wall temperatures peak corresponds with the solar radiation peak.*

The work of Lopez [15] also found similar results with a larger delay in the temperature response of the room. On an average, it was seen that there is a lag of almost 1-2 hours in temperature response, and almost no lag in the response of the wall temperature. According to Lopez's result, this delayed response has a slight influence on the thermal energy predictive model. Thus, this delay was taken into account while improving the MLR models (see chapter 8.6).

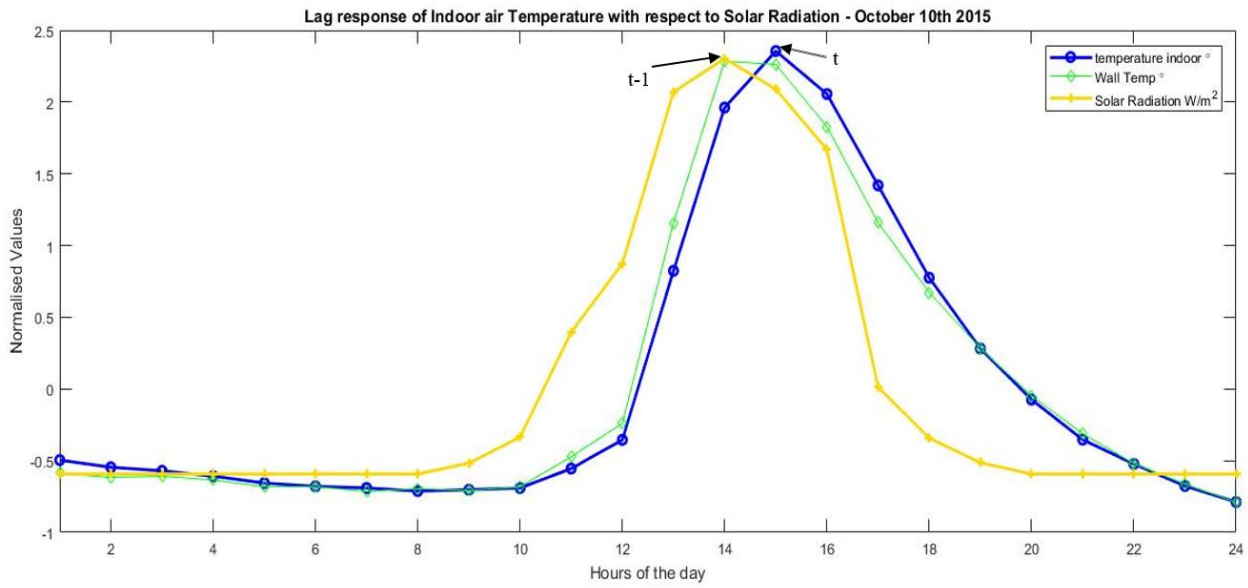


Figure 18 A normalized plot indicating the lag in indoor air temperature response with regards to the increasing solar radiation.

6.5 Seasonal Analysis

In addition to the analyses mentioned above and the ones placed in the appendix 4 (A.4.2), the seasonal analysis of data has proven to be quite useful in some cases. Especially with BEMS supported buildings, there are distinct set points in play during different seasons. The aim of this analysis is to provide automated tools for analyzing a room's energy demand seasonally, and observe and mark any anomalies that one may find in the working of the room and building. The entire year's data can be easily divided into three main seasons, namely midseason, summer, and winter. A few variables were plotted on an hourly average basis as shown below.

The figure below (figure19) is that for the summer period. We can notice;

- *The variation of 15^{o}C of outdoor temperature against the 3-4^{o}C variation in indoor temperature. These variations are lower with no cooling, suggesting the possibility of a high thermal mass and insulation as mentioned before. Since there is no ventilation flow rate during these hours (1100-1500 for example) the possibility of cooling via the AHU is also eliminated.*
- *Once again hours of both occupied and vacant room are witnessed (Summer holidays). However, we do not see a change in set point temperatures.*
- *Thus, thermal floor cooling is prevalent even during the long holiday, with the absence of students. Once again, although this may seem like a wastage of energy, it may be due to the energy needed to*

balance the ATES. A conclusion by an expert could be that the balancing strategy of the ATES must be studied and revised.

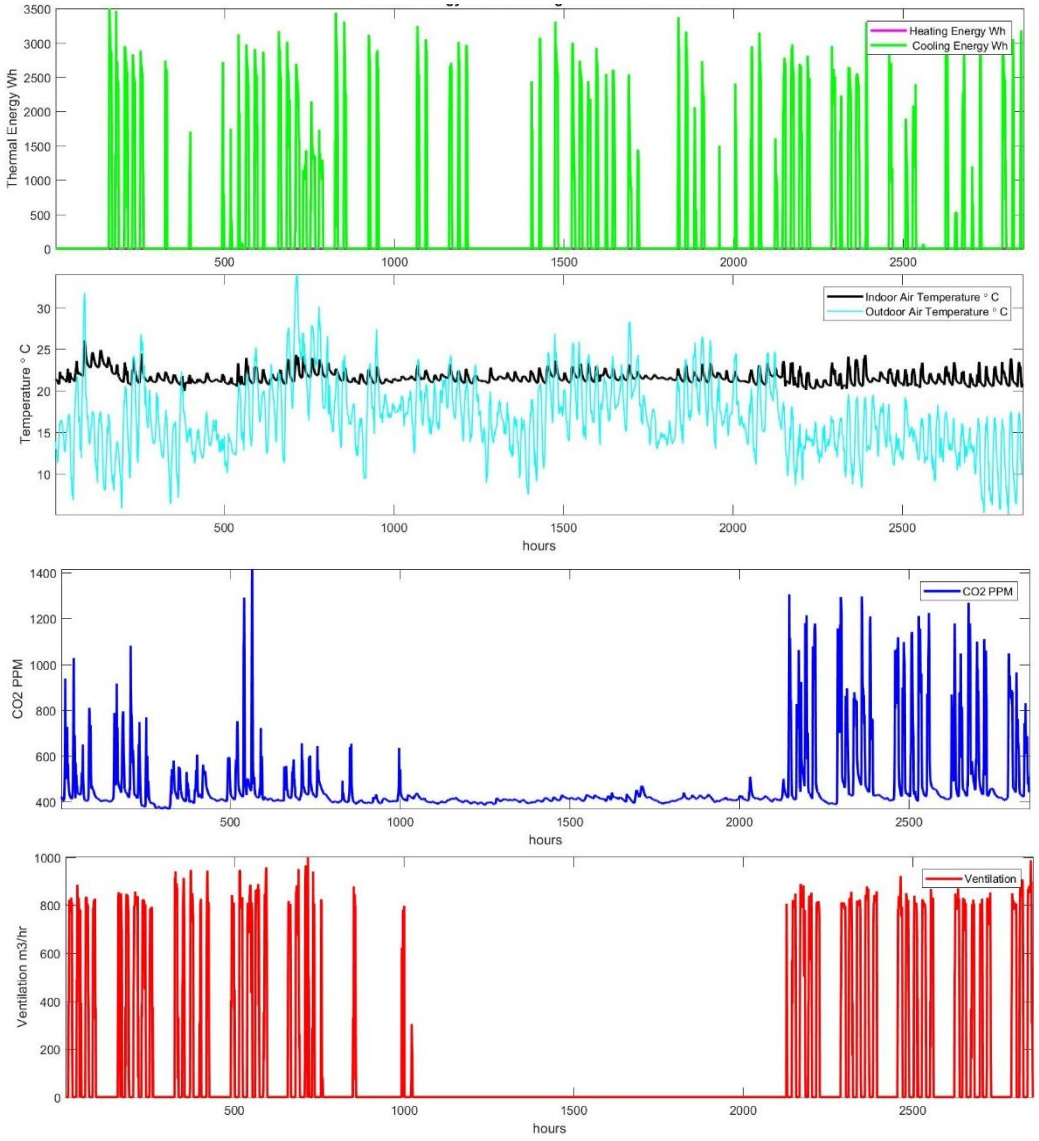


Figure 19 shows the various graphs explaining the readings of presence (CO_2 PPM,) Ventilation air flow rate, Temperatures, and the thermal energy utilization of the class room during the summer period (Note – These are consecutive hours belonging to months May-August 2015)

The graphs for the other seasons, winter and spring, have been placed in the appendix A.4.2.

An interesting point of observation is that, there is no variation in indoor climatic condition with regards to changing seasons. However, it was also mentioned that the building of the HHS is highly controlled and has an almost constant indoor set point through the year. Thus, seasonal analysis may fair well for a building

which do have different indoor set points for distinct seasons.

What follows ahead is a conclusion of the graphical analysis. The reader shall be given an overview of the findings and research questions which have been answered herein.

6.6 Conclusions

The main results of the research reported in this chapter is a toolbox and an understanding into the graphical models developed for the toolbox. It shows the important defining features of a room's operation based on energy use, presence and occupancy rates, ventilation, etc. The following major conclusions and remarks are made.

Regarding the automation of graphical analysis methods:

1. *Quick and useful observations of a building/ room's functioning can be made with an easy to use automated graphical models.* Although results are shown for one room, with the automation, one can generate such qualitative and quantitative results for most sensor-rich room.
2. *Sensor rich environments can provide for a good basis for such a data-driven model.*
3. *The graphical analysis can be used by experts, and need only data obtained from sensor database.*
4. *The need for extensive knowledge about the building itself, as in the case of white box models, is not a necessity.* However, with more information regarding the building and the HVAC systems, the higher the accuracy.
5.

With regards to the types of analysis of the control systems and efficiency in fault detections.

1. *It is clear that the ventilation work only in the case of presence being detected.* -This control strategy can be deduced by the graphs (and is in accordance with the HHS building information).
2. *Room occupancy has an important effect on the overall energy demand, since the HVAC systems are highly controlled by occupancy levels in this building.*
3. *The heating and cooling however, are operational, even in the absence of occupants, in order to maintain a quite constant indoor temperature.* This may be done for good reasons: the floor heating/cooling systems are a low temperature based system, and due to the high time constant of the floors. Thus, turning it off during non-working hours might cause issues during working hours. Also, the balance in the ATES system needs to be maintained with appropriate heating and cooling cycles, else an additional heating. Cooling energy might be needed. These observations could nonetheless point towards more thorough analysis and optimization of the control system.

4. *It is observed that some faults in heating and cooling, temperatures of air supply, and lighting can be easily detected with the model.*

Regarding the type and size of datasets.

1. *It is seen that the size of the data (yearly, monthly, weekly or even daily) is chosen based on the kind of qualitative or quantitative questions which need to be answered.* Whilst yearlong data gives a good estimate of the functioning of the HVAC systems with regards to other parameters, a zoomed in analysis of a month or week, helps understand the functioning of the room and set points such as those of presence and temperatures, etc.
2. *Using working and non-working hours, leads to analyzing the room under different operational conditions.* Public buildings are largely dependent on occupancy, and thus an analysis of such kind can yield a good quality of information about the functioning of the room.
3. *Seasonal analyses are fruitful in the case of public buildings, only if the room has set conditions which alter with every season.* It was seen that since the room 1075 of the HHS and perhaps the entire building maintains their set points almost a constant through the year – however, the seasonal analysis does give an insight into the response of indoor variables (air temperature, wall temperatures) to the outdoor parameters (solar radiation and temperatures).

Overall this chapter investigated the first most important step towards energy optimization in buildings, or building rooms – the graphical analysis. The most important aspect witnessed is that almost every parameter is responsible in some degree towards the increment or decrement of energy demand, typically thermal energy. The important question which follows is, *"whether automated methods can quantify the affect different parameters have on the overall thermal demand of a room?".* What follows is a chapter on the newly developed methodology for obtaining correlation coefficients to quantify this affect.

7.Correlation Analysis

This chapter describes the methodology developed during this research regarding coefficient of correlations. As pointed out earlier in this research, with building systems there are large dependencies of several parameters on the energy demand. The aim of this chapter is to introduce a methodology which allows for precise quantitative measurement of the effect different parameters have, on the thermal energy demand of a room. The study is also a gateway towards analyses on thermal energy prediction, as it first points to the most important factors influencing the thermal demand.

The first section is a small recap on the literature mentioned in chapter 3 (see section 3.2). It also describes the two main types of models (multiple and partial correlation) developed under this step-toolbox. The second and third section discuss results through graphical means, along with a conclusion about the developed model.

7.1 Multiple and Partial Correlations

This research uses Multiple correlation and Partial correlation as a tool to quantify the effect of certain parameters on the overall thermal energy demand of a room. Both correlations are highly generic and therefor are extremely flexible with data analytic systems [33]. This section shall show the correlation plots obtained by using such a mathematical tool. Correlation plots help in identifying the major parameters which have an effect on the dependent variable, and eliminate disturbances that have no impact at all – for example, outdoor wind speed, has almost no correlation with a room's thermal demand, if the room is completely airtight.

Table 8 List of dependent and effective parameters used in the correlation analysis.

Dependent Variable	**Effective Parameters/Disturbances**[7]
Net Thermal Demand	Indoor Air Temperature
Or *Heating Demand*	Outdoor Air Temperature
Or *Cooling Demand*	Floor surface Temperature
	Wall Temperature
	Solar Radiation
	Internal Heat
	Wind Speed
	Presence in CO^2
	Ventilation
	Supply Air Temperature

[7] These variables are called "effective variables" and not independent because there exists some amount of interdependency amongst them. Thus, although these parameters influence the thermal demand, they also influence each other, and thus are termed 'effective variables' through this research.

Table 8 show the various independent parameters (later on also called effective parameters [7]) which will be used during the correlation analysis. Almost all variables available from the sensor-rich environment have been used. However, the inlet and outlet temperatures of the floor heating/cooling pipes have not been used individually, but in the form of floor surface temperature, the calculation technique for which has been described before (see section 6.3). The floor surface temperature has been known to be an effective parameter as per the calculations and recommendations of Lopez [15]. The internal heat load has been calculated by summing up the hourly lighting, and appliance based electrical energy demand. It should be noted that the research does not take into consideration human occupancy levels as an addition to internal heat loads, as the presence (in CO^2) itself is a variable being studied. Moreover, the sensor data did not have a count of individual personnel but only total CO^2 concentrations, making it difficult to approximate the heat generated from occupancy.

A short recap on the mathematics involved in calculating the three correlations has been shown ahead;

Simple correlation – is the strength of a linear relationship of one "effective parameter" on the dependent variable. This value ranges from $-1 \leq r \leq 1$ as it is standardized. Refer to chapter 3.2 and appendix A2 for more details and formula to calculate the simple correlation r. Simple correlation is not the optimum type of correlation method in the case of more than two variables [14] [33]

Multiple correlation is used when there are 2 or more variables jointly affecting the dependent variable. Multiple correlation, denoted by 'R', also ranges between -1 and +1. Refer to appendix A.2 for more details. Multiple correlation is different from Multiple regression (see chapter 8) in the way that it is not able to predict one variable from another, or explain the changes in one variable with regards to another. It is simply a measure of the strength of linear relationship between 2 or more variables on the dependent variable [10] [11]. Multiple Correlation Coefficients are not a factor of causation but only a degree of association between the effective parameters and the dependent variable.

$$R_{z,xy} = \sqrt{\frac{r_{xz}^2 r_{yz}^2 - 2 r_{yz} r_{xz} r_{xy}}{1 - r_{xy}^2}} \qquad \text{Eq-7.1}$$

The multiple correlation coefficient is calculated using the Eq. 7.1 above. For three variables, x and y being the independent variable and z being the dependent variable, the R value is calculated, where, r_{xy}, r_{xz} etc., (see section3.2, Eq. 3.1 for simple correlations) are the simple correlations between the respective variables. As can be seen from the table 8 above, we can judge that the effective parameters are not completely independent either. For example, the solar radiation may have an influence on the outdoor temperature over

a given day. This leads to believe that there exists a certain degree of interdependency between the effective parameters.

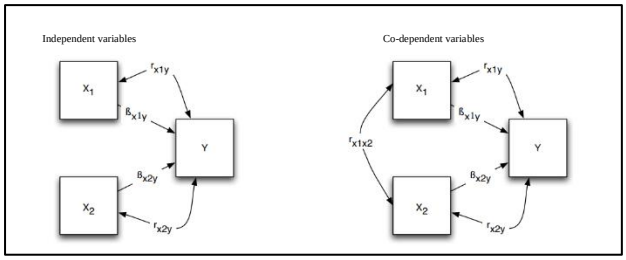


Figure 20 – β and r[8] (simple correlations) between two effective parameters X_1 and X_2 and dependent parameter Y for two given scenarios.

As can be seen from the figure 20 above, if the effective parameters are independent (completely uncorrelated to each other) then each separate variable makes a unique contribution (β^8) to explain Y (dependent variable). However as mentioned, most of the effective parameters are correlated.
Therefore, a secondary tool, called ***partial correlation*** is established (see section 3.2 and appendix A2 for more details). Using partial correlations this model can check for two important factors;

a) *The individual association between an effective parameter and the dependent variable without accounting for the remaining effective parameter* – Under this method the correlation between the dependent and the effective variable is calculated by keeping the other variables constant, thus nullifying any effect they may have.

A partially correlated value however, is not significant in reality. Although they denote the individual affect one parameter might have over the dependent variable, it does not give the true effect since in reality the parameter is interrelated with other effective parameters and does not change individually without the other parameters being altered. Thus, multiple correlation coefficients will always reflect the real relationships between the parameters, and the dependent variable. Having said this, the research looks into finding partial correlations to provide for a tool to understand the individual relationships a parameter has with the overall thermal demand of a room.

[8]In multiple correlations, the β values are standardized and are a measure of how strongly each effective parameter influences the criterion (dependent) variable. Thus, higher a β value, the greater the impact of the effective variable on the criterion variable. Example a β value of -0.91 is stronger than a β value of +0.8. These values are the standardized values of B, or weights which are calculated in the regression models in the following chapter, by using standardize effective variables. It is important to standardize the correlations coefficients because different parameters have different units (W/m^2, Wh, m/s, ^oC). With standardized values we can compare different effective parameters and their linear relationship strength with the criterion variable[10].

b) The individual association amongst effective parameters (co-dependency) – The partial correlations method is also used to calculate the values of co-dependency that may exist amongst two effective parameters. Sine the partial correlation (r) values are much lower in strength than the multiple correlation values between two variables, a significant relationship is guaranteed even with $r_{partial}=\pm 0.1$. For calculating the partial correlation between z and x, keeping y constant, the following formula is used, where r and ŕ are the respective simple and partial correlations;

$$r'_{zx,y\,(partial)} = \frac{r_{zx} - r_{xy}r_{yz}}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xy}^2}} \qquad \text{Eq-7.2}$$

Using this formula (Eq. 7.2), the partial correlation between all effective parameters was calculated. The simple correlation between the effective parameters was also estimated. Significant values, that is with p-value<0.5 were placed in the table 9 shown below.

Table 9 Combinations of interdependent parameters, which were singled out based on their r-value above or below ±0.1 and P-value<0.05 (95% confidence). The table also shows the simple correlation coefficients for the same set of parameters for comparative purposes.

Interdependency of Parameters	Partial Correlation	Simple Correlation
Internal heat and presence	0.37	0.69
Wall Temperature & Indoor Air Temperature	0.93	0.96
Indoor Air Temperature & Supply Air Temperature	0.37	-0.07
Floor Temperature & Supply Air Temperature	0.15	0.24
Supply Air Temperature & Presence	-0.11	-0.42
Presence & Outdoor Air Temperature	-0.18	0.00
Wind Speed & Outdoor Air Temperature	0.20	0.07
Outdoor Temeprature & Solar Radiation	0.20	0.49

To explain partial correlations better, let us consider three variables namely, 'Presence', 'Supply air temperature' and 'Internal Heating'. A partial correlation between Presence and Supply air temperature (variable X_1 and X_2) would be performed after the third variable (Internal Heating) (X_3) is "switched off". Switched off means, that the relationship between one variable (X_1) and the switched off variable (X_3) is calculated and the residuals from such a regression is used to find the partial correlation between X_1 and X_2. That is, to find the partial correlation between presence and supply air temperature, first the correlation of presence with internal heating would be performed. The presence can answer for this correlation but not perfectly, leaving behind a residual. The partial correlation is the relationship supply air temperature has

with the residuals of presence. In this case the simple correlation between presence and supply air temperature is -0.42 and this is without switching off variables such as internal heating (lighting and appliances) and other variables as mentioned. However, the correlation between supply air temperature and presence only, keeping aside the effect of other parameters is much lower but still a negative -0.11. Sometime the reverse is also possible, as seen with wind speed and outdoor air temperature, wherein, the values are lower (almost 0) for a simple correlation, however, individually, these two are correlated weakly at 0.2. Thus, for any two variables there will always be some form of relationship, be it weak (closer to 0) or strong (closer to +1 or -1) but never 0 [33].

The significance of calculating partial correlations is that by analyzing these values we can develop a second opinion regarding a given effective parameter. If this parameter has a significant partial correlation coefficient as compared to its multiple correlation coefficient, it would suggest that the parameter,

1. *has a good linear relationship with the dependent parameter (as all correlation are based only on linearity in data), since it shows a significant relationship under partial and therefore also on multiple correlations.*
2. *is an important factor in analyzing the dependent variable response in terms of a predictive model* [33].

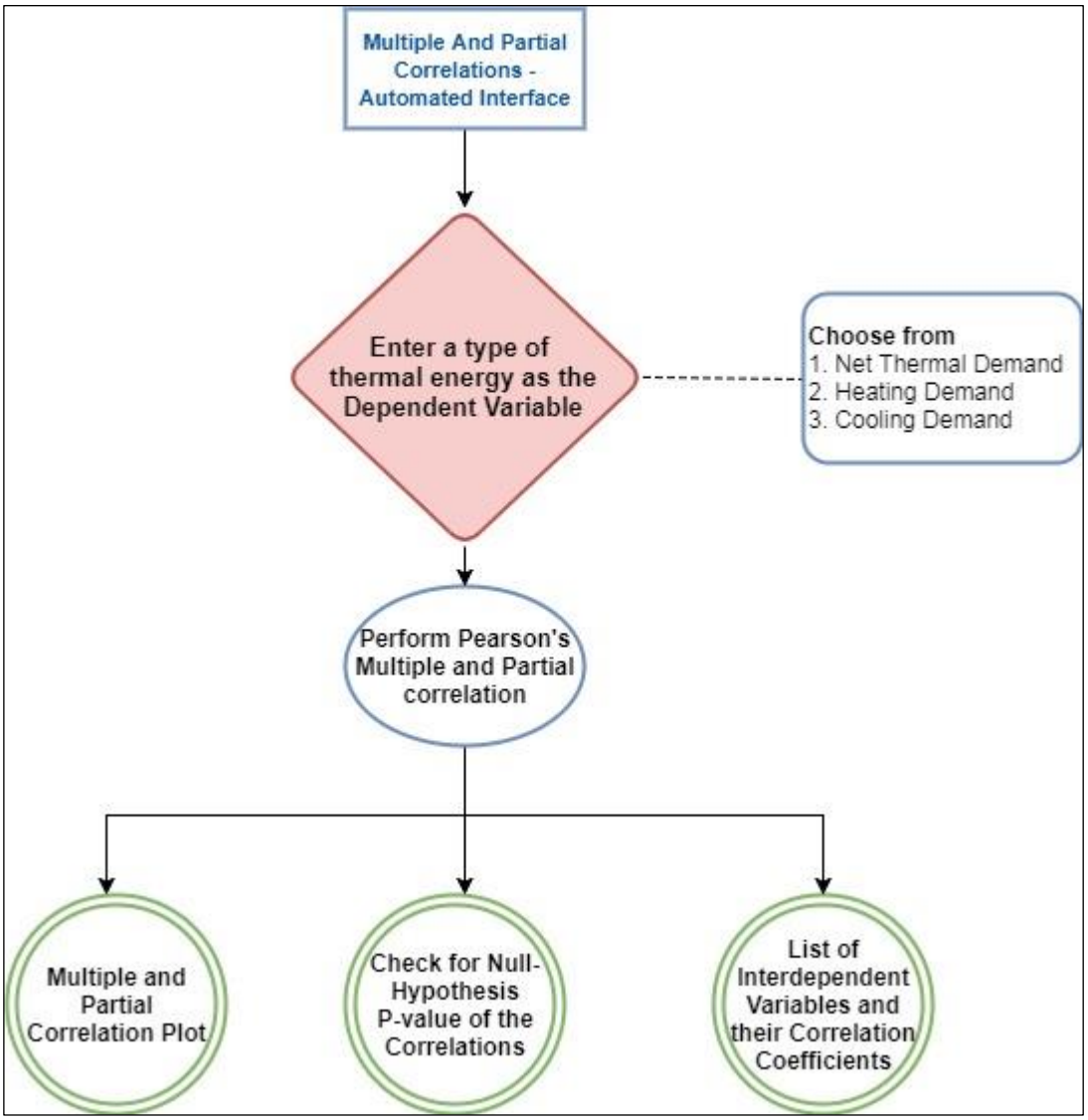


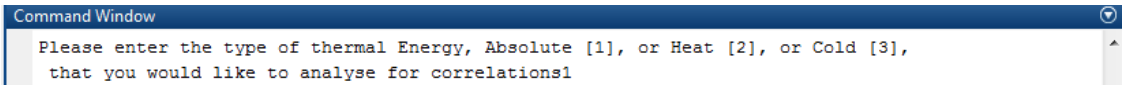
Figure 21 Methodology developed for the correlation coefficient estimation during this research.

An automated methodology (figure 21) has been developed during this research, to help find the multiple and partial correlations of any kind of thermal demand, with respect to the effective parameters recorded from the room. Once again, the interface for this step-toolbox is such that the types and number of parameters can be changed and chosen based on the objectives of the user.

What follows is a description of the correlation toolbox and the interface developed for automating the process. Thereafter the correlation plots and the results for the same are described for the room 1075 as shown in the flowchart above.

7.2 Automated Correlation Step-toolbox

Using MATLAB an automated model has been developed. As pointed out during the literature survey, it was found that there is a lack of understanding of the affect, different parameters have on the thermal demand of a building, in this case a room. Thus, a generic algorithm is used to find these correlations, of both multiple and partial nature.



Herein the command menu asks the user to input the type of thermal energy which should be set as the dependent variable. Thereafter, the Pearson's correlation is calculated, leading to the following type of result.

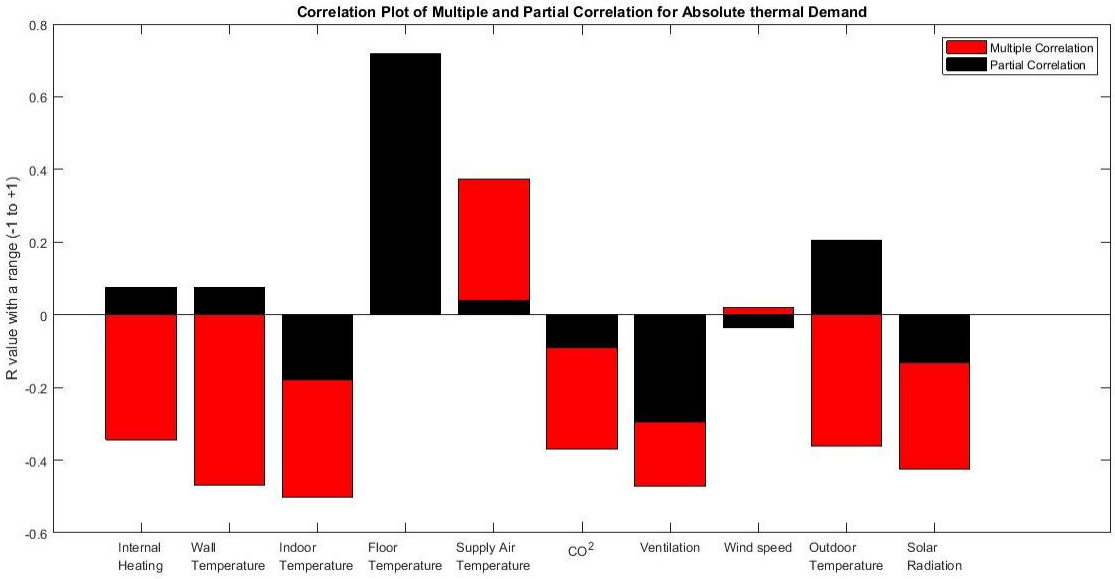


Figure 22 Multiple and Partial correlations of net thermal demand with the effective variables for a period of 1 year. Naturally the most correlated is the floor temperature as it is the main source of thermal energy

The figure 22 above is a graphical plot of the multiple and partial correlation coefficients, the exact values of which have been placed in the table 10 below. The bars represent the strength of the relationship between the effective variables and the dependent variable chosen (in this case the net thermal energy demand). Please note that this image does not incorporate any values of interdependency that may exist between the effective parameters.

Table 10 shows the obtained Multiple and Partial correlation coefficient (R-Value) and the P-values for the various parameters as stated.

Parameters →	Internal Heat	Wall Temperature	Indoor Air Temperature	Floor Surface Temperature	Supply Air Temperature	Presence	Ventilation	Wind Speed	Outdoor Air Temperature	Solar Radiation
Multiple Correlation Coefficient (R)	-0.34	-0.469	-0.503	0.719	0.374	-0.37	-0.47	.019	-0.362	-0.425
P-Value	0.0	0.0	0.0	0.0	0.0	0 .0	0.0	0.090	0.0	0 .0
Partial Correlation Coefficients ($r_{partial}$)	.074	0.08	-0.176	0.720	0.04	-0.097	-0.30	-0.03	0.205	-0.132
P-Value	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0

Parameters →	Internal Heat	Wall Temperature	Indoor Air Temperature	Floor Surface Temperature	Supply Air Temperature	Presence	Ventilation	Wind Speed	Outdoor Air Temperature	Solar Radiation
Partial Correlation R^2 Values	1%	1%	3%	52%	0.00	1%	0	0	4%	2%

The R^2 values describe the percentage of variance in the values of the dependent variable explained by the independent variable(s) (see chapters 8 and 9 for detailed explanation on R^2 values). The R^2 value is simply calculated by squaring the R values estimated by a simple or a partial correlation – these have been shown above in table 10. However, the R^2 for a multiple correlation needs a linear model to calculate the overall R^2 value. The total R^2 value of the Multiple Correlation Coefficients estimated was 74.2%.

The overall R^2 value is obtained by placing the effective parameter/variables into a linear model (FITLM or Stepwise fit, see chapter 8 for detailed information). The model trains itself on this input matrix, and develops an equation to define the value of the dependent variable (net thermal energy demand -NTD). Thus, based on the values obtained above, the equation for each variable with the partial coefficient would be as follows (Eqs. 7.3-7.5);

$$NTD_i = C_{internalheat+}0.074 * InternalHeat_i \rightarrow R^2 = 1\% \quad \text{Eq-7.3}$$

$$NTD_i = C_{walltempearture+}0.08 * WallTemp_i \rightarrow R^2 = 1\% \quad \text{Eq-7.4}$$

$$NTD_i = C_{floortempearture+}0.72 * FloorTemp_i \rightarrow R^2 = 54\% \quad \text{Eq-7.5}$$

And so on, where *i* represents a given hour. It is seen that only the floor temperature has a high enough partial correlation with the net thermal energy demand. However, other variables are not able to partially answer for their relationship with the thermal demand, and have a very low R^2 value. This is once again

because, in reality these variables are not partially related to thermal demand, but are interrelated as seen from the high R^2 value of multiple regression equation (Eq. 7.6) obtained from FITLM.

$$NTD_i = C - 0.34InternalHeat_i - 0.047walltemp_i - 0.50indoortemp_i + 0.72floortemp_i - 0.37Prsence_i - 0.47ventiatlion_i - 0.36outdoortemp_i - 0.425solarradiation_i \rightarrow R^2 = \sim 75\% \quad \text{Eq-7.6}$$

Thus, the first step towards obtaining multiple regression equations is to understand the strength and weaknesses of each of the predictor or effect parameters being used. This shall be explained further in the chapter on Multiple Linear Regression – chapter 8.

7.3 Analysis of the results for the case study

In this discussion, we shall first analyze the results of the multiple correlation followed by those of the partial correlation coefficients.

The dependent variable chosen for this case study is the net thermal demand of the room, and for a period of 1 year. Please note that the net thermal demand can be a positive or negative value, as shown in figure 12 in chapter 6. This positive (heating energy demand) or negative (cooling energy demand) could lead to certain explanation of the analyses provided ahead. The variables/parameters affecting this thermal demand are listed in table 8 above.

1. ***With regards to temperatures*** – the net thermal demand of a room is inversely proportional to both *indoor and outdoor temperatures*. This seems to make sense, as the thermal demand rises or falls with decreasing and increasing temperatures respectively. Contradictory to this, with a high *floor temperature*, there shall be a higher amount of heating, which implies a higher thermal demand. A low floor temperature would mean a higher demand in cooling and therefore a more negative thermal demand.

2. ***With regards to outdoor wind speed*** - A very low correlation of thermal demand with wind speed suggests that the building perhaps and in this case this specific classroom is immune to large infiltrations. By looking at the high P-value for wind speed in the table xx below, we can conclude that the correlation coefficient did not pass the null hypothesis test, and thus this variable has no effective correlation with the thermal energy demand of the room.

3. ***Regarding occupancy, internal heating and ventilation*** – The ventilation is coupled highly to the occupancy, at almost 0.69. Internal heat is also coupled to the occupancy at 0.37 (see table 9). With regards to a multiple correlation we see an inverse relationship between these three variables and the net thermal demand. This is natural as increase in any of these variables leads to an increase in the internal room temperature, leading to a negative demand (cooling demand) in thermal energy.

Other graphs, based on only heating or only cooling hours have been placed in appendix A5. These graphs are accompanied by their tabulated values and the set of interdependent effective variables in their case.

7.4 Conclusion

The use of such a model is essential in generating quick analysis regarding the most effective parameters on the thermal demand of a room. The chapter shows the use of multiple and partial correlations on building energy demands. It should be kept in mind that these coefficients are just a measure of effect, or degree of effect and not the strength of causality.

The chapter also introduces the strengths and weaknesses of these two correlation methods with regards to building energy, occupancy and outdoor climatic parameters.

1. ***Regarding Effectiveness of multiple correlations*** - Multiple correlations although includes interdependent parameters, it is still able to supply significant results, with a certain degree of error. These results hold true as in reality, variables are interrelated and without multiple correlations, the overall effect of a variable would not be understood.
2. ***Regarding effectiveness of partial correlations –*** The use of partial correlations helps tighten the number of significant parameters effecting the thermal demand. However, one major disadvantage in using partial correlations with so many parameters is that the reliability of the results begins to diminish as can be seen above and was also seen in the book of J. Cohen [10]. So therefore, under partial correlations, certain parameters cannot be "switched off", as the results then tend to become insignificant. Multiple correlations thus need to be used to get a quantitative measure of the relationship between effect and dependent parameters.

The next chapter shall finally move into the predictive models developed during this research based on the findings of chapters 6 and 7. These chapters were important to analyze the building, which shall now be followed with energy predictions.

8.Multivariate Regression & Predictive Modelling

This chapter describes the developed predictive models based on Multivariate Linear Regression. From the findings of the previous chapters it has become clear that the thermal energy of a room depends on a lot of factors, and the strength of each of these factors varies non-linearly. Using this knowledge, the types of data and variables for the predictive model will be selected. This model shall focus on prediction of heating and cooling demands of the case room 1075 and developing an automated model that can be applied to other rooms and buildings as well.

In order to provide a wholesome explanation of this 3rd step-toolbox, the section shall be divided in the following manner. The first subsection shall explain the thermal energy balance prevalent in buildings/rooms (related to Heat Transfer Fundamentals section 3.1) and the corresponding set of parameters involved in defining Multivariate Liner Regression (MLR) equations. The second section includes a detailed brief on model development, continuing the literature survey in chapter 3 (see section 3.3). Once the predicted model is setup with the appropriate equations, it is followed by a description of the statistical methods used for validation of the predicted values and coefficients. An explanation of the significance of the coefficients will be provided. Using the results of both FITLM [41], and STEPWISEFIT [42], two MATLAB functions - multivariate linear modelling, shall be showcased, along with a descriptive and comparative analysis.

8.1 The Principle of Thermal Energy Balance

This section shall expand upon the literature of heat transfer in the built environment, and explain the most important mechanisms through which thermal energy balance is affected. This is vital, to understand the set of equations which should be associated with the predictive modelling through multiple linear regressions in the next subsection.

The thermal maintenance of a room is a chief factor for a comfortable indoor microclimate. However as pointed out in the literature section there are several fluxes of heat being transferred in and out of the room which influences the balance of thermal energy. The five main categories of heat flux are: internal heat gains ($Q_{internal}$), envelop losses/gains ($Q_{floor}+Q_{envelope(s)}$), ventilation ($Q_{ventilation}$), solar ($Q_{solar}$), and infiltrations($Q_{infiltratinon}$). There is a sixth term which is responsible for both gains and losses of thermal energy, and is associated with the thermal mass of a building/room ($Q_{thermal\ mass}$). This thermal mass absorbs the solar radiation, and internal heat gains. The thermal energy balance can thus be written as a simple linear equation shown below;

$$(Q_{thermal\ demand})\ [W] = (Q_{internal}) + (Q_{floor}+Q_{envelope(s)}) + (Q_{ventilation}) + (Q_{solar}) + (Q_{infiltratinon}) + (Q_{thermal\ mass})$$

Eq-8.1

Here the $Q_{thermal\ demand}$ is the overall thermal demand of a room (the net thermal demand). *The* $Q_{internal}$ is the thermal energy introduced to the room via occupants, appliances and lighting. Q_{floor}, Q_{ground} and $Q_{envelope}$ are heat transmissions through the envelope of the building, and are driven by the difference in temperatures between the indoor and the outdoor. The $Q_{ventilation}$ and $Q_{infiltration}$ are the thermal energy transmittance through air ventilation and infiltration respectively. Q_{solar} is the solar heat gains of a room through solar radiation. Lopez [15] summarizes a table of the physical description and the parameters associated with each of the heat fluxes. A modified summary corresponding to this research has been shown in the table 11 below.

Table 11 Physical description of distinct heat fluxes affecting the thermal demand of a room together with the parameters associated with each physical attribute.

Heat Flux	Physical Description	Parameters Associated
$Q_{internal}$	$Q_{internal} = n_{people} . Q_{body} + Q_{lighting} + Q_{appliances}$	n_{people} is the number of people Q_{body}, $Q_{lighting}$ and $Q_{appliances}$ corresponds to the heat gain per person, total heat gain from lighting and appliances, respectively [W]
Q_{floor}	$Q_{floor} = U_{floor} . A_{floor} . (T_{floor} - T_i)$	U_{floor} is the convective and radiative heat transfer coefficient of the floor at the indoor side [W/m^{2}K] A_{floor} Area of the floor [m^2] T_{floor} hourly floor temperature [K] T_i indoor air temperature [K]
$Q_{envelope(s)}$	$Q_{envelope} = \sum_i U^i . A^i . (T_o - T_i)$	i is the number of façades of the room including roof, excluding floor U_i the overall heat transfer coefficient for this envelope [W/m^{2}K] A_i the area of the facades [m^2] T_o the outdoor temperature [K]
$Q_{ventilation}$	$Q_{ventilation} = M_{vent} . C_{p\ air} .(T_{outAHU} - T_i)$	$C_{p\ air}$ heating capacity of air (J/kg.K) M_{vent} mass flow rate of the ventilation air (kg/s). T_{outAHU} Temperature of the ventilation air coming into the room [K]

Q_{solar}	$Q_{solar} = Q_{sol\ direct} + Q_{sol\ diff.} + Q_{reflective}$	The diffused and direct solar radiation are calculated from the obtained global horizontal solar radiation obtained from the KNIM source[9]. $Q_{reflective}$ is neglected.
$Q_{infiltrations}$	$Q_{infiltrations} = (m_{openings} + m_{cracks}) . C_{p\ air} (T_o - T_i)$	$m_{cracks} = V_{room} . 0.15. \left(\frac{V_{wind}^2}{V_{reference}^2}\right)^{2/3}$ [Kg/s][*Lopez 15, ,NEN-EN 12207]* $m_{openings} = V_{openings} . \rho_{air}$ [Kg/s] $V_{reference} = 5$ [m/s] V_{room} = volume of the room [m^3] ρ_{air} = *density of air 1.225* [Kg/m^3]
$Q_{thermal\ mass}$	$Q_{thermal\ mass} = \alpha_i . A_{indoor\ surfaces} . (T_w - T_i)$	$A_{indoor\ surfaces}$ = total area of all the facades in contact with indoor air [m^2] α_i = the indoor combined heat transfer coefficient for convection and radiation T_w = wall temperature [K]

It is important to note that not all these heat fluxes are applicable in this research. From the previous chapter we found for example, that the wind speed has almost no effect on the thermal demand of this case room 1075. The P-value for the extremely low correlation coefficient was high indicating that the coefficient did not pass the null-hypothesis. This means that $Q_{infiltrations}$ will have no effect on the thermal energy balance for this specific room.

From such a table, it is evident that to utilize the thermal energy balance equation 8.1 as stated above, we need to estimate or know a lot of the building's physical parameters such as dimensions and building properties such as U values. This becomes a bigger issue with already existing buildings where such information is difficult to retrieve. However, making a regression analysis using independent variables T_w, T_i, T_o, T_{floor}, T_{outAHU}, Q_{solar} **and** $Q_{internal}$ should allow to develop a predictive model with physical meaning. The equation used here is the same as developed by Lopez[15] , who already demonstrated the potential of this method. However, she had no actual data at her disposal and made the demonstration based on data produced by an emulator (a white box model), which produced a surrogate for the actual data. Furthermore, in her study, there was no floor heating or cooling, and the floor temperature was approximated by using indoor temperature, internal heat load and solar radiations at the hours before.

[9] The use of solar radiation at t-1 hours (as seen in the graphical analysis chapter 6 section 6.3.3) shall also be taken into consideration at a later stage in this chapter (see section xx).

In the present research, we can test the approach on actual data and, we can validate that the black box model of Multivariate Linear Regression has a good fit with the thermal demand of the rooms.
Ahead a detailed description of the MLR has been given followed by the major first results of the predictive model.

8.2 Multivariate Linear Regression – An overview

The literature on MLR has already shown the approach of using a mathematical black box model. This model consists of a linear equation with one dependent variable (prediction term) and other independent variables (predictor terms). The general expression which defines such a linear relationship is as shown;

$$Q_h = constant + \sum_{i=1}^{n} C_i \,.\, X_i \qquad \text{Eq-8.2}$$

Wherein, Q_h is the hourly thermal energy demand in Watt (W), C_i is the coefficient estimated for the i^{th} parameter X_i. The equation can also be rewritten in the matrix form as [11]

$$Q = constant + X \times C$$

Where

1. Q is the matrix of the prediction term (net thermal demand) which in this case is hourly (n hours) making it n by 1column matrix
2. C is the matrix of coefficients of the predictor variables (m variables) (these differ for differing combination of predictor variables), a m by 1 matrix
3. X is the multi columned matrix (n by m) which denotes one column for each predictor variable used for this predictive model. Example, indoor temperature, outdoor temperature, solar radiation, etc.

The n is the number of hours, and m the number of predictor variables. Thus, for one entire year of 8760 hours, the matrix would be represented as shown;

$$\begin{bmatrix} Q_1 \\ \vdots \\ Q_{8760} \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{8760,1} & \cdots & X_{8760,m} \end{bmatrix} \begin{bmatrix} C_0 \\ \vdots \\ C_m \end{bmatrix} \qquad \text{Eq-8.3}$$

Where,

C_0 is the constant obtained of the entire MLR equation.

The matrix set of equations are solved by estimating the constants through training and fitting the model with historical data, explained ahead.

8.2.1 Regression and Fitting

Using equation 8.4 shown below, the model can calibrate the constants of the equation to match the actual values of a historical dataset of the dependent variable, in this case the thermal demand.

$$\begin{cases} Q_1 = C_0 + C_1.X_{1,1} + \cdots + C_m.X_{1,m} \\ Q_{8760} = C_0 + C_1.X_{8760,1} + \cdots + C_m.X_{8760,m} \end{cases} \quad \text{Eq-8.4}$$

The coefficient values obtained represent the numeric contribution of each parameter (X) on the overall thermal demand (Q) [43] [11]. These when normalized (-1 to +1) are similar to the correlation coefficients estimated before in chapter 7. This is a type of inverse modelling; wherein statistically valid coefficients are estimated to fit each parameter into the linear equation 8.2. Once the fit is made (or training of the model is complete) the Left-Hand Side of the equation can be estimated for future values of X.

8.2.2 Model development

A description of the types of data retrieved from the Octalix and Priva database has been provided before in chapter 5. A secondary table of the sample of available data has been placed in the appendix (A6) as well. This data set was measured data from sensors and lacked any informative data about the building, such as the building physical parameters, or the insulation U values, etc. In this research, the actual data of temperatures, electrical consumption and solar radiation, etc have been used to develop the major linear equation solved by MATLAB.

Based on the knowledge of available data, we can relate the equation of thermal energy balance (eq. no. 8.1), and the regression equation number (8.4) as follows;

$$Q_{thermal\ demand}[W] = C_0 + C_1.(T_{outdoor} - T_{indoor}) + C_2.(T_{floor} - T_{indoor}) + C_3.(T_{wall} - T_{indoor}) + C_4.(T_{out\ AHU} - T_{indoor}) + C_5.V_{wind}.(T_{outdoor} - T_{indoor}) + C_6.Q_{solar} + C_7.Q_{internal} \quad 8.5$$

This equation is a transformation of the general equation (8.2) of multivariate linear regression with eight constants (coefficients) and seven parameters written in the form of temperature differences [^oC], solar radiation [W/m^2], internal heat gains [W] and wind speed [m/s]. The overall result obtained is a total thermal power in the hour in Watts, which translates to the thermal energy demand of the hour in Wh.

The values of the coefficients relate to an individual category of heat flux as mentioned in section 8.1 of heat transfer above. The values C_1, C_2 and C_3. relate to the room's physical parameters, and are constant for a given room. C_4 is dependent of the ventilation time frame, which in turn is dependent on the occupancy profiles. C_5 is dependent on the infiltration rate in the room. C_6 and C_7 are the coefficients for

solar radiation and internal heat gain respectively. The physical significance of each of these coefficients have been summarized in the table below.

Table 12 Physical significance of the coefficients estimated by using MLR models, with regards to the equation presented before (equation number 8.4).

Coefficients	Physical Significance	Associated Variables
C_1	$C_1 \sim \sum_i U^i . A^i$	i is the number of façades of the room including roof, excluding floor U_i the overall heat transfer coefficient for this envelope [W/m^{2}K] A_i the area of the facades [m^2]
C_2	$C_2 \sim U_{floor} . A_{floor}$	U_{floor} is the convective heat transfer coefficient of the floor [W/m^{2}K] A_{floor} Area of the floor. [m^2]
C_3	$C_3 \sim \alpha_i . A_{indoor\ surfaces}$	$A_{\text{indoor surfaces}}$ = total area of all the facades in contact with indoor air [m^2] α_i = the indoor combined heat transfer coefficient for convection and radiation. [W/m^{2}K]
C_4	$C_4 \sim M_{vent} . C_{p\ air}$	$C_{p\ air}$ heating capacity of air [J/kg.K] M_{vent} mass flow rate of the ventilation air [kg/s].
C_5	$C_5 \sim (m_{openings} + m_{cracks}) . C_{p\ air}$	$m_{cracks} = V_{room} . 0.15 . \left(\frac{V^2_{wind}}{V^2_{reference}}\right)^{2/3}$ [Kg/s] [Lopez 17, NEN-EN-12207] $m_{openings} = V_{openings} . \rho_{air}$ [Kg/s] $V_{reference} = 5$[m/s] V_{room} = volume of the room [m^3] ρ_{air} = *density of air 1.225* [Kg/m^3]
C_6	$C_6 \sim$ effect of solar radiation on the thermal demand	_
C_7	$C_7 \sim$ effect of total internal heat gains on the thermal demand.	_

The predictive accuracy of a model or set of equations depends upon the following two main factors:

1. *Type and number of parameters introduced as independent variables.*
2. *The statistical significance of the obtained coefficients of each parameter*

The significance of the coefficients estimated above are validated from the p-value and the t-statistics tests (explained ahead in section 8.3 in detail) performed by the models. Two main multiple regression functions were studied during this research namely, FITLM and STEPWISEFIT on the MATLAB 2017b [30]

8.2.1.1 FITLM

FITLM [41] is a simple function in MATLAB used in order to develop and estimate values for a simple multivariate linear regression problem. The research focuses on comparing FITLM to the second more complex function namely, STEPWISE FIT [44]. FITLM works on the principle of least-square fitting and uses the backward principle of estimating statistically valid coefficients and significant parameters for a predictive model. Herein, all the parameters are first placed in the MLR equation, and based on trial and error method, the less significant parameters are removed from the regression equation [11]. During research, it was noted that the major disadvantage of this function is that all validation and insignificant parameter removal must be done manually. This is not optimum as the research calls for more automation.

8.2.1.2 STEPWISE FIT

Stepwise fit is a much more automated regression function in MATLAB. It is used to fit parameters onto a model, and does so with the forward principle. It uses an interface (see appendix A7.1) wherein, the parameters (predictor variables) are added one by one to the regression equation, chosen based on optimality [11]. The effect on the overall models R^2 value (goodness of fit) and the significance of each of the parameters can be judged instantly. This method allows for automated removal of parameters having no significant effect on the predictive accuracy.

This research has made use of both the functions for predictive analysis of thermal demand at the room level. A short comparative study of these two methods regarding their use and accuracy has been missing from literature and thus, falls under an objective of this research. The detailed explanation of these two models and their user interface can be seen in the appendix (A7.3).

The next subsection deals with the statistical validation of the results necessary to deem the model practical.

8.3 Statistical Validation

With the use of inverse modelling using Multivariate Linear Regression the values obtained must be statistically sound. The Stepwise and FITLM functions use statistical significance in regression models to obtain the best fit for the dependent prediction variable. The statistical significance of the model can be checked using three main methods -

1. ***Analyzing the residuals of the data*** – The residuals of the data can be verified statistically by analyzing their mean, variance and distribution profiles. The mean of the residual data set should be nearly zero. The variance should be approximately constant for all values of X and there should be a normal distribution associated with the values of X [11].
2. ***Significance levels of the estimated coefficients*** – the individually obtained coefficients of the regression model must be statistically significant for the entire prediction to be significant. This is checked using the p-values and the t-statistics tests.
    a. The *P-value* as mentioned before, is to test the null hypothesis of the significance of the estimated values. This research uses a minimum p-value of 0.05. This means that a p-value >0.05 would result in the fact that the values of the particular coefficient estimated is not statistically significant, and thus might induce an error to the overall model prediction [11]. The p-value and t-stats value change each time a new parameter is introduced into the equation. Variables with a p-value above 0.05 shall be removed from the equation, as there is less than 5% chance of the values being significant.
    b. The t-statistics is a ratio measuring the difference in estimated and actual values to the standard error of the estimated value [11]. Thus, a lower value of t-statistics means that the variable has a lower error and a better contribution to the fit. The stepwise fit uses the t-stats value to add new variables into the regression equation.
3. ***Significance of the entire model*** – The entire model as a whole needs to be statistically sound to be accepted as a valid prediction model. A large number of parameters can be added as predictor variables; however, a large number of parameters could lead to overfitting, which showcases random errors rather than addressing the relationship between dependent and independent variables [11] [45]. This would lead to poor-predictive performance of the entire model. By using the goodness of fit (R^2) value and the RMSE the significance of the model is validated.
    a. $\boldsymbol{R^2}$ measures how close the data is to the fitted regression line. It is a measure of the total variation in data that can be explained by the model. It varies from 0 to 1, with 1 being a perfect fit. It is 1 minus the ratio of the sum of square of prediction errors and the sum of squared deviations from the mean [45]. The Adjusted-R^2 uses the variances instead of the variations, that is, it takes the sample size and the number of predictor variables into

considerations. The adj. R^2 value should be used in case of comparing different sub-datasets[10]. The goodness of fit increases by increasing the number of significant parameters, but leads to a reduced value with adding random predictor input variables (see section xx). Also, there is a tight relationship of the R^2 value with the RMSE.

b. *The RMSE* is the square root of the mean variance of the residuals. It needs to be minimized. With increasing parameters there is a high chance of overfitting, leading to an increase in the RMSE [11]. The RMSE hold the unit of the predicted and actual values (in this case Watt) and its value is proportional to the values of the actual data. The adjusted R^2 is thus a modified version of R-squared adjusted for the number of parameters in the equation. The adjusted R^2 increases when a model is actually improved by a new parameter, rather than increasing due to probability [11].

8.4 Data set selection

Chapter 5 has shown the types of data retrieved from the sensors. Data regarding the room 1075 is available at an hourly timestamp (see appendix A.6). Based on the equations formed in the sections above, and the knowledge of correlation coefficients generated regarding the most influential parameters the appropriate predictor variables are selected. The important *parameters* of data selected for the models developed are;

1. *Temperatures of indoor air, walls/envelope and floor*
2. *Outdoor Climatic data (solar radiation, wind speed [10]and outdoor air temperature)*
3. *Use and operation (Time of use, ventilation profiles etc.)*

The dataset for room 1075 is a total of 8327 hours our of 8730 hours of a year, due to some missing data in the Octalix and Priva databases. The sub-datasets sets were varied with each obtained result, which has been discussed ahead.

The choice of the size and time period of data selected to form the sub-data set is extremely crucial towards a good training and fit, and this shall be explained ahead with the results from this research section.

[10] Although it was proven for this dataset, that the windspeed is insignificant in forming a linear relationship with net-thermal demand, it was used nonetheless to show the functioning of FITLM and STEPWISE FIT with regards to removal of insignificant parameters.

What follows is a flow chart (figure 23) summarizing the entire process of the automated MLR step-toolbox, developed during this research.

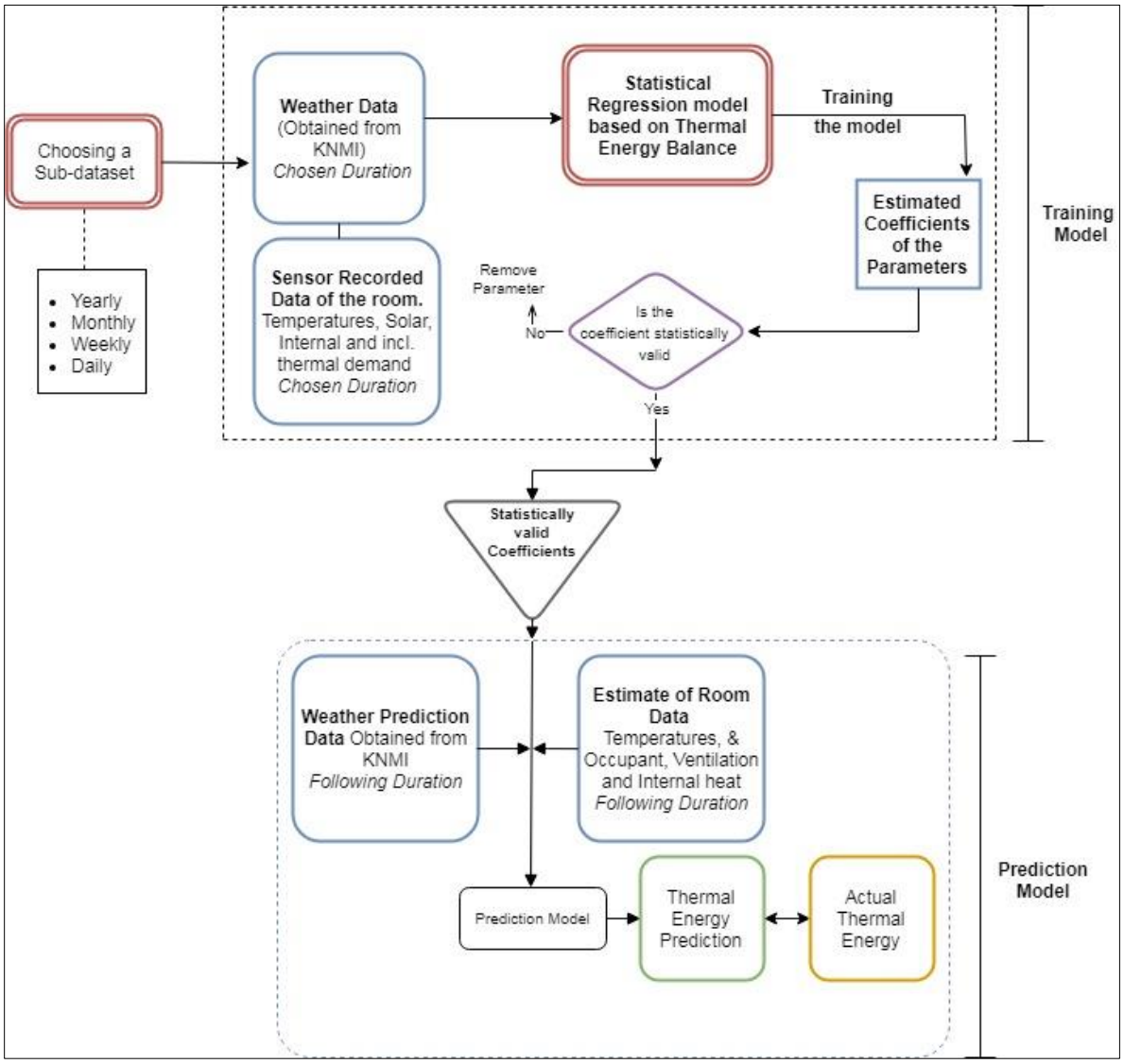


Figure 23 Descriptive flow chart of the steps followed in the entire methodology of the MLR step box, to obtain high comparative prediction model.

8.5 Results and Discussions

This section describes the results obtained by using two functions, FITLM and STEPWISEFIT in MATLAB 2017b and their corresponding significance with relation to real-time data. The section is divided into two major parts showcasing the results of fitting and training models on different sub-data sets first, followed by a subsection on prediction of data. Thereafter, the conclusions of this chapter are accompanied by a brief comparative study.

8.5.1 Fitting and Training the Models

Using the FITLM and STEPWISEFIT the fitting profiled over an entire year's data was analyzed. The adjusted R^2 value of such large dataset was 64.7% and a RMSE of 399Wh for the FITLM while for the STEPWISEFIT, a similar adjusted R^2 value of 64.9% with RMSE being 397Wh was obtained. This means that the regression models can explain 65% of the total variance in the data. The values for the correlation coefficients and the adjusted R^2 value has been placed in the table 13 below. A graphical representation of the fitted data for the entire year can be seen in the figure 24.

Table 13 Estimated coefficient values for the fit over an entire year[11].

	FITLM		STEPWISE FIT	
Constant	**Estimate**	**P-value**	**Estimate**	**P-value**
Intercept (C_0)	16.8	0.01	16.86	0
C_1	294.52	0	290.96	0
C_2	334.08	0	335.68	0
C_3	10.54	0	10.85	0
C_4	0.008	0.17	3.89	0.517
C_5	0.004	0.97	0	0.1
C_6	-0.20	0	-0.20	0
C_7	0.13	0	0.10	0
Adjusted R-Squared	**64.7%**	-	**64.9%**	-
RMSE	**399**	-	**397**	-

[11] Please note that there is not Prediction, but a mere fit of data with the yearly time period.

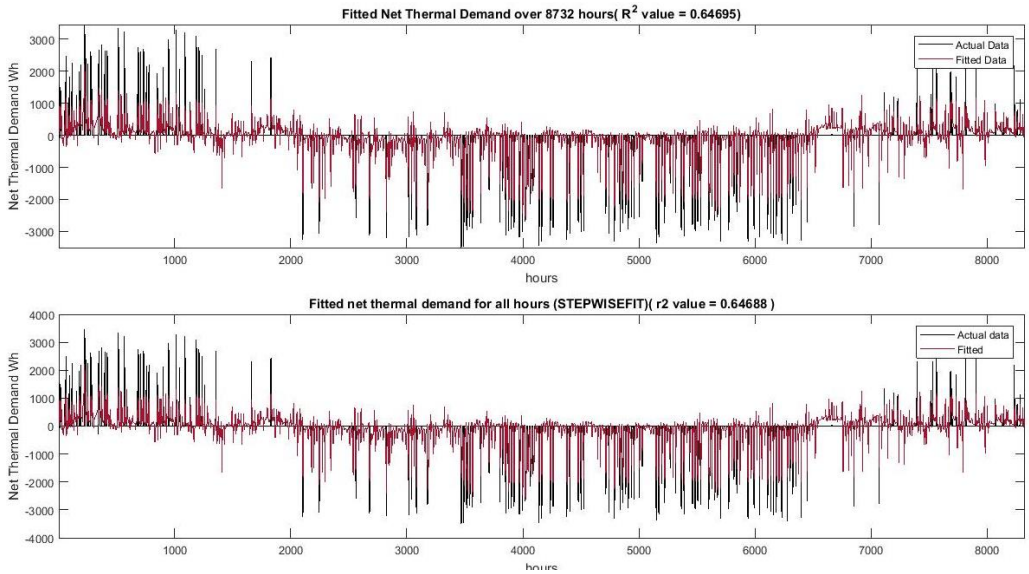


Figure 24 Graphical representation of the fitted net thermal demand over the entire year of 8372 hours data size. The R^2 value and RMSE of this fitting were similar for both functions.

Observations

1. *The first most important observations are the p-values, which in this case showcase a poor test result for the coefficients related to ventilation (C_4) and infiltration (C_5).* Since these are statistically insignificant, they are removed from the equation manually for the FITLM, whilst automatically from stepwise fit.
2. *The walls and the floor surface temperatures are playing an important role in determining the fit of the data for both functions.* When the training was performed with just these two variables the model was able to fit up to approximately 64.1% of the data at almost the same RMSE (407).
3. *The values of net-thermal demand range up to a high of 3600Wh. An RMSE of 400Wh, states an approximate statistical error of a maximum of 11.1%.*
4. *The residuals obtained were put to a normality test, wherein it was found that they were normally distributed.* (See Appendix A.7.2 for explanations and graphical images) Also, the residuals had an extremely low value for mean, making it equivalent to 0.

Inferences from the analysis of figure 24

1. *In general, the fit is quite poor (even though the adj. R^2 of the fit was fairly high) and there is the lack of estimation of peak values. This is seen especially during the heating hours.*
2. *This could be due to the following reasons:*
    a. ***Type and size of dataset*** *- The actual profiles of heating and cooling demand as mentioned before is only a total of 1250 hours out of the 8732 hours (see table 5).* Of the entire dataset

used (8372 hours) a majority of these hours were without any thermal demand. Therefore, the fit on 1250 heating/cooling hours, whilst training on 7 times the data size could lead to such a poor fit. Out of this, the heating hours are a bare minimum of 411, making the training over 8732 hours even more difficult during the heating period. This could also explain why the fit on the cooling hours is slightly better than those on the heating hours.

b. ***Interdependent predictor variables.*** As shown in chapter 6 and 7 (see table 9) there is a degree of linear and non-linear relationships between certain predictor variables. We already know that MLR equations needs the predictor variables to be as independent as possible.

c. ***Non-linear relations between predictor and target variables*** – It is also known that the predictor variables must be linearly dependent on the target variable (net thermal demand). However, this is never the case in the real world, as can be seen from the plots in chapter 6. In fact, non-linearity increases whilst using a sub-set data of both heating and cooling period. (Example the relationship between thermal demand and outdoor temperature varies during heating periods and cooling periods, and is not always linear) It would be better to fit a model over heating periods and cooling periods individually.

This is the reason why the automated MLR step-toolbox created during this research makes the use of different *sub-datasets.*

To understand the fitting abilities of the MLR functions the model takes the following ranges into account;

- Datasets of a whole year (already tested)
- Data set belonging to working hours (based on opening hours of the building)
- Data set belonging to nonworking hours (based on closing hours of the building)

Based on operational use, i.e., working and non -working hours

The working and non-working hours were defined based on the building operational time of use and not room occupancy (see figures in the appendix 7.2.2). Sundays were maintained as non-working and other days were maintained from 6am-7pm. Table 14 shows 1st estimation is for working hours and the second for non-working hours, with their respective p-values. The results obtained by using either of the two functions were yet again very similar, with slight improvements in STEPWISE FIT with regards to the overall RMSE of the model. See appendix section 7.2 for all the obtained results including those for FITLM.

Table 14 Estimated coefficients for working and non-working hours using the Stepwise Fit function only.

	Working hours of the year		Non-working hours of the year	
Constant	**Estimate**	**P -Value**	**Estimate**	**P-value**
Intercept (C0)	-68.42	0.00	-7.07	0.00
C1	143.45	0.00	133.08	0.00
C2	382.38	0.00	201.39	0.00
C3	13.57	0.00	4.48	0.00
C4	-68.18	0.00	-6.27	0.00
C5	0.00	1.00	0.00	1.00
C6	-0.21	0.00	0.09	0.00
C7	-0.06	0.07	0.02	0.66
Adjusted R-Squared	**69.7%**		**30.0%**	
RMSE	**507**		**192**	

Inference

1. ***Higher R-squared value for working hours*** -There is an increase in R-squared value for fitting over the working hours (69.7%), as compared to the entire year (64%). The non-working hours experience very few hours of heating/cooling demand, (see the point made below). On removing these hours, the ratio of actual thermal hours to the total training hours has been increased unlike in the full year's training set. Thus, a better fitting has been achieved by eliminating the non-working hours.
2. ***Poor performance with non-working hours*** - Due to such low number of hours of thermal demand during the non-working set of hours through the entire year, the fitting of data was extremely difficult. From a total of 4155 hours of non-working (night hours and Sundays) in the year, there were a total 128 recordings of hourly thermal demand consumption. Thus, training data over 128 hours from a set of 4155 hours leads to extremely poor results, with excessive errors. Thus, the model is nullified.
3. ***Smaller Timesteps for higher efficiency*** – An important inference is that a higher ratio of actual target values to the total number of training values, leads to a better training of the data. Further reducing the sub-dataset sizes to month, week or day might lead to even better R-squared values and lowered RMSE for MLR models and better visualization of a fit. However, there may be issued of overfitting which must be taken into consideration. Overfitting occurs either when the model is too complex, and includes more than needed parameters for training the model. It leads to a highly efficient fit of the model on the dataset, however, the model

reacts poorly to changing parameters whilst predicting future value (see xx for more information on overfitting)

Since the results obtained from the two functions are comparable in nature, the research focusses on using *STEPWISE FIT* for further prediction of data.

8.5.2 Predictive power of the MLR models using different sub-datasets.

This section shall now look into predictive capabilities of the MLR model. Based on the findings above, it has become clear that proper visualization and higher accuracy of training could be obtained with smaller sub-datasets. A higher accuracy in training also could also lead to a higher accuracy in prediction of reliable data, except when there is overfitting. To understand the fitting as well as predictive abilities of the MLR functions the model takes the following smaller ranges into account based on user input;

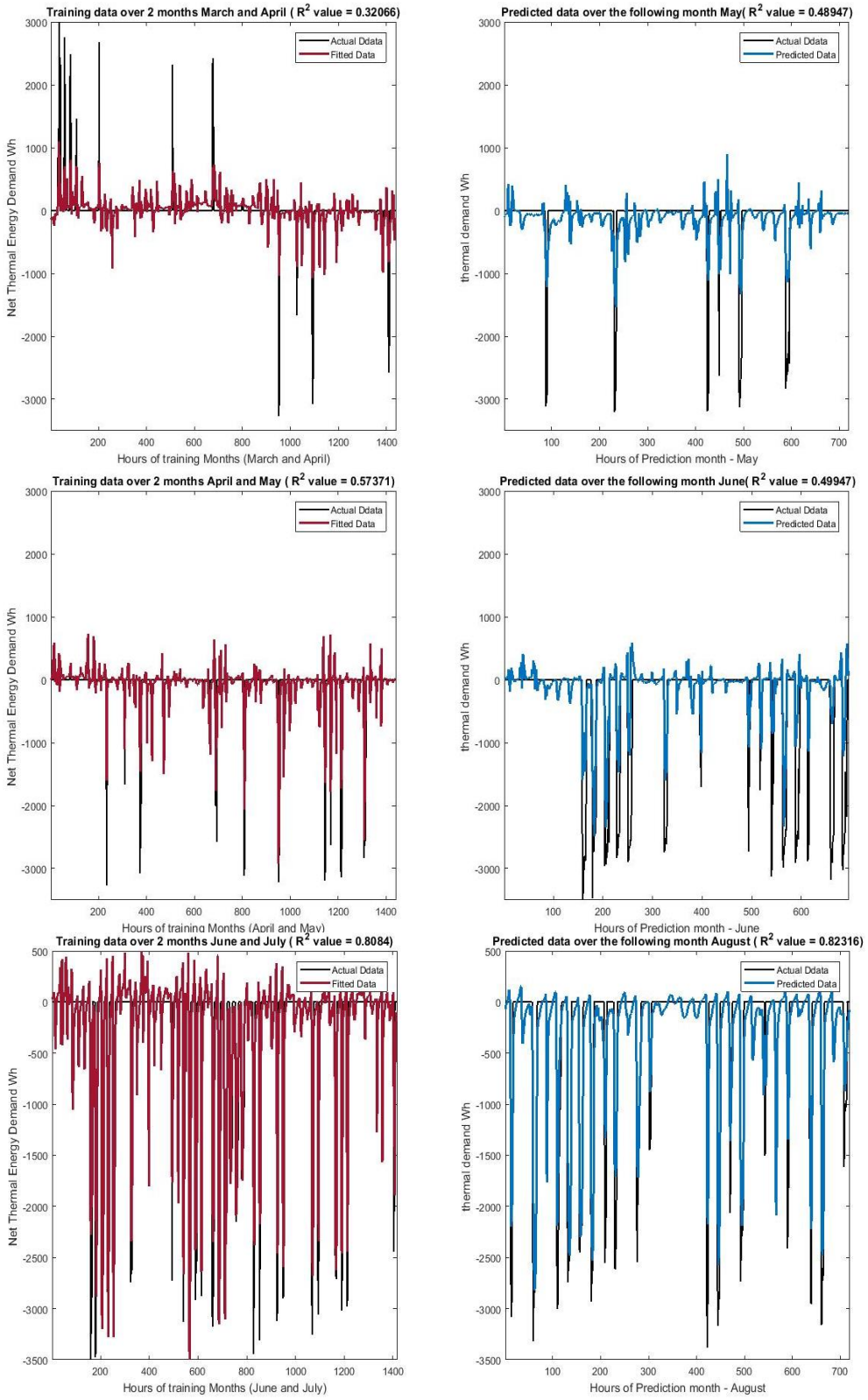
- Datasets belonging to month (May and June 2015) for fitting/training and thereafter 2 months (July-August) for prediction.
- Week-wise datasets – weeks 24 and 25 for training purpose, and week 26 for prediction. These are dates from 7th to 20th June 2015 for training and 21st to 28th June for prediction.
- Single Day wise datasets from any month. This is followed by a day-ahead prediction.

Monthly Data-sets

Using the two months as training period for the MLR model, the prediction of thermal energy demand for the following month was made. During the research three different sets of data were chosen, see table 15 below, and all three sets had a vast difference in the R-squared (or goodness of fit) of the data.

Table 15 shows the varying timeframes chosen for training and predicting the data and the associated adjusted R-squared values obtained from the MLR. (addto appendix how r2 is calculated)

Data Set	Training Months	Adj.R-Squared [%]	RMSE Wh	Prediction Month	RMSE Wh	Adj.R-squared [%]
1	March & April	32.0%	298	May	371	48.95%
2	April & May	47.1%	283	June	614	50%
3	June & July	80.5%	331	August	342	82.3%



Figures 25 a, b and c Training and data prediction of the monthly datasets used on the Stepwise fit MLR model. These three graphs showcase the three different sub-datasets considered during research and have varied adjusted R^2 values.

Inferences

1. ***Regarding the groups of months chosen*** – these months represent different periods of the data, during which the building performed in different modes to maintain the indoor comfort level. For example, during March it can be seen (figure 25 a) that there is a heating demand for half the month, followed by a cooling demand for the rest of the month, and for the whole of April. This is the trained model, applied to the following month of May, where there is only cooling demand. Thus, the predictive power is low. However, for the months of June and July, there is a prevailing constant cooling demand, and thus predictions of this trained data, over August, which have the same characteristic, is much accurate. Accuracy of such models decreases when training data deviates from testing data [32]. Thus, the training should be fixed over either cooling or heating to reduce the non-linearity when using MLR.
2. ***Regarding the significance of the model*** – the training models and the predicted data exhibit normally distributed residuals. The B values (weight of the effective parameters) and the corresponding p-values have been placed in the table 16 below.

Table 16 estimated values of the coefficients and their corresponding p-values for the three groups of training over monthly data-subsets. The red markings are insignificant parameters.

		March and April		April and May		June and July	
Constants	**Physical Significance**	**B (coefficients)**	**P-value**	**B (coefficients)**	**P-value**	**B (coefficients)**	**P-value**
C1	$C_1 \sim \sum_i U^i . A^i$	2.25	0.37	15.50	0.00	26.34	0.00
C2	$C_2 \sim U_{floor} . A_{floor}$	190.69	0.00	229.01	0.00	406.97	0.00
C3	$C_3 \sim \alpha_i . A_{indoor\ surfaces}$	360.92	0.00	304.13	0.00	-26.67	0.57
C4	$C_4 \sim M_{vent} . C_{p\ air}$	2.25	0.37	0.00	1.00	0.00	1.00
C5	$C_5 \sim (m_{openings} + m_{cracks}) . C_{p\ air}$	0.00	1.00	0.00	1.00	0.00	1.00
C6	*Solar radiation effect*	-0.06	0.13	-0.14	0.00	-0.20	0.00
C7	*Internal heating effect*	0.04	0.45	-0.22	0.00	-0.26	0.00
R2-fit [%]	-	**32%**	-	**57%**	-	**84%**	-
R2-pred. [%]	-	**49%**	-	**50%**	-	**82%**	-
RMSE fit [Wh]	-	**298.60**	-	**283.38**	-	**331.36**	-
RMSE pred. [Wh]	-	**371.41**	-	**614.41**	-	**342.62**	-

From the table, it becomes evident that the floor surface temperature, the indoor air temperature and the outdoor air temperature are the variables which have the most influence in training and therefore also on the prediction of the thermal energy demand. The linear equations thus developed are;

$$Q_{thermal\ demand\ March\text{-}april}\ [Wh] = -32.12 + 190.69(T_{floor} - T_{indoor}) + 360.92(T_{wall} - T_{indoor}) \quad \text{Eq-8.6}$$

$$Q_{thermal\ demand\ april\ -\ may}\ [Wh] = -21.12 + 15.50(T_{outdoor} - T_{indoor}) + 229.01(T_{floor} - T_{indoor}) + 304.13(T_{wall} - T_{indoor}) - 0.14(Q_{solar}) - 0.22\ (Q_{internal}) \quad \text{Eq-8.7}$$

$$Q_{thermal\ demand\ june\text{-}july}\ [Wh] = -29.79 + 26.34(T_{outdoor} - T_{indoor}) + 406.97(T_{floor} - T_{indoor}) - 0.20(Q_{solar}) - 0.26\ (Q_{internal}) \quad \text{Eq-8.8}$$

The coefficients vary for each group of sub-datasets as the correlation between variables differs over different periods of the year.

The next step of this research is to eliminate trainings over simultaneous heating and cooling periods to check for the accuracy of the MLR models in predicting thermal energy demand. One method of doing this is by using weekly. The benefits of such a plot are that *firstly* the variation in outdoor climatic conditions are much lower over the span of a week. *Secondly*, a much sharper and defined visualization of the data can be made during a smaller plot over a week.

Weekly Data-sets

The MLR model was run on two periods of the year, one belonging to the summer month of June 2015, and one sub-dataset belonging to January 2015[12]. The model was fitted onto the two sub-datasets, and thereafter, the following week for prediction. The significant coefficients obtained and the predictive values have been shown in the table 17 and graphs 26 a and b below.

Table 17 estimated values of the coefficients and their corresponding p-values for the two groups of training over weekly data-subsets.

		Summer (weeks 24-25)		Winter (weeks 3-4)	
Constants	**Physical Significance**	B (coefficients)	P-value	B (coefficients)	P-value
C1	$C_1 \sim \sum_i U^i . A^i$	21.10	0.00	34.30	0.02
C2	$C_2 \sim U_{floor} . A_{floor}$	410.66	0.00	436.73	0.00
C3	$C_3 \sim \alpha_i . A_{indoor\ surfaces}$	-158.01	0.10	-293.46	0.10
C4	$C_4 \sim M_{vent} . C_{p\ air}$	0.00	1.00	0.00	1.00
C5	$C_5 \sim (m_{openings} + m_{cracks}) . C_{p\ air}$	0.00	1.00	0.00	1.00
C6	*Solar radiation effect*	-0.16	0.09	-0.28	0.28
C7	*Internal heating effect*	-0.54	0.00	1.95	0.00
R2-fit [%]	-	**88%**	-	**82%**	-
R2-pred. [%]	-	**78%**	-	**80%**	-
RMSE fit [Wh]	-	**320.39**	-	**446.29**	-
RMSE pred. [Wh]	-	**423.98**	-	**353.54**	-

[12] During the heating period, the week in January dated 7th -15th of Jan 2015, was chosen instead of two weeks for training, as the months of Jan -February have quite a few missing days in the data obtained from the room 1075.

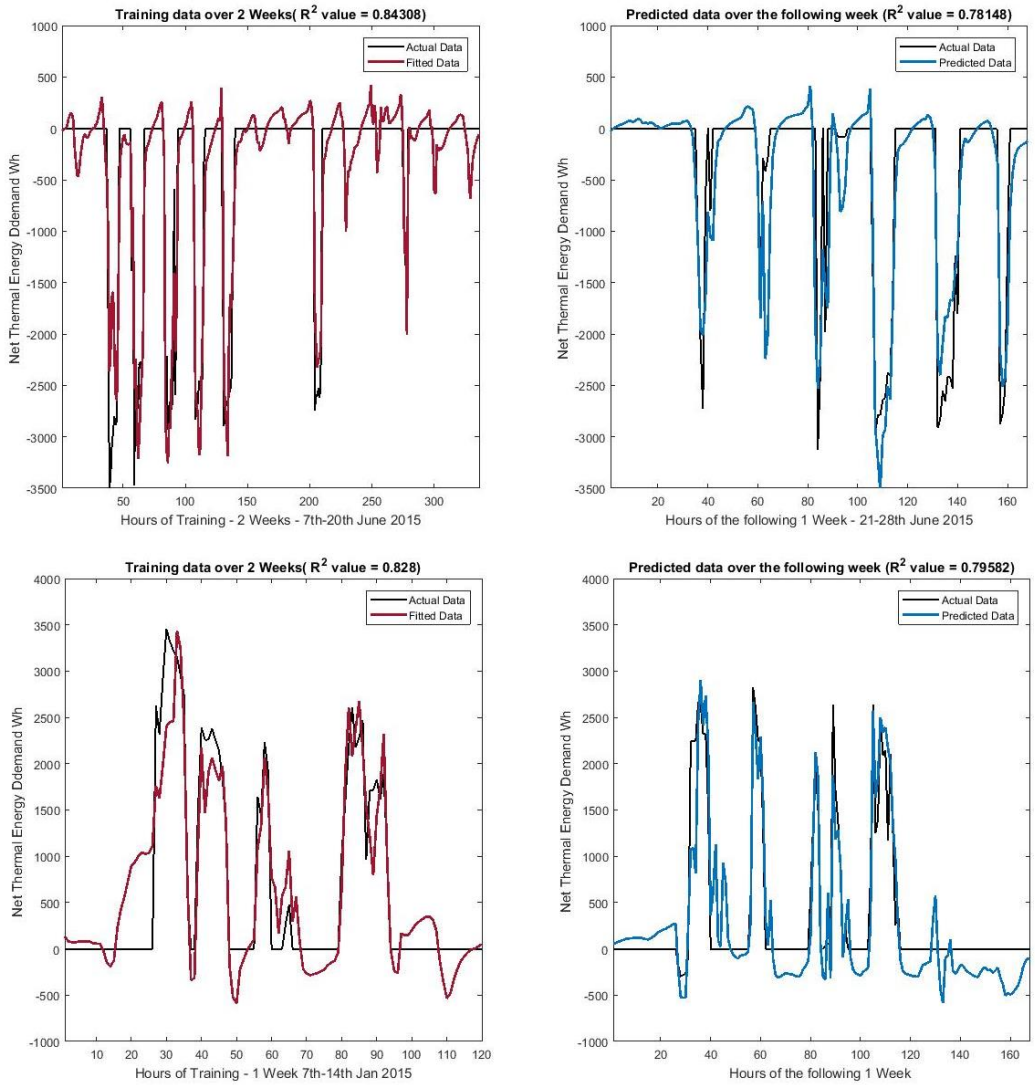
From the table 17, it becomes evident that again the floor surface temperature, the indoor air temperature and the outdoor air temperature are the variables which have the most influence in training and therefore also on the prediction of the thermal energy demand. There is a slight influence of the internal heat since both the sub-datasets were accompanied by occupancy. The linear equations thus developed are Equations 8.8, and 8.9;

$$Q_{thermal\ demand\ week\ 24\text{-}25\ (summer)}\ [Wh] = -2.9 + 21.1.(\ T_{outdoor} - T_{indoor}) + 410.65(T_{floor} - T_{indoor}) - 0.54\ Q_{internal}$$

Eq-8.8

$$Q_{thermal\ demand\ week\ 3\text{-}4\ (winter)}\ [Wh] = -21.9 + 34.3.(\ T_{outdoor} - T_{indoor}) + 436.73(T_{floor} - T_{indoor}) + 1.95\ Q_{internal}$$

Eq-8.9



Figures 26 a and b are the training (weeks 24-25 and 3 respectively) and data prediction (weeks 26 and 4 respectively) of the weekly datasets used on the Stepwise fit MLR model. These two graphs showcase the two different sub-datasets considered during research and have varied adjusted R^2 values.

As expected, the adjusted R^2 value for the training data is high (80-85%), with a prediction adjusted R^2 value of 78-80%

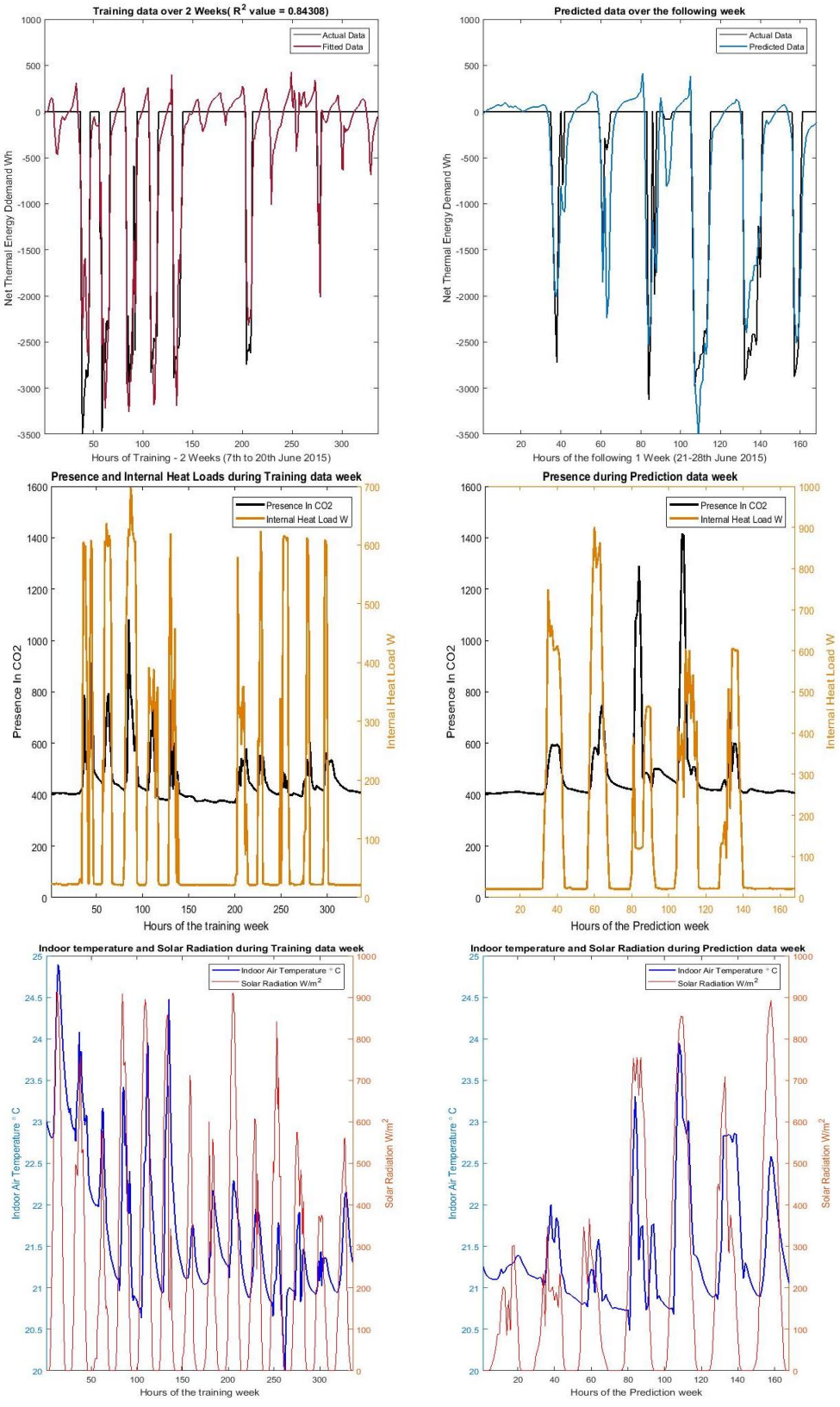
Observations

1. The weekly plots show a higher consistency in data. The model is able to train sharply for the weekends, as it can be seen that there are days with thermal demand corresponding to presence and weekdays (figure 26 a hours ~25-145), followed by two days of no thermal demand, belonging to a weekend (hours ~146-195).
2. During the second week of training, the actual thermal demand is 0W compared to the trained dataset. However, a quick analysis of the room (see figures 27 b below) during that week shows that there was a certain degree of presence and internal heat load from lighting and appliances in the room leading to a small amount of thermal demand in the model.

Inferences

1. ***Higher accuracy in predictions*** – The predicted data is highly accurate with statistically valid coefficients as seen in table 17 above. This high accuracy (R^2 84%) in fitting is due to consistency in data over the training period, unlike the monthly plots, (figure 26 a) wherein the discontinuity (sudden an abrupt changes) in the heating and cooling demand was observed due to a shift in the outdoor climatic conditions or the schedules of the room.
2. ***Absence of cooling during high solar hours*** [13] – During hours – *245 to 260*, there is a high amount of solar radiation and certain presence level which has led to almost 1000W of cooling demand by the model. However, the indoor temperature has been low enough for the floor systems to not perform cooling actions, thus actual thermal readings have been 0. This leads to believe that the model shrinks in accuracy with the addition of large variation in patterns of predictor variables, introducing non-linearity in them.
3. ***Discrepancies in Fitting*** – Although for both monthly and weekly plots, the MRL model was able to account for the important parameters, there seems to be a variance of 20-25% in the data-sets which were not answered by the fitted model. This could either be pure noise, not being explained by the MLR input parameters, which may suggest to another parameter(s) missing in the dataset.

[13] Another probable reason for not supplying cooling energy, as compared to the reason sated above in point two, could be because the floor heating and cooling systems is a slow system and does not react as immediately towards sudden excitations (See section 6.3)



Figures 27 (a, b and c) show the two training weeks followed by the prediction week and the associated thermal energy, air temperatures and solar radiation, and the occupancy levels.

Similarly, even smaller data-subsets involving day ahead prediction were developed in this MLR step-toolbox.

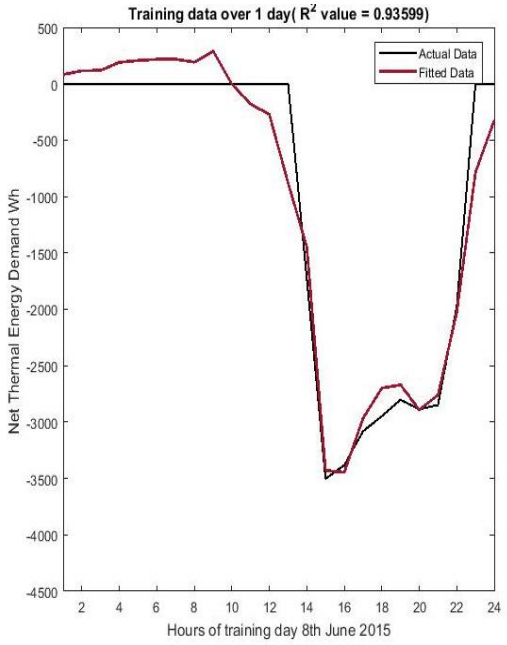
Daily Data-sets

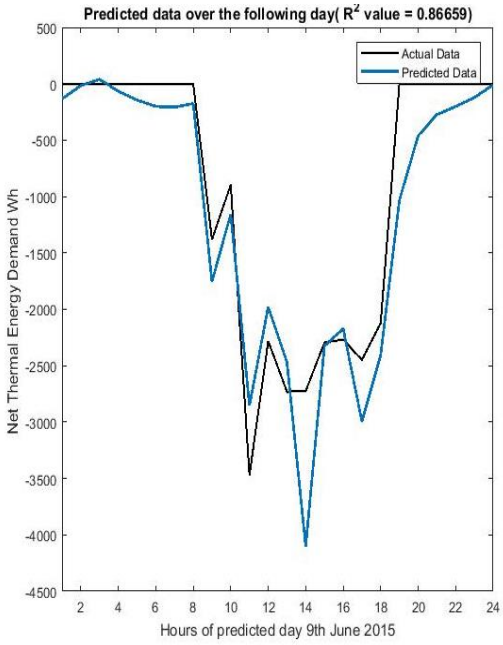
The stepwise fit function was used towards training a 24-hour model, to predict the following day. This resulted in an extremely good fit and prediction since in each week, the classroom faces much lesser discrepancies in values in terms of climatic conditions outdoor, or indoor temperature –

Two separate groups of days were chosen prediction.

- Training on a day with occupancy (8.6.2015) followed by prediction of the next day *with* occupancy. (9.6.2015)
- Training on a day with occupancy (12.6.2015) followed by prediction of the next day *without* occupancy (13.6.2015)

The adj. R^2 values for both the training plots were almost 93% and 96% respectively with a RMSE of approximately 300-400Wh. The R^2 values for the predicted data were 87% . The R^2 for the secondary plot is infinity, as the squared sum of deviation of the data itself is 0, since the net thermal demand on this day was 0. Thus, the only way of measuring the efficiency in prediction for 13th June 2015, is by the help of the RMSE value of 278 Wh. Figures 28 a and b below show the two sets of training and prediction.





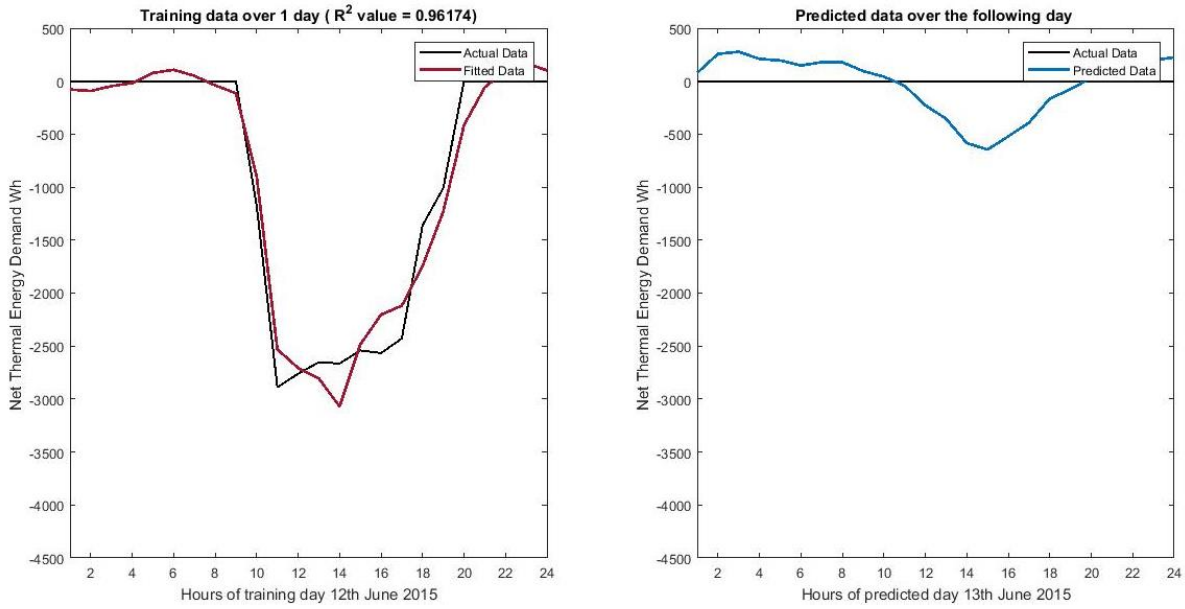


Figure 28 a and b show the day wise training and prediction using the Stepwise fit function. Figure a is for the first group of dates 8th and 9th October, whereas the figure b is for the second set of days, 12th and 13th of October 2015.

Table 18 estimated values of the coefficients and their corresponding p-values for the two groups of training over weekly data-subsets.

	8TH 9TH JUNE 2015		12TH AND 13TH OCT 2015	
	B (coefficients)	P-value	B (coefficients)	P-value
C1	22.82	0.55	44.49	0.01
C2	665.02	0.00	355.78	0.00
C3	-1005.18	0.01	-59.18	0.80
C4	22.82	0.55	0.00	1.00
C5	0.00	1.00	0.00	1.00
C6	0.02	0.97	-1.32	0.00
C7	0.79	0.06	-0.28	0.47
R2-FIT	0.95		0.97	
R2-PRED	0.87		-	
RMSE FIT	325.41		234.45	
RMSEPRED	437.50		278.14	

$$Q_{thermal\ demand\ 8th\text{-}9th\ June\ 2015}\ [Wh] = -535.76 + 665.02(T_{floor} - T_{indoor}) - 1005.18(T_{outdoor} - T_{indoor}) \quad \text{Eq-8.10}$$

$$Q_{thermal\ demand\ 12th\text{-}13th\ June\ 2015}\ [Wh] = 222.20 + 44.5(T_{outdoor} - T_{indoor}) + 355.78(T_{floor} - T_{indoor}) - 1.32\ Q_{solar} \quad \text{Eq-8.11}$$

1. *It is inferred that the models are not overfitted while training.*
2. *The models respond much better to abrupt changes in routines when trained over smaller time periods. However, this can vary if there is inconsistency in data between two days (for example using a sunny day in winter to predict during a cold cloudy day).*

An important aspect to account for is that the coefficients calculated by the model, for all the different sub-datasets differ. However, these coefficients are relevant only when similar data is chosen as input predictor variables. Therefore, the efficiency and significance of the MLR toolbox is not dependent on the size of the sub-dataset, but on the type of sub-dataset. Using data from colder months to train a model, will not help in predicting periods belonging to the warmer months.

The reason for choosing the periods of sub-datasets are as follows;

1. *The monthly sub-datasets* are used to show the affects of discontinuity in data (March and April) over the pros of fitting a model on continuous and less abruptly varying data (June and July).
2. *The week-wise periods* were chosen to show training and predictive capabilities over both heating and cooling demand – this was not done with the monthly sub-datasets as there are very little heating hours scattered over the entire year. It also proves that training over two continuous weeks is good enough to predict over the following week.
3. The day-wise sub-datasets were chosen at random.

8.6 Improvements in MLR models

The MLR models developed over the three different periods were further improved by adding and/or improving on certain parameters such as;

1. ***Accounting for the delay in solar radiation*** **–** During chapter 6 (graphical analysis) it was seen that the indoor temperature peaks response is delayed by an hour to that of the solar radiation peaks (see section 6.3). This was also noted by Lopez [15], who was able to improvise on her model by taking into account this lag in response. Thus, a lag of 1 hour and 2 hours was introduced in the data with respect to solar radiation.

2. ***Using the indoor air temperatures of the side rooms to evaluate their effect on the room in question*** **–** Since the heating and cooling of the rooms is done via a priority based system (see section 6.3 and Appendix A1) perhaps accounting for the adjacent rooms may help improvise the model. This would be taking into consideration if the side rooms need heating or cooling demand.

Results and Discussions

With regards to accounting for the delay in solar radiation

The model was trained over June and July to predict for August, weeks 24 and 25 to predict for week 26, and over 8th June 2015, to predict for 9th June 2015.

There was only a slight increase in the R^2 value in prediction and fit and an improvement seen with regards to the RMSE values for the monthly sub-dataset. With regards to the weekly and day-wise sub datasets, no real increase in prediction could be noticed. However, the underlying relationship between the effective parameters ad the dependent variable was seen to alter slightly as shown in the table 19 below.

The coefficient for solar radiation was slightly more significant by weight for the monthly and weekly sub-datasets. However, almost no change is observed during the day-wise sub-dataset. The graphs for the training and prediction over these three sub datasets has been placed in the appendix A.7.5.

The lack of improvement in the MLR models of this particular room could be due to the fact that the effect of solar radiation on the thermal demand, is only slightly significant, as can be seen from figure xx in the appendix. The coefficients estimated by the MLR for all models above also shows that solar radiation is has a very small significance in answering for the thermal energy demands.

Table 19 - estimated values of the MLR models, on accounting for the lag in solar radiation, as compared to the values estimated without accounting for the delay.

		June and July		June and July (with 1-hour Delay in data)		Summer week 24-26 2015		Summer week 24-26 2015 (with 1hour delay in data)		8^{th} -9^{th} June 2015		8^{th} – 9^{th} June 2015 (with 1-hour delay in data)	
	Physical Significance	B coefficients	P-value	B coefficients	P-value	B coefficients	P-value	B coefficients	P-value	B coefficients	P-value	B coefficients	P-value
C1	$C_1 \sim \sum_i U^i . A^i$	26.34	0.00	20.74	0.00	21.10	0.00	21.67	0.00	22.82	0.55	22.82	0.55
C2	$C_2 \sim U_{floor} . A_{floor}$	406.97	0.00	460.07	0.00	410.66	0.00	468.44	0.00	665.02	0.00	665.02	0.00
C3	$C_3 \sim \alpha_i . A_{indoor\ surfaces}$	-26.67	0.57	-113.95	0.01	-158.01	0.10	-321.50	0.00	-1005.18	0.01	-1005.18	0.01
C4	$C_4 \sim M_{vent} . C_{p\ air}$	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	22.82	0.55	22.82	0.55
C5	$C_5 \sim (m_{openings} + m_{cracks}). C_{p\ air}$	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
C6	***Solar radiation effect***	**-0.20**	**0.00**	**0.42**	**0.00**	**-0.16**	**0.09**	**0.49**	**0.00**	**0.02**	**0.97**	**0.20**	**0.76**
C7	*Internal heating effect*	-0.26	0.00	-0.37	0.00	-0.54	0.00	-0.56	0.00	0.79	0.06	0.79	0.06
R2-fit	-	**84%**		**84%**		**88%**		**88%**		**95%**		**95%**	
R2-pred	-	**82%**		**83%**		**78%**		**78%**		**87%**		**87%**	
RMSE fit [Wh]	-	**331.36**		**329.00**		**320.39**		**326.6**		**325.41**		**325.41**	
RMSE pred [Wh]	-	**342.62**		**334.00**		**423.98**		**430.57**		**437.50**		**437.5**	

The MLR equations obtained were as follows;

$$Q_{thermal\ demand\ june\text{-}july}\ [Wh] = -26.79 + 20.74(T_{outdoor} - T_{indoor}) + 460.07(T_{floor} - T_{indoor}) - 113.95(T_{wall} - T_{indoor}) + 0.42(Q_{solar}) - 0.37\ (Q_{internal})$$

$$Q_{thermal\ demand\ week\ 24\text{-}25\ (summer)}\ [Wh] = -5.9 + 21.67(T_{outdoor} - T_{indoor}) + 468.44(T_{floor} - T_{indoor}) - 321.50(T_{wall} - T_{indoor}) + 0.49(Q_{solar}) - 0.56\ Q_{internal}$$

$$Q_{thermal\ demand\ 8th\text{-}9th\ June\ 2015}\ [Wh] = -515.76 + 665.02(T_{floor} - T_{indoor}) - 1005.18(T_{outdoor} - T_{indoor})$$

With regards to using the indoor air temperatures of the side rooms – It was seen that the model did not improve, but only became more complex leading to increased R^2 and RMSE values for the weekly and monthly dub-datasets. Refer to Appendix A.7.5 for more details.

8.7 MLR Using Sub-Datasets from the Previous Year

One of the major objectives of this research is to provide a thermal energy predictive model with physical and practical meaning. The models described in the former sections were done so with known, recorded parameters. The prediction too is based upon known future values of the very same parameters. Thus, though the research provides a detailed analysis of the predictive capability of the MLR models, it is important to understand the applicability of such models.

8.7.1 Methodology and Data Selection

This chapter focuses on another classroom of the HHS (1087) [14] with two years of data. A weekly timestep is chosen, and MLR models are trained with known parameters from the year 2015. The trained model is then used to predict future thermal energy demand values of the following week with the help of

1. Predictor variables from the same year – 2015 (as done before)
2. Predictor variables from the past year – 2014

The availability of input data (predictor variables) for prediction purposes is summarized below:

1. *Using forecasts from weather stations regarding outdoor temperature, and solar radiation.*
2. *Using occupancy related patterns based on schedules of a classroom to predict* $Q_{internal}$
3. *Using data from previous year as input predictor variables.*

The scope of this research limits to using previous year data. The use of forecasts from weather station and predicted patterns of ventilation and $Q_{internal}$, and other significant parameters has been discussed in the chapter 11 future recommendations.

Keeping this in mind, the model is trained over two weeks (7^{th}-20^{th} June 2015), and the thermal energy demand for week number 26 (june 21^{st} to 28^{th} 2015) is predicted. The objective is to predict for week 26 2015, by using the input parameters from week 26 of 2014.

Comparing the two weeks of prediction in both years is important to witness any major differences in some important input parameters. We see (figure 29 below) that the input data for the predictive week in 2015 (week 26), are varied in terms of values to those obtained from 2014 for the same week. There is a large difference in outdoor air temperatures for the first few days of the week. There is also a large difference in ventilation flow rate and solar radiations.

[14] Data from 2014 for the room 1075 was not in order or complete. Thus, a new room is shown.

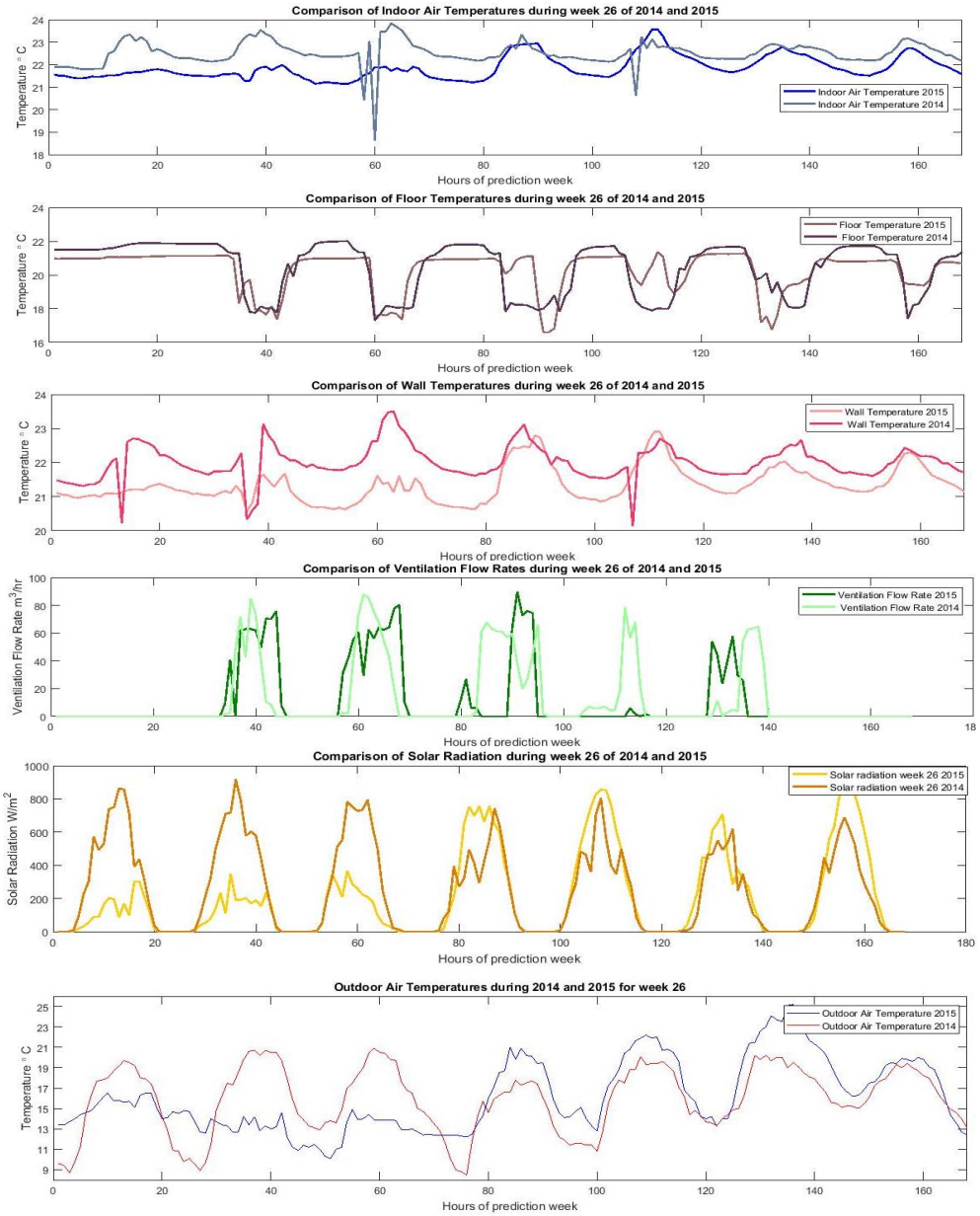


Figure 29 Graphical images comparing different recordings of room 1087 during the same week (week number 26) of two years 2014 and 2015. This week is used for prediction of thermal demand in 2015.

8.7.2 Results and Discussions

The obtained coefficients from training the data over week 24 and 25 (7th June 2015, to 20th June 2015) were significant only for the 4 thermal fluxes, namely, Q_{floor}, $Q_{envelope}$, $Q_{ventilation}$ and $Q_{internal}$ *(see the equation 8.12 below)*. The adjusted R^2 value for this training was 81%. The predictive capability of this model was (R^2 prediction - 71%) close to the actual demands of 2015, as can be seen from figure 30 a below. This was done by using the input parameters of week 26 from the year 2015. However, on using input parameters from the data of week number 26, *2014* the predictive capability -77%, indicating an extremely high amount

of error in thermal demand prediction as compared to the mean variation in the thermal demand of 2015 week 26 itself (figure30 b).

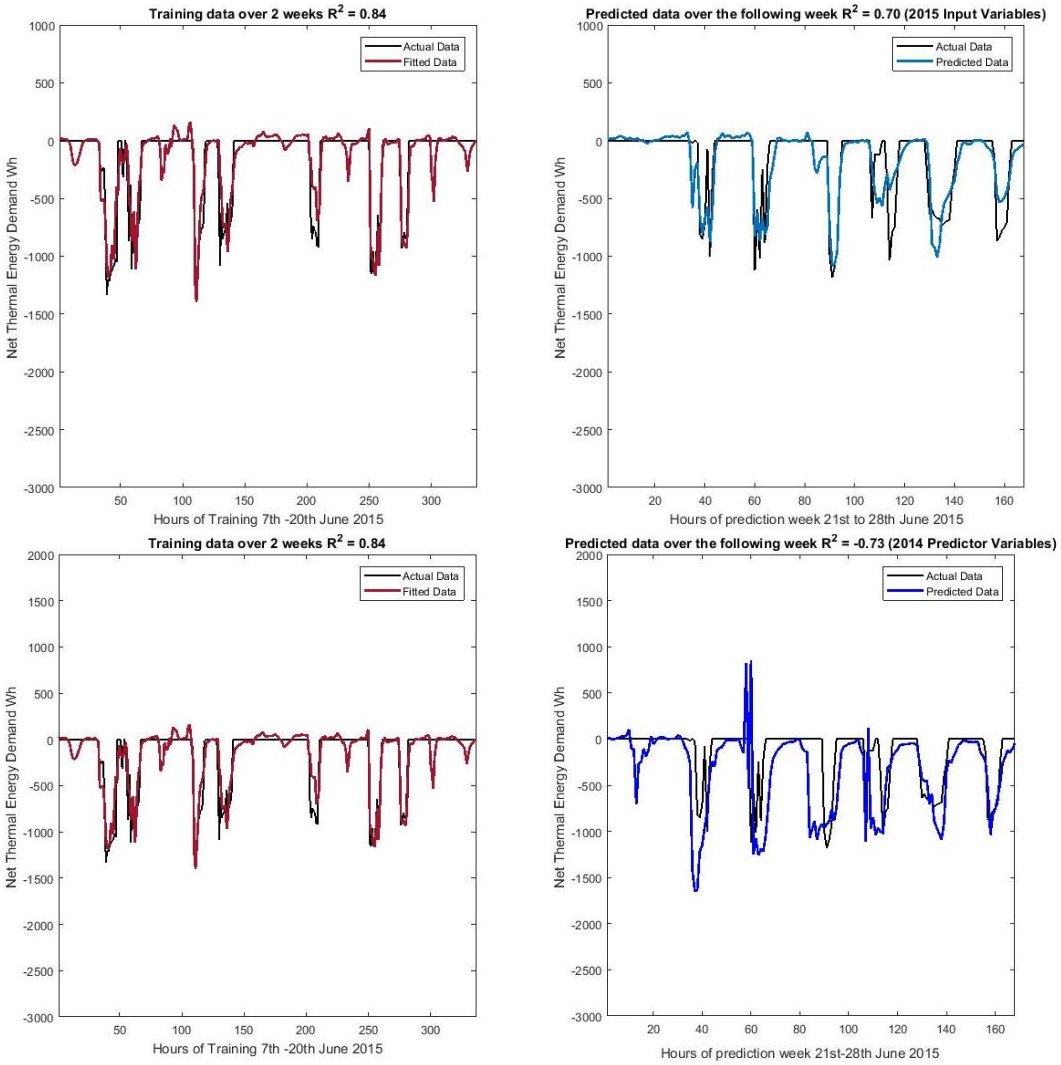


Figure 30 a. Training over data from two weeks (7th to 20th June 2015), and prediction of the following week 21-28th June 2015 using input data from 2015 and b. shows the training over data from two weeks (7th to 20th June 2015), and prediction of the following week 21-28th June 2015 using input data from 2014.

$$Q_{thermal\ demand\ week\ 26}\ [W] = 225.58 + 203.37.(T_{floor} - T_{indoor}) + 4.80.(T_{outdoor} - T_{indoor}) + (-32.52).(T_{out\ AHU} - T_{indoor}) + (-0.45).\ Q_{internal}$$

Eq-8.12

It is important to note that according to the MLR equation, only a few parameters were major contributors towards the fit and prediction. Most important ones being (T_{floor}, T_{indoor} and $T_{out\ AHU}$). Apart from the ventilation flow rates, the (T_{floor}, T_{indoor}) are extremely comparable over the two years. This leads to believe

that normalization of the input data might be able to predict better due to lowered distribution between certain parameters such as the ventilation or $Q_{internal}$. The normalization was carried out by using the *'Z score'* function on MATLAB [46], where the input and target values were normalized between [0 to 1].

The result of normalization was indeed better. The R^2value for the fit was 83% and the predicted R^2 value was now at 20%, on using data from week 26 from the year 2014 as the input variables (see figure 30 c below). The results for using 2015 as input predictors was the same as above.

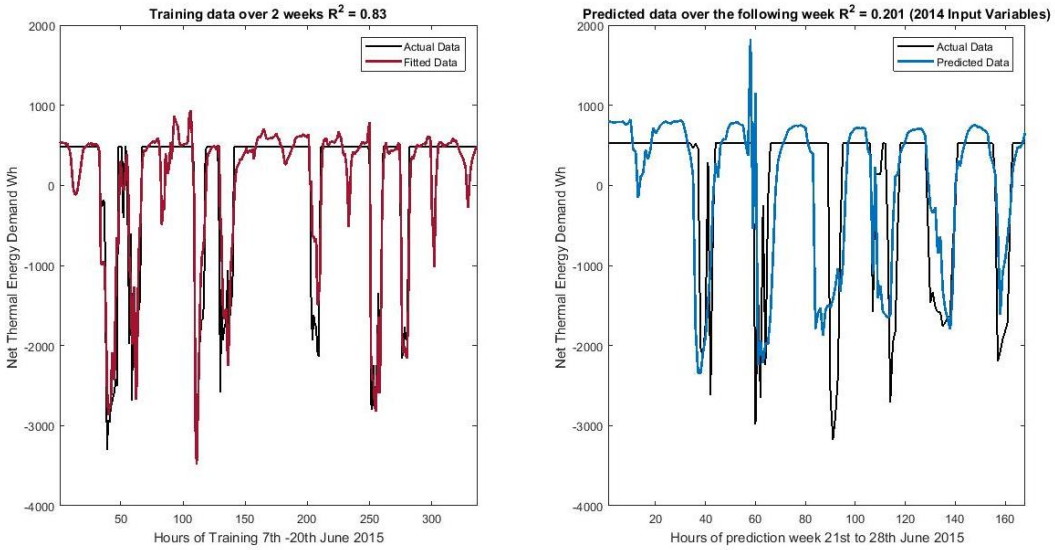


Figure 30 c training over normalized data from two weeks (7th to 20th June 2015), and prediction of the following week 21-28th June 2015 using input data from 2014.

Inferences

1. *Poor prediction by using the input data from previous year*. – The R^2 value of 20% is extremely low, and could be due to the following reasons –
    a. *Changes in occupancy profiles* - Especially for school buildings wherein the room schedules for the HHS Delft change quarterly. Thus, with occupancy, the ventilation flow rate and internal heat loads can vary.
    b. *Variations in outdoor climate* - Also, as noticed there is quite a lot of variation in the outdoor climatic data over the two years for the two same weeks.

What follows is a conclusive brief study of the use, efficiency and results of the Multivariate Linear Regression models

8.8 Conclusion

This chapter has shown the use of a MLR in thermal energy balance models, for training and prediction of thermal energy demands at the room level, using actual data. Major conclusive points are described below, answering to some of the major research questions and sub-questions.

1. ***Regarding the most Important parameters*** **–** Through the entire chapter, it can be deduced that the most important parameters affecting the thermal energy training models are the floor surface, indoor, wall and outdoor air temperatures. This is related to the fact that the indoor thermal comfort is being maintained by floor heating and cooling systems, and the thermal demand is varying based on indoor air temperatures. This is similar to the results obtained in the chapter on correlation, chapter 6. However, it should also be noted that there could have been important parameters missing in the data sets which could explain for the variance not explained by the R^2 values.
2. ***Regarding discontinuity in data of predictor variables-*** It was found throughout the research, that the model efficiency is questionable whenever there is a high amount of discontinuity in the values of the predictor variables. This is the main reason why the R-squared value is not able to reach 90% predictive power. Especially over large datasets, like the yearly datasets, it is seen that the erratic behavior of the predictor variables leads to a poor fit with lowered R-square values. However, using yearly datasets makes it possible to identify heating and cooling periods, which is not possible when training data only on heating periods or cooling periods.
3. ***Efficiency in training–*** The model efficiency, measured by the R^2 value, is seen to grow with smaller datasets, at least for this particular highly efficient HHS building. The classrooms are scheduled 4 times a year, deciding upon the occupancy and use-time. The datasets could be less erratic if chosen from amongst these scheduled months. This can be related to the point above, wherein there is less erratic behavior of predictor variables leading to slightly higher efficiency in training. Having said that, the models were developed on real-time data and achieved a high efficiency of up to 85%. This helps us use the model for real time applications as well, (see the next chapter 9).
4. ***Efficiency in Prediction -*** The predictive efficiency, also measured by the R^2 values, does not simply depend upon the efficiency of the trained model, but also on how similar the independent parameters are in the period to be predicted as compared to the training period. A training and prediction cannot be performed over two varied sets of data in terms of occupancy, and climatic conditions. The prediction of August is lower in error than those of May and June, due to the fact that the training months June and July belong to the same summer season, and witness similar profiles of outdoor climatic conditions and indoor temperatures.

5. ***Regarding significance of coefficients –*** The calculated coefficients are physically equivalent to the area times U value of the envelope (C_1) or area times u value of the floor (C_2) etc., when trained over data belonging to a similar timestep. The chosen period for training the model, must be similar in terms of input or target variables, to the period to be predicted, to deem the estimated coefficients significant for prediction.
6. ***Regarding Use of Previous Year's data –*** in this research, it can be concluded that for a building of school or office origin, the use of previous year's data might not be the best option to produce a practical MLR prediction model, since not only the outdoor climatic condition, but also the occupancy profiles tend to change. Having said this, the past year data might work for certain office rooms, where the occupancy profile does not change much, and the room has no direct contact with the outdoor climate (room towards the inside of the building).

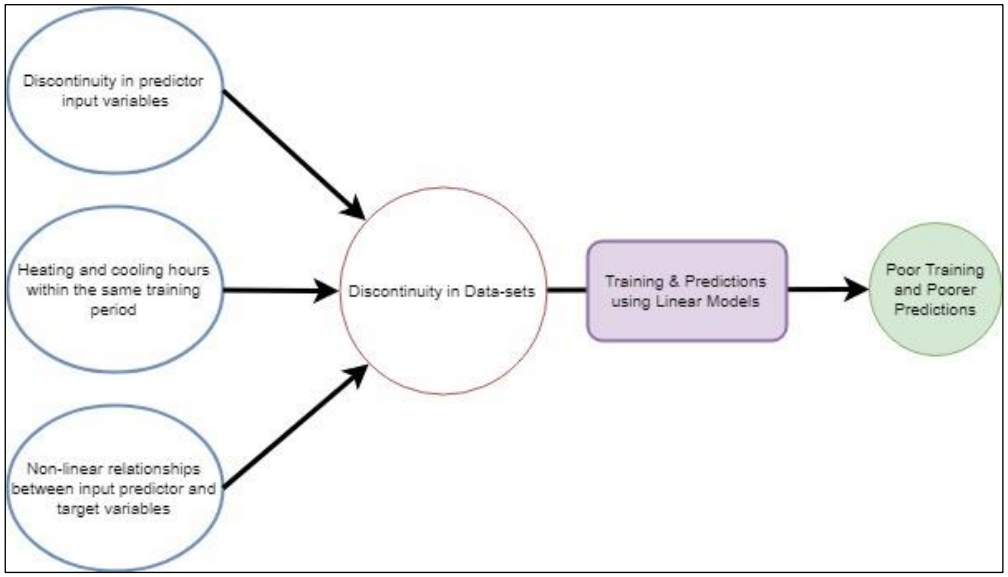


Figure 31 Summarizes the various sources of discontinuity in data seen in real-time data sets which lead to the reduced efficiencies of MLR models.

The MLR is not able to achieve efficiencies of 100%, even with certain improvements, as there is a certain amount of noise which is not being accounted for by the Linear Equations. This could be purely due to a missing important parameter, and/or the presence of discontinuity in input data, and/or due to the portions of relationships between the input and target variables, which are non-linear in nature.

Thus, the research looks in the direction for another black box model – one which can account for complex discontinuous data. The next chapter is about Artificial Neural Networks, as a secondary predictive black-box modelling technique to find thermal demand predictions at the room level of a school/office building.

9.Artificial Neural Networks

It has been shown by previous researchers that Artificial Neural Networks (ANNs) are now effectively being used to predict energy more reliably than traditional simulation models and regression techniques [47]. This chapter deals with a comprehensive small scale Artificial Neural Network (ANN) model development, to estimate higher efficiency prediction of thermal energy demand at room level for the building of HHS. From the findings of the previous two chapters, it is evident that the non-linear relationships of the predictor variable with the dependent variables are a major cause for lowered efficiencies (maximum up to 85%). Since ANN are nonlinear in nature, the predictive model developed using ANN shall become a solution for the unaccounted efficiency gap [32].

The first section shall give a brief overview of the ANN and its use in predictive modelling for thermal energy at the room level. The second section discloses the model architecture developed for the purpose of this research. This is followed by results obtained and the statistical validation of ANN models. Finally, a conclusive section talks about the comparative analysis of ANN and MLR models with regards to the case study.

9.1 Neural networks - An Overview

Several researchers have applied ANN successfully to building energy models and discovered it to be a more reliable prediction then other traditional statistical methods due to the ability of ANN to model non-linear patters [32] [47] [48] [49] . Recapping back to the literature on ANN (section 3.2), by definition, "A Neural Network is a non-linear mapping of the space between an input data set and an output data set and consists of three parts - an input vector (independent variables), an output vector (dependent variables), and an algorithm that maps the input space to the output space" [47]. The objective function of an ANN predictive model is to minimize the error between the actual and desired or predicted outputs of the network.

9.1.1 Working Principle of ANN

A typical neural network has been shown below. The figure 32 shows an input matrix, which is connected to 3 hidden layers which in turn is connected to an external output matrix (not shown). The number of layers can be chosen as per user discretion. ANN is a simulated connection of neurons. Similar to the function of a neuron in a human brain, these simulated neurons are used to receive and send input and output signals via their connections or synapse. Each hidden layer is made up of n-1 neurons, where n is the number of input variables. The synapses connect the layers to the external layers using weights (W). These weights are at first chosen randomly by the neural network. Along with the weight the model also chooses a random bias (b_1 -b_3) for each layer. The product of the input variable and its corresponding weight

are added to the bias of the particular neuron (denoted by ∑). A sigmoid transfer function *f* is applied to each neuron in each layer enabling them to uniformly approximate any continuous function [47]. The outputs of each neuron of a layer is the input for the corresponding neuron of the following layer [50]. Using a backpropagation (see section 9.1.2) algorithm, the weights and biases are adjusted to minimize the error in the fitted values.

The image 31 also shows the corresponding formula for each layer, wherein a^1 is the output of the first hidden layer, or the input to the 2nd hidden layer. First the input to the 1st hidden layer is calculated (denoted by n^1)

Therefore, $$n^1 = (\Sigma iw^{1,1} \times p) + b^1$$ Eq-9.1

Where $iw^{1,1}$ = the weight for the first input to the first neuron (1,1) in layer i.

p = the input (P_1-P_R)

b^1 = the bias applied to all the neurons of the 1st layer.

A logistic function is applied to this value n^1 to obtain the output of the first neuron of the first hidden layer.

$$a^1 = \frac{1}{1+e^{-n^1}}$$ Eq-9.2

a^1 is the sigmoid function (used in this research). However, there are several other functions to choose from.

For other training models and functions under ANN, refer to the book by Howard Demuth [50]

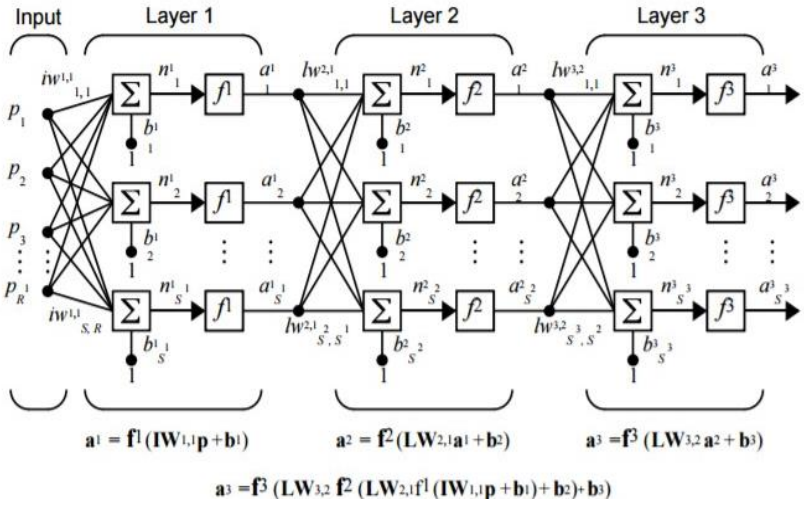


Figure 32 shows a diagrammatic explanation of a 3-layered Artificial Neural Network along with the input matrix and the corresponding formulas for the outputs of each layer [50].

9.1.2 Backpropagation Algorithm

The training of data in this research is done by using a standardized back-propagation algorithm. The objective of backpropagation is to optimize the weights and biases so that the neural network can learn how

to correctly map the calculated outputs, to the target outputs [48]. It corresponds to inverse modelling, wherein the weights are iteratively adjusted across the hidden layers in order to minimize the objective function. The model comes to a halt when the best fit has been achieved within limits of the overall error. This is done to ensure that there is no probability of overfitting the data (see section 9.2.2 ahead).

The obtained weights and biases from each layer belong to specific neurons and not to the input parameters This means that the values calculated during the entire operation, is not significant in defining the underlying relationship between the input parameter and the target variable, rendering the ANN models less useful for model predictive control designs.

The neural networks can be developed using n number of layers. However, an unnecessary addition to the number of layers adds to the model complexity of ANN. It also increases the chances of over-fitting training data, reduces the generalization capability of ANN and also increases the training error [32]. Therefore, it is important to choose the number of hidden layer neurons appropriately. Once again, like linear regression models, ANN models can be judged by the overall RMSE and R-squared value of the achieved fit of the estimated values over the actual data (see section 8.3 for more details on these two terms).

What follows is a detailed study of the model developed for this research and the methodology adopted towards making this model more automatic.

9.2 ANN Architecture – Model Development

There are two major categories of ANN models existing in literature [48]. The first one being *static* in nature, wherein the prediction model uses historical data to train and then predict future demand. This training does not change over time when new information becomes available. The second is *dynamic* in nature, wherein the predicted data along with future information is added to the historical data, for retraining the model. Dynamic models are said to be useful for short term prediction [48].

This research focuses on *simplified static models*, but automated in a way that the user can choose the period for training and prediction.

9.2.1 Static Automated ANN Models

The earlier chapters have mentioned the type and form of data available for research from the HHS building. The objective function of the model is to minimize the error associated with the estimated values of thermal demand for the chosen timeframe of data. Also from previous chapters it has been established that the net thermal demand of a room is a function of certain variables, which shall be used as inputs to the ANN model to predict future demands. The table explained in thermal energy balance principle section 8.1

(table 11-12) above, shows the various input parameters being used in order to estimate the output parameter.

As stated, from literature the backpropagation method was found to be most effective network for non-linear solutions under complex set of variables [49]. It was also found that the use of Levenberg-Marquart (LM) algorithm [48] is best suited for fast training of datasets in the case of a linear system of equations. The thermal balance principle section has already showcased the linear equation solved using Multivariate regression models. The same set of linear equations has been used for the ANN.

Thus, the model incorporates the backpropagation algorithm as the network type along with the use of Levenberg-Marquart (LM) algorithm as the training function.

Below a figure (figure 33) has been shown which represents the 3-layered model developed for this research. It consists of an input layer pointing to the number of input variables (7) added to the ANN model (there were seven predefined predictor variables in the previous chapter). The output layer contains one neuron consisting of the predicted value. The hidden layer consists of 2n+1 neurons (15).

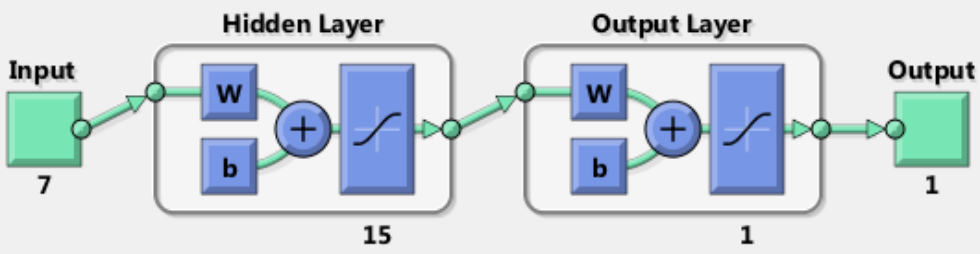


Figure 33 Snapshot of the developed neural network during this research. It shows a set of 7 inputs used for training the data within the hidden layer. There were 15 neurons placed in the hidden layer of this neural network.

9.2.2Overfitting

An important issue which leads to poor predictive abilities of a model is overfitting. In this case the training set is well embedded into the model. The error of training and prediction is extremely low, however when new value is added to the model for prediction the model fails considerably. The network thus does not generalize well. This leads to an extremely high R-squared value, due to almost 0 error, however, a very high RMSE and low R-squared for when used in predicting future values.

It has been found from literature that network generalizations, or overfitting can be prevented by choosing an algorithm that is just large enough to provide the best fit. An extremely large network leads to more complexity and increase in power to overfit [50]. This model makes use of Bayesian automated regularization

method to prevent overfitting of data [50] [51]. In this method, the weights are first chosen at random with specified distribution. The network parameters, (weights and biases) are related to the unknow variances related to the distributions and can be found using statistical methods [50]. An explanation of this method along with Levenberg-Marquardt training, can be found in Foresee's paper [51]. In MATLAB, this algorithm is embedded in the function '*trainbr*' [52]. Using this function, the number of network parameters can be judged efficiently, giving an idea of the appropriate size of network needed.

9.2.3 Data selection

The model uses the exact same data as used during the regression analysis. As pointed out, the static model shall be utilizing the Levenberg-Marquart (LM) algorithm to solve a linear system of equations, in this case the thermal energy balance equation (Equations 8.1-8.4). The important *parameters* of data needed are;

1. *Temperatures of indoor air, walls/envelope and floor*
2. *Outdoor Climatic data (solar radiation, wind speed and outdoor air temperature*
3. *Use and operation (Time of use, ventilation profiles etc.)*

One important aspect related to ANN is to normalize the input and output parameters before being used in the model. The numerical range of the input parameters are highly varied. It is therefore advised to normalize the input and output variables to prevent any sort of severe numerical roundoffs which might affect the overall training efficiencies [53]. Also by doing this, the variables have zero mean and unity standard deviation [50]. The model standardizes the values of the inputs and the predictor variable between [-1 to 1]. This was done by using a function called mapminimax and for normalizing and de-normalizing the data [54].

The toolbox carries out three experiments under the static ANN model, in order to predict the net thermal energy demand at time (t). Similar to the linear regression models from the previous chapter, a set of four-time periods were chosen. The first being a yearly time frame. The model was trained over the data from 2015. This is followed by a monthly time frame, wherein the training is performed over two months to be able to predict data over the next month. Further the data trains over a period of two weeks to predict over the following week and finally, a day timeperiod is used to predict the next day.
Ahead the results obtained over monthly, and weekly plots shall be discussed along with the variation in input variables that were used.

9.3 Results and Discussions

9.3.1 Results using Real-time data over ANN models

The backpropagation network model was trained and built for predicting the net thermal demand of the room. The model operational parameters were set to stop training once its MSE has reduced below 10^{-3} or if the training has occurred for 500 epochs (iterations). Based on the RMSE and the R-squared value, the models' effectiveness can be judged. A figure showing the screenshot of the network in operation has been placed in the appendix A8 from MATLAB.

This section shall showcase the results of the ANN model and make a comparative study against the results of the Linear Regression models. For this, the same months of data were chosen for training and prediction during the monthly, weekly and day-wise time periods.

The model developed over the yearly data was much better fitted than the fit experienced in MLR. This model has an R^2 value of 75% with a RMSE of 334, compared to the R^2 value of 64% and RMSE of 397Wh under the MLR models. The image for ANN fit (figure 34) is shown below.

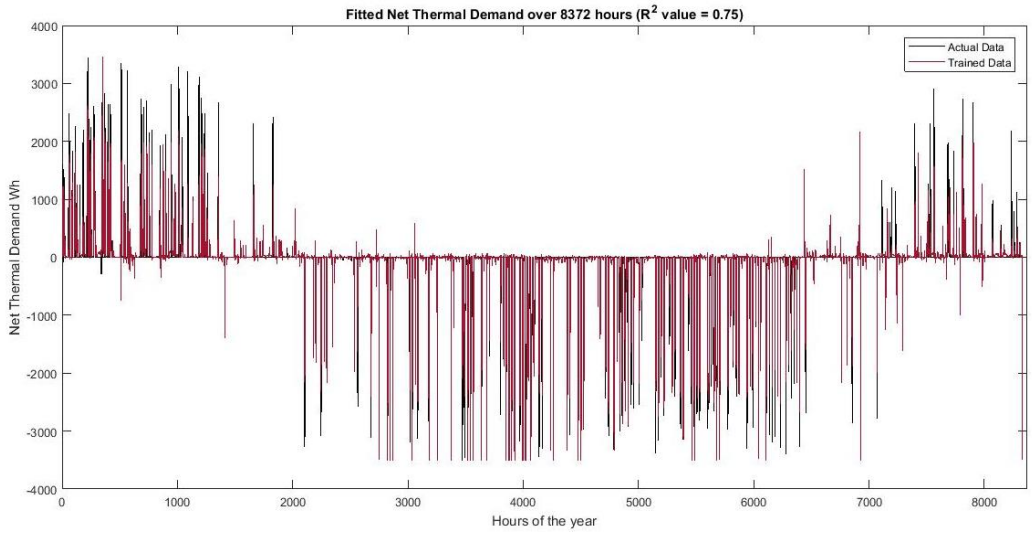
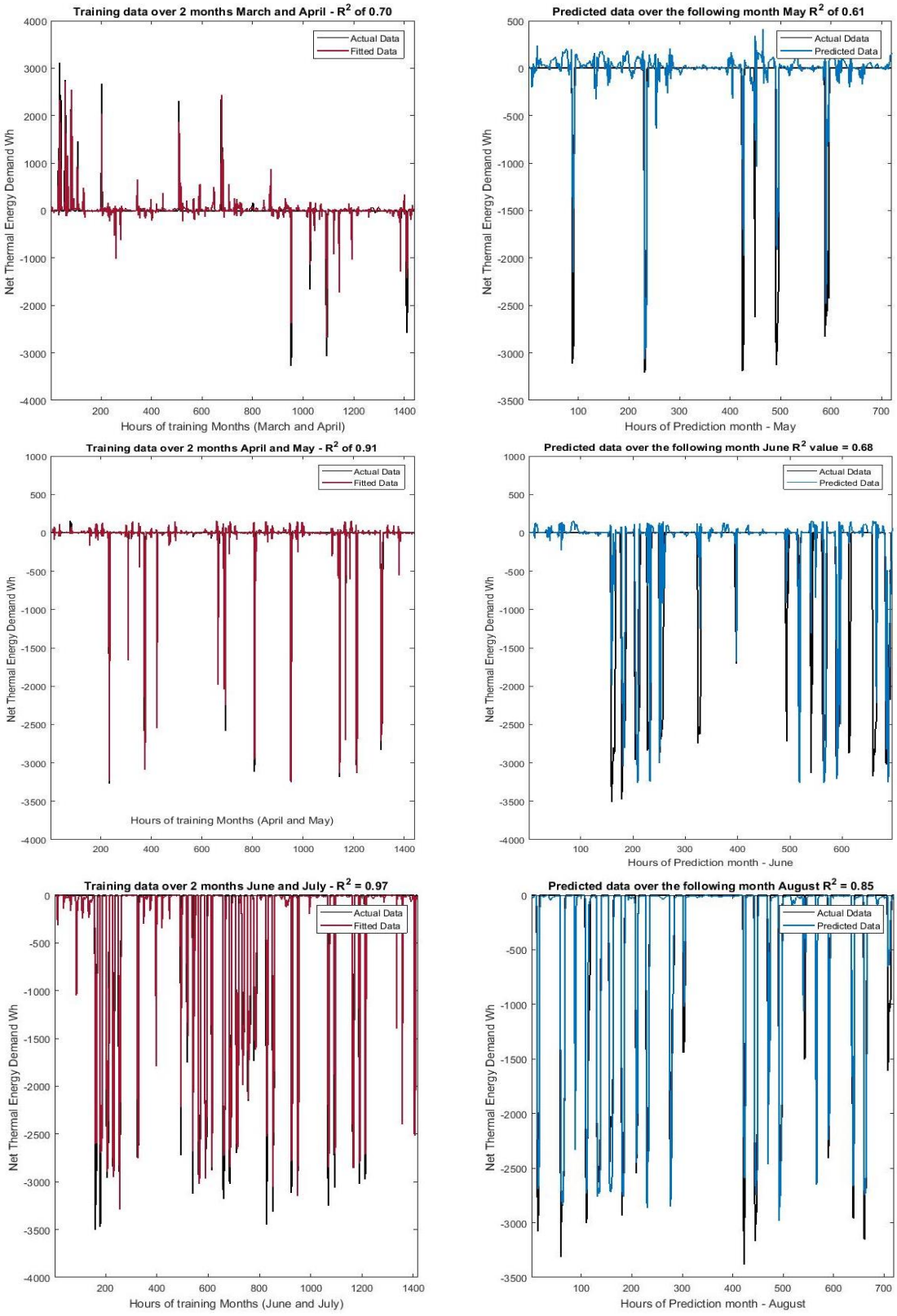


Figure 34 Estimated model fit over the entire year's data of 8372 hours using ANN.

9.3.2 Monthly Training and Prediction

The monthly time period models are high in efficiency with an R^2 value of up to 95%. The same monthly time periods have been used under ANN to compare the results with those of the linear regression models. The results have been validated using the R-squared and RMSE values obtained from the fits. Table 20 shows these results.



Figures 35 a, b and c Training and data prediction of the monthly datasets used on the ANN model. These three graphs showcase the three different data sets considered during research and have varied R^2 values.

Inferences

1. Similar to the fitting and predictive capabilities of the MLR models, the ANN models seem to fit and predict better with lesser discontinuity in data.
2. The ANN models have lesser noise around the 0-thermal energy mark, as compared to the MLR models.

Table 20 RMSE and R-squared values of the ANN and MLR models compared over monthly sub-datasets.

Training Months	Adj. R^2 MLR	RMSE MLR [Wh]	Adj. R^2 ANN	RMSE ANN [Wh]	Prediction Month	Adj. R^2 MLR	RMSE MLR [Wh]	Adj. R^2 ANN	RMSE ANN [Wh]
March & April	32%	298.60	70%	197.99	May	49%	371.41	61%	325.52
April & May	57%	283.38	91%	194.63	June	50%	614.41	68%	390
June & July	84%	331.36	97%	253	August	82%	342.62	85.6%	309

Week-wise training and Prediction

Similar to MLR models shown in chapter 8, two sub-datasets were chosen from the summer period (weeks 24 and 25) and from the winter period (week 3) for analyzing the ANN potential over weekly time periods. There was a higher R^2 value obtained for the ANN as compared to the MLR models.

Table 21 RMSE and R-squared values of the ANN and MLR models compared over weekly-sub-datasets.

Training Weeks	Adj. R^2 MLR	RMSE MLR [Wh]	Adj. R^2 ANN	RMSE ANN [Wh]	Prediction Month	Adj. R^2 MLR	RMSE MLR [Wh]	Adj. R^2 ANN	RMSE ANN [Wh]
3rd week 2015	82%	446.29	95%	229	4th week	80%	353.54	87%	321
24th -25th Week 2015	88%	320.39	97.1%	154	26th week	78%	423.98	81.2%	486

The figures 36 a and b below show the two plots of the ANN model training over the two sub-datasets.

The data obtained for the sub-datasets over a 24-hour period (again similar to the days chosen in the MLR models) have been placed in the appendix, A8.2, figure 60.

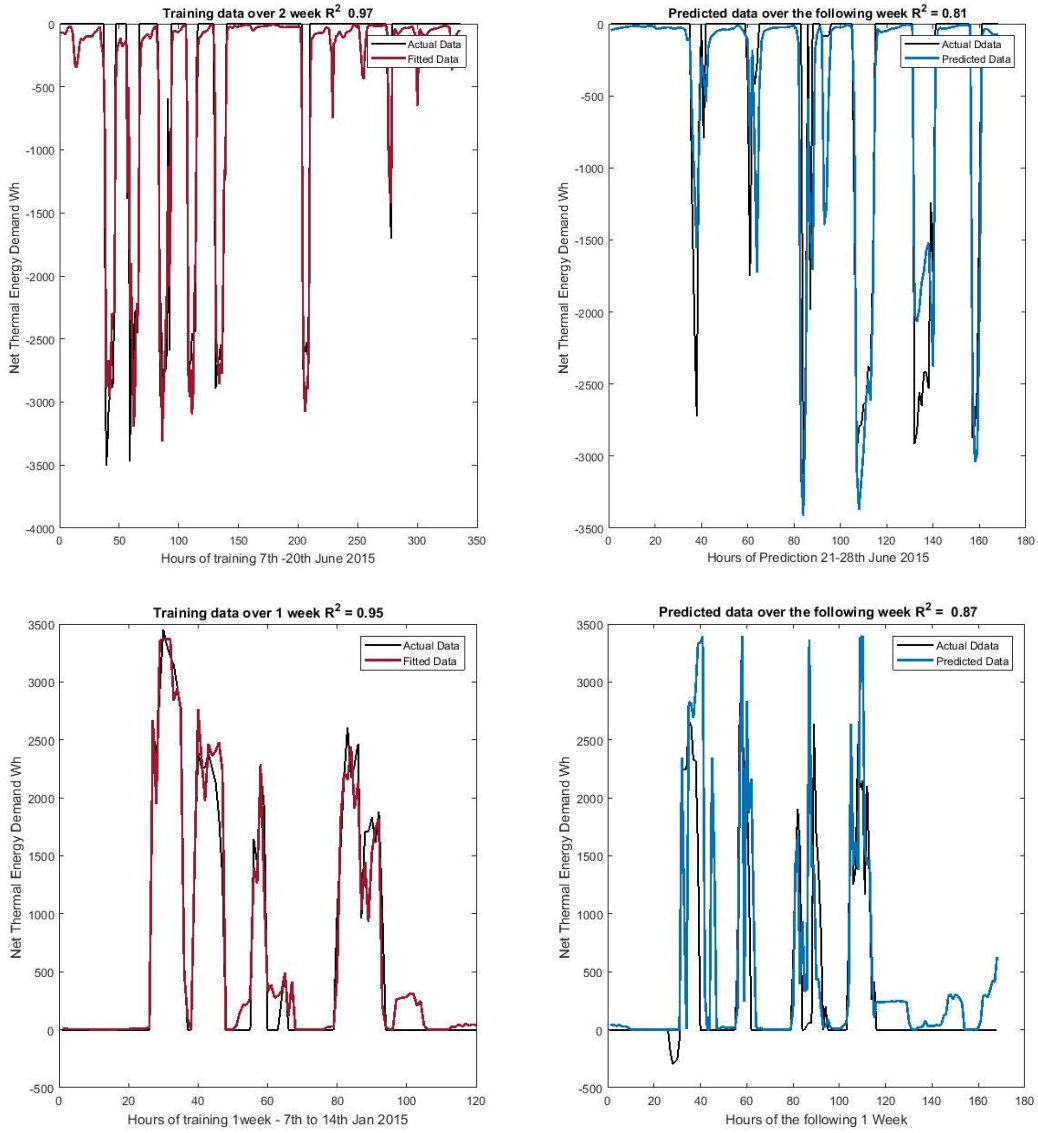


Figure 36 a and b Fitted and predicted values of the ANN models over weekly timestamps. The fit and prediction are extremely good, in comparison to the MLR models.

Inferences

1. ***With regards to the chosen period –*** It is seen that for month ahead or week ahead predictions, the ANN network fits extremely efficiently. Similar to the MLR models, a reduced period, leads to a better training and prediction of the model. However, the models must belong to a range of similar data, with either thermal heating or cooling, to prevent unnecessary increments in model error (figure 35 a).
2. ***With regards to discrepancy in target variable data –*** When both heating and cooling hours are placed over a training period, there is a certain degree of noise generated in the fit and prediction.

Thus, ANN networks are most effective in prediction when there are lesser variations in the type of thermal demand.

9.4 Comparative Analysis – MLR and ANN models

This chapter involves a research on the use of Artificial Neural Networks on prediction of thermal demand of a school room. This section shall focus on concluding the results obtained from ANN networks and compare them to the results of the MLR models.

Table 22 is a descriptive comparison table between the MLR and ANN training models based on the findings of this research.

Category	MLR	ANN
Flexibility in Training Models	*MLR models although train well with shorter timeframes, they are not able to address the discontinuity in predictor variables and their non-linear relationships with the target variable.*	*ANN models can address the continuity better and being more complex in nature, can address the existing non-linear relationships between the predictor and target variable.*
RMSE efficiency	*MLR training models witnessed the lowest RMSE at 283 to a maximum of 450 Wh and the lowest prediction RMSE at 342 up to 500 Wh*	*The training under ANN networks for the same months/weeks experienced RMSE values of a minimum of 154 to a max of 250Wh, whilst prediction models ranged from 250-250Wh.*
Adj. R^2 Efficiency	*The MLR models have R^2 values up to a mx of 88% in fitting the model, and 82% in prediction.*	*The ANN models can account for complexities in the input data, and are slight but significantly higher with 97% R^2 values over fitted models and 87% for prediction.*
Ease of Use	*Both Fitlm and Stepwise have been developed into automated algorithms, with easier to use interfaces for stepwise models.*	*Needs calibration of data, normalization and then de-normalization for plots. Difficult to make an appropriate network of nodes with the right model functioning parameters.*
Physical Significance of estimated coefficients and weights	*The estimated coefficients of the MLR model are dependent on the size of the data (timestep chosen) and the period of data chosen. However, these values are physically significant for a given model.*	*The ANN models calculate weights for each neuron, and this weight depends upon a lot of factors in addition to MLR. Thus, ANN models are able to account for non-linear nature in the data up to a certain degree. Also, the weights obtained are for each neuron, and not assigned to any specific input. Thu,s these weights are*

		not physically significant as the coefficients estimated in MLR models.
Based on the objective of the models	*The MLR models are reliable in bringing out the underlying relationships between various parameters and a target variable.*	*The ANN models are more reliable for prediction purposes, and providing high accuracy while doing so.*

To conclude, ANN are a fitting model concept in predicting thermal energy demands of a room. They exhibit a high efficiency in training and prediction of data, and can be used for practical purposes in determining month ahead or week-ahead thermal demands. However, as mentioned in the last point of the table the parameters on which this thermal demand can be tuned (like T_{AHU} or T_{floor}) are hidden, which makes the use of ANN difficult for Model Predictive Control.

The models were not developed based on past-year data as that was beyond the scope of this research. It was limited to understanding the simplistic use of ANN on predictive models and comparing them with MLR techniques. In turn, the ANN could answer for the error gap seen during MLR predictions.

10. Research Conclusions

10.1 Research Highlights

- ✓ An automated graphical tool for analyzing and visualizing data based on room level sensors of a school building
- ✓ Multiple and Partial Correlation Coefficients is a reliable methodology to gather quantitative and qualitative information on the most influential parameters affecting a room's thermal energy demand.
- ✓ Multilinear Regression Models (MLRs) are an efficient method in training and predicting datasets of monthly, weekly, and daily time-steps.
- ✓ MLR models are much better at defining the underlying relationship
- ✓ Using Stepwise Fit function is much more automated, simple, and efficient (in some models) than the FITLM function of MATLAB.
- ✓ A dynamic automated tool can be used for prediction of data based on forecasting of weather, occupancy schedules, and simulated data of the temperatures of a room.
- ✓ Using Past year data is not the most optimal type of data for prediction of thermal energy in the present year – however similar periods between two years might lead to better results.
- ✓ Artificial Neural Networks can be used to provide higher efficiency energy predictions and training than MLR networks. ANN can take into account the non-linear aspects of input variables and thus lead to higher efficiencies.
- ✓ All these step boxes described above, can be summed into one major toolbox for automated and effective room level analysis of a school building with the right sensor-based environment.

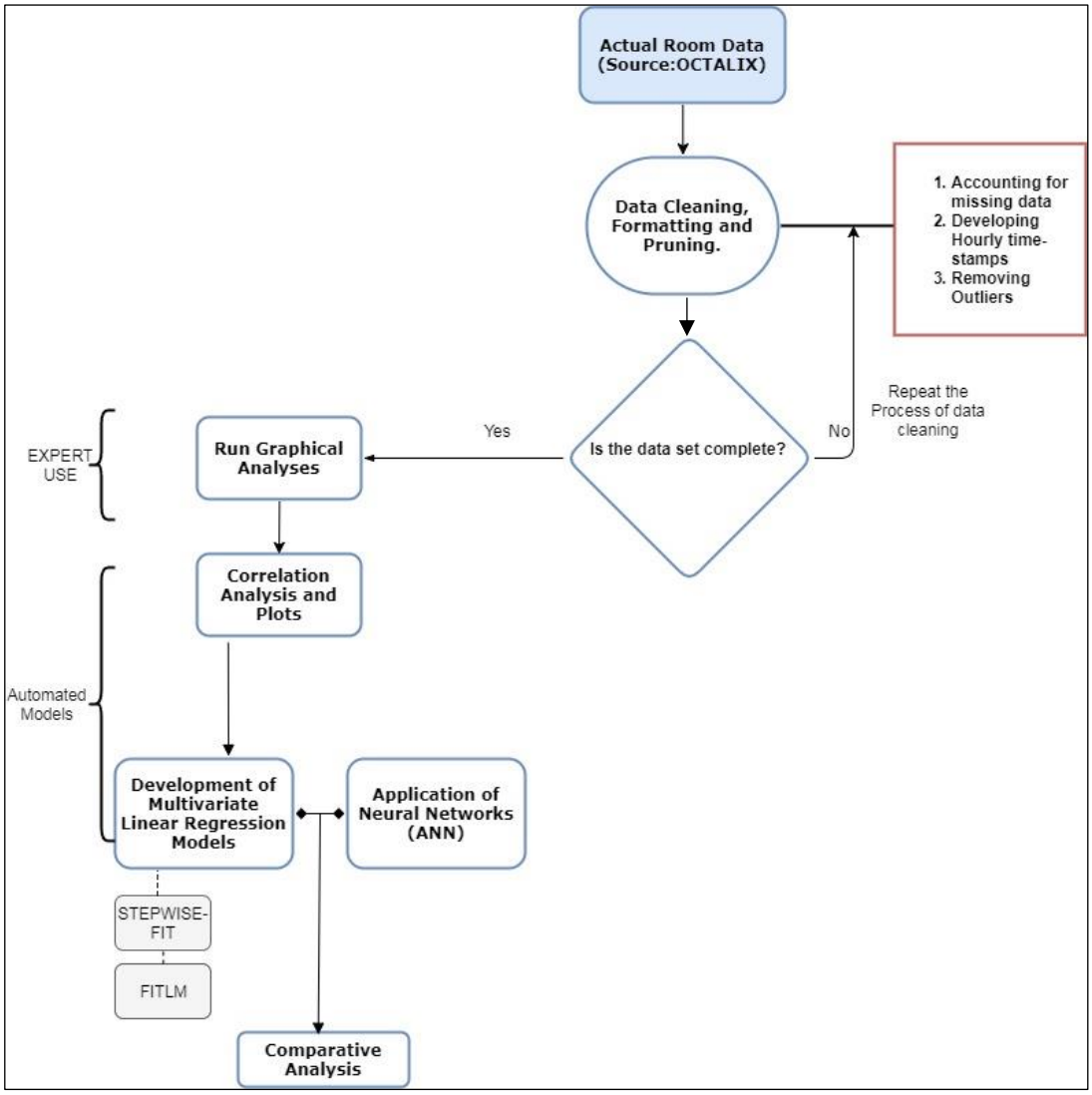


Figure 37 – The adopted methodology in a Flow chart representing the entire Toolbox.

10.2 Research Conclusions

Amongst this research work, there were several major and minor conclusions and results obtained. The investigations carried out during this research have been vital towards the development of an analysis and energy diagnostics automated tool. The use of 4 different step-toolboxes have been explained over the last 5 chapters. The use of each step-toolbox has been of practical importance for the research. The toolbox itself can be used in practice.

This chapter shall present the most important conclusions in coherence with the main research questions which have being addressed in section 4.2. The layout of this conclusion is scheduled chapter wise, addressing the important questions answered from each of them.

The graphical Analysis

In the *first phase* of the research, automated graphical analysis tools were developed to analyze a room's functioning with the help of sensor-based data. The objective of this research with regards to this phase was ***Objective b) Establishing a generic automated model to provide detailed graphics with regression analyses of the functioning of the room and the building.***

- ✓ ***With regards to the type of graphical analyses that help determine the functioning of a room in a school building*** **–** Several techniques were developed on different periods of data. This allows for monitoring the function of a building/room even in the absence of a large dataset. Graphical analysis of a room gives a quantitative measure of the varying effects of outdoor climate on the indoor climate. Analysis of the ventilation and airflow temperature explains the functioning of the HVAC system, and shows some faults in the HVAC or in the sensor system.

- ✓ ***With regards to types and periods of data and their need for building analyses*** **–** It was seen that segregating the data into working and non-working hours, helps to explain the functioning of the HVAC and also spot faults If any in the system. It is important to see how the building functions with and without occupancy as occupancy related energy demands play a big role in commercial buildings.

By using smaller periods of data, such as monthly plots, the functioning of a room with regards to presence, and ventilation can be explained more in detail. The case study helps in drawing conclusions that this building room is very well controlled, except that there is room for improvement with regards to the supply air temperature during the cooling periods (see chapter 6.3)
The most important aspect witnessed was that almost every parameter is responsible in some degree towards the increment or decrement of energy demand, typically thermal energy.

Correlation Coefficients

The *second phase* of the research was to quantify the effect parameters have over the thermal demand of a room. The objective associated with this phase was,
Objective c) Providing a methodology involving the use of correlation coefficients to understand the most probable and effective parameters influencing the heating and cooling demands at room level. This helps determine the most prominent losses and gains of thermal energy in classrooms.

A new methodology has been developed in this research, which incorporates a Multiple and Partial correlation model, to determine the percentile effect of one variable on the thermal demand.

- ✓ ***With regards to the effectiveness of mathematical tools such as Multiple and Partial correlations in determining the affect different parameters***- Almost every parameter in a building or room is affected by several other parameters. A methodology to quantify this effect has given a broader opinion about a building and its properties. For example, insignificant values of wind speed correlation with thermal demand, suggests that the building is quite unaffected by infiltration losses. It is seen from this research that the Multiple and partial correlations explain the dependent nature of thermal demand on other factors. This section of the research also explained;
    - *Interdependency of parameters*
    - *Partial correlations*

Using Multiple correlations, the true underlying relationship between dependent and independent parameters was calculated. This was also the first step towards explaining the factors which determine the thermal balance equations in the Multilinear Regression Models.

- ✓ ***Concerning the most important parameters needed for such a toolbox*** -From this chapter, it was seen that not all parameters used have a significant effect on thermal demand – thus concluding that certain parameters are not required to be measured by the room or building, in order to make predictive models. In this case study, the wind speed seemed to have no significant contribution towards the multiple correlations or the Regression models ahead.

Multiple Regression Predictive Modelling

Objective d) Utilization of the established parameters to perform multivariate regression analyses. These could be used in turn to train models for prediction of thermal energy demand.

This research has focused on determining the predictive capabilities of MLR functions on real-time room level data. The purpose of this research was to develop suitable models over distinct sub-datasets of year,month, week and day and understand the pros and cons of choosing different sub-datasets, towards training and prediction of thermal energy demand of a room.

- ✓ ***With regards to the overall efficiency of MLR –*** Linear Regression models, running on the thermal energy balance principle equation were exhibiting good training and predictive responses. However, due to discontinuity in datasets, it was seen that the predictive ability of the model would drop.

- ✓ ***With reg4ads to major difference between backwards (FITLM) and forward(STEPWISEFIT) propagation algorithms in Multilinear Regression Models -*** It was also found that stepwise fit, (a forward propagation algorithm function) is an efficient and faster function on MATLAB, for MLR training of models that the backwards propagation algorithm based FITLM function. The stepwise fit function also provides a user interface, to eliminate certain parameters, and understand the varying R^2 and RMSE values during operation.

- ✓ ***With regards to the discontinuity and the complex nature of data -*** Knowing that a room has a complex network of thermal energy balance, and is always altering based on several factors, the predictive power of the model would decrease. The MLR models can best account for the linear components of the input variables. When the input parameters are highly non-linear to the dependent parameter, or in the presence discontinuity in data, the model begins to crumble in efficiency as it begins to adopt towards more error in training and thus higher errors in predictions. It was also seen that heating and cooling periods of the year cannot be simultaneously trained by using MLR models,

- ✓ ***With regards to prediction using past year data as input variables for predicting in the present year*** – It was seen that the past years data are not the most optimum methodology to deliver predictive results regarding the present year. The outdoor climatic conditions, and the occupancy profiles tend to change, thus making it difficult to predict data at the room level for a school building. Having said this, the past year data might work for certain office rooms, where the occupancy profile does not change much, and the room has no direct contact with the outdoor climate (room towards the inside of the building). Also, since the results only ventured upon a particular group of weeks, it might be better to use the model on similar period between the two years of data, to give more accurate results.

- ✓ ***Estimation of Input variables*** – The tight relationships between T_{indoor}, $T_{out,}$ $T_{wall,}$ etc., can help estimate the room data beforehand for prediction purposes. This way a Model Predictive Control (MPC) could be developed which would be able to estimate T_{floor} and T_{AHU}, which are two major sources of thermal energy supply to the room, and are controllable with the parameters which can be actuated by the HVAC. Thus, in a real way the energy demand can be optimized, and the peaks reduced.

- ✓ ***With regards to the significance of estimated constants*** – the MLR models though not as efficient in predicting as the ANN models, were able to provide significant relationships between the various input parameters and the dependent parameter. Also with efficient training, the model is able to show significant estimations of the constants of the thermal balance equation.

Artificial Neural Networks

This phase dealt with the last objective of the research

Objective e) Utilization of ANN techniques to improvise on the prediction demand of energy.

The use of ANN models on real-time room level data was performed to investigate their use and efficiencies in prediction of thermal demand at the room level. The investigations yielded the following results

- ✓ ***With regards to the efficiency of ANN in training over complex sub-datasets -*** ANNs are much better at training over complex, sub-datasets. However, they do have the tendency to fall into the pit of overfitting. This leads to high training efficiencies but poorer predictive capability. Accounting for overfitting lead to certain reductions in overall model efficiencies however, also gave much better results in terms of predictions.

It was seen that the model works better but not the best in predicting for months with both heating and cooling demand.

- ✓ ***Regards to the overall use of ANN in this toolbox and for predictive model development for a room's thermal energy demand –*** By using the ANN it was found that the training of sub-datasets is much better as compared to the MLR models. Although, the ANN models have predictive prowess, they lack the ability to provide significance the coefficients, or any linear equation.

11. Future Research and Recommendations

This research was a combination of several mathematical and statistical operations towards optimizing energy performance at the room level of a school building, The HHS – Delft. The resulting objective was an automated Toolbox, inclusive of several other step-toolboxes, each with a unique operation.

The research found several methods of improvement, some were undertaken whilst others have been left for future purposes. The major recommendations for this research are pointed out below.

1. With regards to the Graphical Analysis

a) Using Working and Non-working hours of the rooms – Although this building functioned with working and non-working hours of the entire building, the model might be able to show some graphical results of interesting nature, if the hours of presence and absence from the rooms, were considered as working and non-working hours.

b) Developing Pattern Recognition for Fault Analysis **–**The graphical Analysis step-toolbox, could also include pattern recognition algorithms, such as The Gaussian Cluster, to find certain important patterns yet not visible regression graphs developed in this step-toolbox.

2. With regards to the Multiple Regression Predictive Models

a) ***With regards to predicting input variables for practical applicability*** – It was seen during the MLR step-toolbox, that there are methods in which certain input parameters can be estimated for predicting thermal demand. For example, the indoor air temperature could be a set point temperature. Seen a close relationship between the indoor air temperature and wall temperature, the latter could be estimated with high efficiency. The occupancy profiles can be known beforehand (schedules for the classroom) with which ventilation flow rates can be estimated. With the help of the weather forecasts a prediction can be made.

b) ***With regards to initial MLR analysis*** **–** It was also found that an initial MRL model can be developed over the entire year. This would help in segregating the data into heating, cooling and noon-heating-cooling hours. The MLR models could then be trained solely on the heating and cooling hours to understand their predictive capabilities. This would also prevent training or predicting data over heating and cooling hours simultaneously.

c) ***With regards to optimizing the physical parameters estimated by the MLR models*** **–** the coefficients estimated by the MLR models tend to vary with variations in the period of data chosen However, these parameters, with the help of iterations, converge to a constant for certain specific periods of the year. Once estimated, the constants might be efficient when applied on data from similar rooms (in size and functions), for predictive purposes

3. With regards to the ANN models

a) ***With regards to using different functions*** **-** The ANN model developed was a simple on over the available room level dataset. IT is highly recommended to fit models with other algorithms, and activation functions with the help of MATLAB or other programming languages.

b***) With regards to the estimated weights*** – it is recommended to use the weights calculated by the ANN models, and link it to the input variables of the model. Also changing these weight is a possibility on the ANN toolbox of MATLAB, Perhaps, doing this improves the predictive capacity of the model.

.

BIBLIOGRAPHY

1. International Energy Agency. Transition to Sustainable Buildings - Strategies and Opportunities to 2050. *OECD/IEA* **1,** 37 (2013).
2. Jos G.J. Olivier , Greet Janssens-Maenhout , Marilena Muntean, J. A. H. W. P. Trends in global CO 2 emissions 2015. *PBL Netherlands Environ. Assess. Agency* 12 (2015).
3. Magpantay, P. *et al.* Energy Monitoring in Smart Buildings Using Wireless Sensor Networks. 78–81 (2014).
4. Pohl, W. *Energy Performance of Buildings Directive*. *Energy Performance of Buildings* (2016). doi:10.1007/978-3-319-20831-2
5. Rijksoverheid. Plan of Action Energy Saving in Built Environment. Pg 8 (2011).
6. Opstelten, I., Bakker, E., Kester, J., Borsboom, W. & Elkhuizen, B. Bringing an energy neutral built environment in the Netherlands under control.
7. Doukas, H., Patlitzianas, K. D., Iatropoulos, K. & Psarras, J. Intelligent building energy management system using rule sets. *Build. Environ.* **42,** 3562–3569 (2007).
8. Taal, A.; Itard, L.; Zeiler, W. . Z. Automatic detection and diagnosis of faults in sensors used in EMS. *(CLIMA 2016)* 1–10 (2016).
9. Williams, K. T. & Gomez, J. D. Predicting future monthly residential energy consumption using building characteristics and climate data: A statistical learning approach. *Energy Build.* **128,** 1–11 (2016).
10. Cohen, J. & Cohen, P. *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences Third Edition*. (Routledge 2002-08-01, 1983).
11. Sa, J. P. M. de. *Applied Statistics - Using SPSS, STATSISTICA, MATLAB and R.* (SPRINGER, 2007).
12. Sandels, C. *Modeling and Simulation of Electricity Consumption Profiles in the Northern European Building Stock.* (2016).
13. Yuan, L., Ruan, Y., Yang, G., Feng, F. & Li, Z. Analysis of Factors Influencing the Energy Consumption of Government Office Buildings in Qingdao. *Energy Procedia* **104,** 263–268 (2016).
14. Zou, K. H., Tuncali, K. & Silverman, S. G. Correlation and Simple Linear Regression. *Radiology* 617–622 (2003).
15. Lopez, C. J. Data-driven Predictive Control for Heating Demand in Buildings. (Technical University of Delft, 2017).
16. Dong, B., Dong, B., Lee, S. E. & Sapar, M. H. A holistic utility bill analysis method for baselining whole commercial building energy consumption in Singapore A holistic utility bill analysis method

for baselining whole commercial building energy consumption in Singapore. (2005). doi:10.1016/j.enbuild.2004.06.011

17. Wei, X., Li, N. & Zhang, W. Statistical Analyses of Energy Consumption Data in Urban Office Buildings of Changsha , China. *Procedia Eng.* **121,** 1158–1163 (2015).
18. Kramer, R., van Schijndel, J. & Schellen, H. Simplified thermal and hygric building models: A literature review. *Front. Archit. Res.* **1,** 318–325 (2012).
19. Rao, K. R. & Lakshminarayanan, S. Partial correlation based variable selection approach for multivariate data classification methods. **86,** 68–81 (2007).
20. Široký, J., Oldewurtel, F., Cigler, J. & Prívara, S. Experimental analysis of model predictive control for an energy efficient building heating system. *Appl. Energy* **88,** 3079–3087 (2011).
21. Li, Z., Han, Y. & Xu, P. Methods for benchmarking building energy consumption against its past or intended performance : An overview. *Appl. Energy* **124,** 325–334 (2014).
22. G. Mustafaraj, G.Lowry, J.Chen. Prediction of room temperature and relative humidity by autoregressive linear and nonlinear neural network models for an ... (2011). doi:10.1016/j.enbuild.2011.02.007
23. Parab, V. Thermal Modelling of Existing Residential Buildings in North-Western Europe. *Msc. Thesis* (TU Delft, 2016).
24. Carbonari, A., Vaccarini, M. & Giretti, A. Bayesian Networks for Supporting Model Based Predictive Control of Smart Buildings. (2014).
25. Bacher, P. & Madsen, H. Identifying suitable models for the heat dynamics of buildings. *Energy Build.* **43,** 1511–1522 (2011).
26. Jiménez, M. J., Madsen, H. & Andersen, K. K. Identification of the main thermal characteristics of building components using MATLAB. *Build. Environ.* **43,** 170–180 (2008).
27. Trčka, Jan L.M. Hensen, M. Overview of HVAC system simulation. *Autom. Constr.* 93–99 (2010).
28. de, Nijs, J. M. Inverse modeling of buildings with floor heating and cooling systems for benchmarking operational energy use. (Eindhoven University of Technology, 2016).
29. Jiménez, M. J. & Madsen, H. Models for describing the thermal characteristics of building components. *Build. Environ.* **43,** 152–162 (2008).
30. MATLAB - MathWorks. Available at: https://nl.mathworks.com/products/matlab.html. (Accessed: 10th December 2017)
31. Jr, M. P. M. & Ã, G. A. B. Neurocomputing Long-term time series prediction with the NARX network : An empirical evaluation. **71,** 3335–3343 (2008).
32. Afram Abdul,Farrokh Janabi-Sharifia, Alan S. Funga, Kaamran Raahemifar. Artificial Neural Network (ANN) based Model Predictive Control (MPC) and Optimization of HVAC Systems : A State of the ... *Energy Build.* **141,** 96–113 (2017).

33. Neumann, C. & Jacob, D. *Results of the project Building EQ Tools and methods for linking EPBD and continuous commissioning*. *Solar Energy* (2010).
34. Rafsanjani, H. N., Ahn, C. R. & Alahmad, M. *A Review of Approaches for Sensing, Understanding, and Improving Occupancy-Related Energy-Use Behaviors in Commercial Buildings*. (2015). doi:10.3390/en81010996
35. Pe, L. A review on buildings energy consumption information ´. **40,** 394–398 (2008).
36. World Health Organisation. *Indoor Environment: Health Aspects of Air Quality, Thermal Environment, Light and Noise*. (1991).
37. Agentschap NL. *Gebouwmonitoring met Energieprofielen*. *Agentschap NL ENergyie en Klimaat* (2015).
38. Jones, P. & Salleh, E. Evidence base prioritisation of indoor comfort perceptions in Malaysian typical multi-storey hostels. *Build. Environ.* **44,** 2158–2165 (2009).
39. Least-Squares Fitting - MATLAB & Simulink - MathWorks Benelux. Available at: https://nl.mathworks.com/help/curvefit/least-squares-fitting.html. (Accessed: 10th January 2017)
40. Center for Disease Contorl and Prevention - NIOSH. Chapter 2. in *CDC - Factors Affecting Indoor Air Quality* 5–12
41. Create linear regression model - MATLAB fitlm - MathWorks Benelux. Available at: http://nl.mathworks.com/help/stats/fitlm.html. (Accessed: 5th January 2017)
42. Stepwise regression - MATLAB stepwisefit - MathWorks Benelux. Available at: https://nl.mathworks.com/help/stats/stepwisefit.html. (Accessed: 12th January 2017)
43. Xu, G. HVAC system study : a data-driven approach. (2012).
44. Stepwise regression - MATLAB stepwisefit - MathWorks Benelux. Available at: https://nl.mathworks.com/help/stats/stepwisefit.html. (Accessed: 5th January 2017)
45. Hawkins, D. M. The Problem of Overfitting. *J. Cheminstry Inf. Comput. Sci.* **44,** 1–12 (2004).
46. Standardized z-scores - MATLAB zscore - MathWorks Benelux. Available at: https://nl.mathworks.com/help/stats/zscore.html. (Accessed: 11th July 2017)
47. Datta and S. A. Tassou, D. Application of Neural Networks for the Prediction of the Energy Consumption in a Supermarket. *ASHRAE Trans. 99* 505–517 (1993).
48. Yang, Hugues Rivard, Radu Zmeureanu, J. BUILDING ENERGY PREDICTION WITH ADAPTIVE ARTIFICIAL NEURAL NETWORKS Department of Building , Civil and Envr . Engineering , Concordia University , Department of Construction Engineering , ETS , 1100 Notre-Dame Street West , COMPUTATIONAL EXPERIMENTS. in *Ninth International IBPSA Conference* 1401–1408 (2005).
49. Ekici, B. B. & Aksoy, U. T. Prediction of building energy consumption by using artificial neural networks. *Adv. Eng. Softw.* **40,** 356–362 (2013).

50. Demuth, H. *Neural Network Toolbox - User's Guide Version 4*. (The MathWorks, Inc, 2001).
51. Foresee, F. D. & Hagan, M. T. GAUSS-NEWTON APPROXIMATION TO BAYESIAN LEARNING ** School of Electrical and Computer Engineering. *Network* 1930–1935 (1930).
52. Bayesian regularization backpropagation - MATLAB trainbr - MathWorks Benelux. Available at: http://nl.mathworks.com/help/nnet/ref/trainbr.html. (Accessed: 6th May 2017)
53. Yang, J., Rivard, H. & Zmeureanu, R. On-line building energy prediction using adaptive artificial neural networks. **37,** 1250–1259 (2005).
54. Process matrices by mapping row minimum and maximum values to [-1 1] - MATLAB mapminmax - MathWorks Benelux. Available at: http://nl.mathworks.com/help/nnet/ref/mapminmax.html. (Accessed: 23rd May 2017)

Appendix

Appendix 1 – Images of the HHS and the case study room 1075.

A.1.1 – Floor Heating Mechanism of the BMES at the HHS

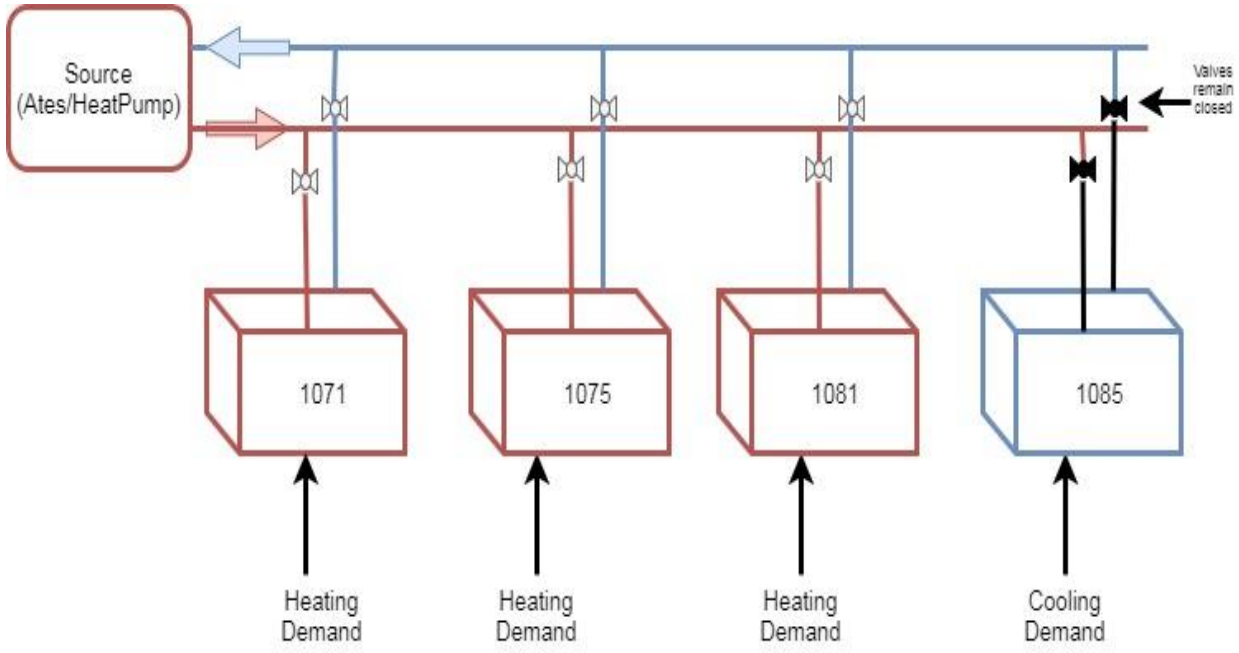


Figure 38 shows the mechanism of priority based heating or cooling system adopted in the HHS.

Here it can be seen that the room 1085, needs cooling. However, due to the other rooms having a heating demand, the room 1085, shall not be cooled. This would lead to a training deficiency in models (see section 8.5.2 – weekly datasets) wherein, even during the need for thermal cooling, the model is forced to train with no thermal demand. Having said this, it should be noted that since this is a school building and is heavily maintained in terms of the indoor climate being constant, this occurrence of such a scenario is not quite often.

A1.2 – Images of the Case room

Figures 39 a and b – Classroom 1075 and the adjoining corridor respectively. The Classroom has a south facing wall (with windows seen on the left of figure a).

Appendix 2 Correlation Coefficients and the literature associated with it.

The research focuses on multiple and partial correlation (explained ahead). Both techniques give a much broader understanding of the individual effect the parameters on the overall thermal energy demand. Correlation coefficient as mentioned is used to measure and interpret the strength of a linear or non-linear relationship between two continuous variables. This research focuses on the Pearson Correlation, and Partial Correlations. The formula adapted to compute the simple Pearson Correlation Coefficient 'r' for a given set of X_i and Y_i values (i = 1…n) where n is the number of samples is,

$$R = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\,\sum_{i=1}^{n}(y_i-\bar{y})^2}}\;; \qquad \text{Eq-A2.1}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the $x_i$and $y_i$values, respectively.

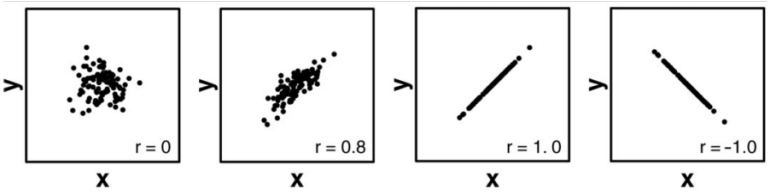


Figure 40 shows four scatterplots with the Pearson Correlation Coefficients (from left to right): r = 0 (uncorrelated data), r = 0.8 (strongly positively correlated), r = 1.0 (perfectly positively correlated), and r=1 (perfectly negatively correlated) [14].

An important test required for validating the correlation coefficient value is the statistical hypothesis tests. The p-value is the measure of the significance of the null hypothesis [14] [15]. The significant level generally chosen for the null hypothesis is 5%, or (p-value < 0.05). This means that there is a confidence level of 95% that the coefficient is a correct value for the correlation between the dependent and the independent variable. The *statistical Hypothesis Tests* for Correlation coefficients check for null hypothesis. This hypothesis states that "the underlying linear correlations has a hypothesized value, p_o." [15]. The alternative hypothesis is that the underlying value is greater or lesser than the p_o. Using a z-test the statistic value $(\text{p} - \text{value}) = (r - p_o)/s_r$ is calculated, where s_r is the standard error of the calculated value.

$$s_r = (1 - r^2)/\sqrt{n} \qquad \text{EqA.2.2}$$

where n is the sample size.

If this p-value is <0.05 the null hypothesis is rejected and the correlation coefficient is deemed significant [15] [11].

Table 23 shows the interpretation of correlation coefficients. The sign signifies the direction of the relationship. The absolute value is the indication of the strength [14].

Correlation Coefficient Value	Direction and Strength of Correlation
-1.0	Perfectly Negative
-0.8	Strongly Negative
-0.5	Moderately Negative
-0.2	Weakley Negative
0	No association
+0.2	Weakly Positive
+0.5	Moderately Positive
+0.8	Strongly Positive
+1.0	Perfectly Positive

Appendix 3 – Yearly Plots from Graphical Analysis

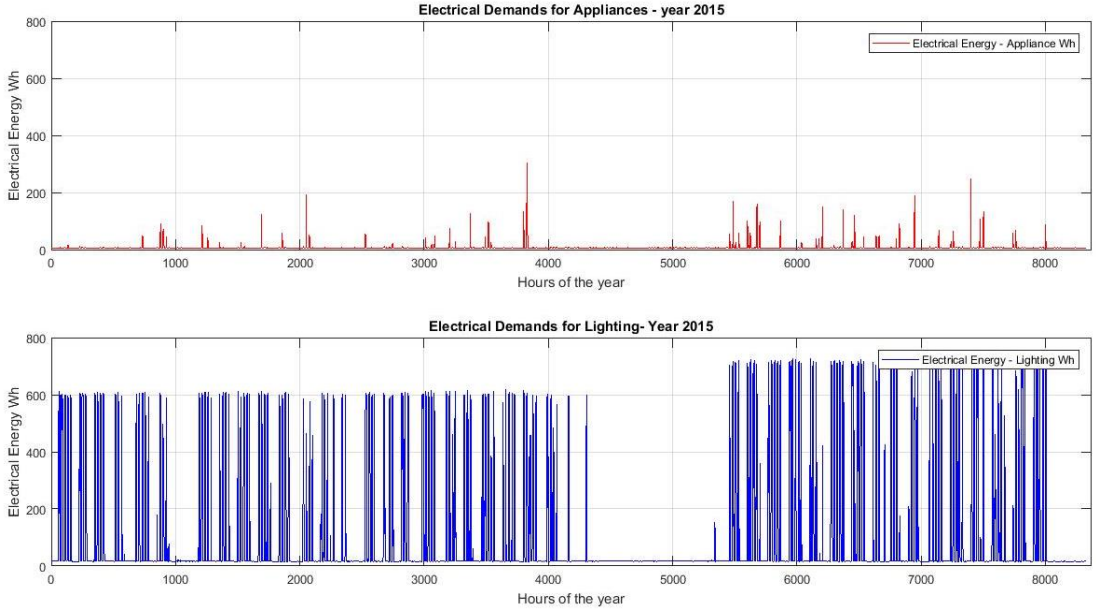


Figure 41 Graphical image of the electrical demands for the entire year. The electrical demand from lighting is the highest and almost constant, representing different weeks of the year. A gap in August (4500-5500 can be seen which represents the break period of the school.

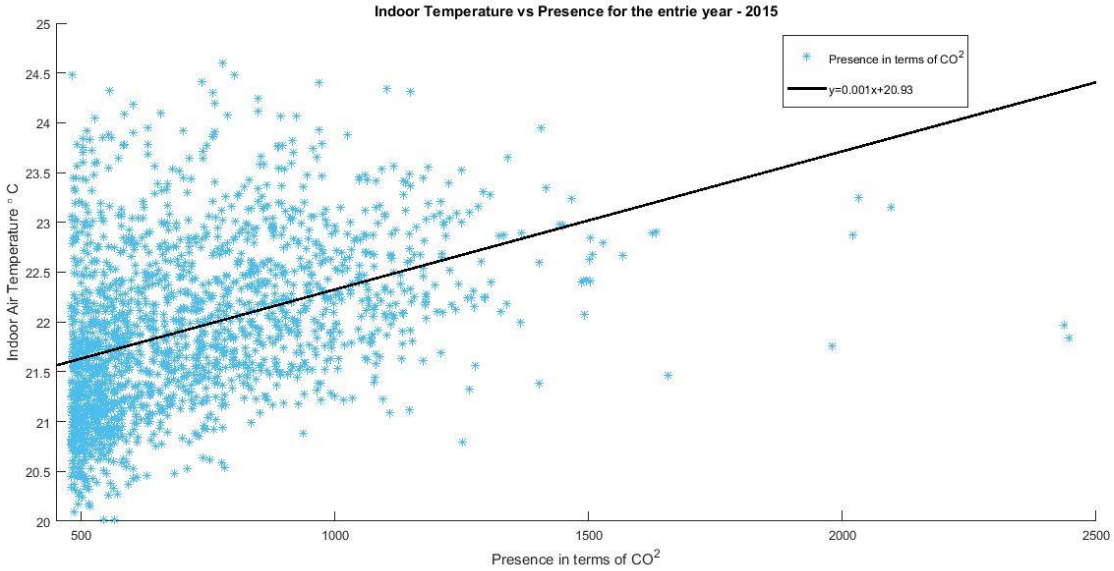


Figure 42 Sensitivity of the indoor air temperature with rising carbon dioxide. It is seen that the indoor air temperature does not vary much with increment in occupancy of students. This shows that the room is extremely well ventilated to nullify the heating effect due to occupant behavior.

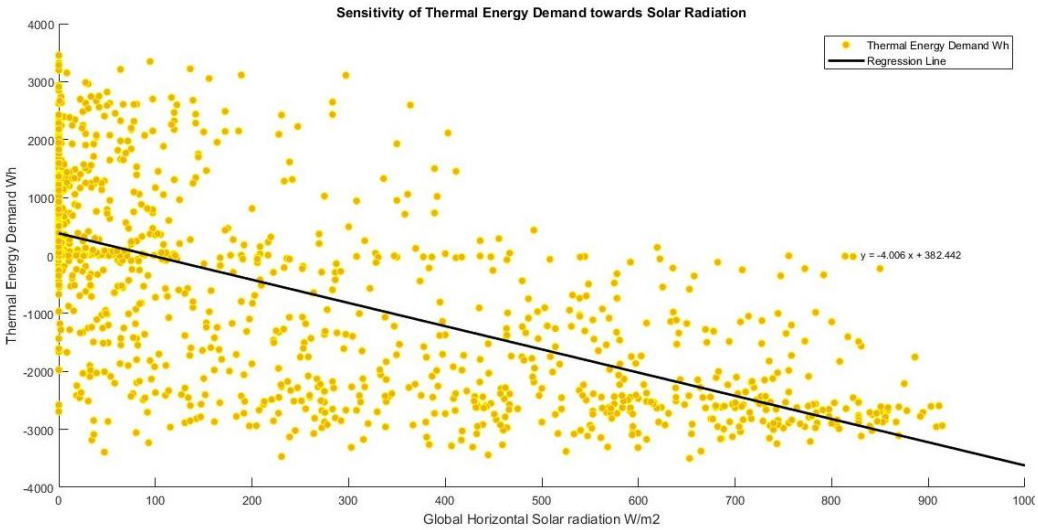


Figure 43 Sensitivity of Thermal Energy demand with the solar radiation.

Appendix 4 – Miscellaneous Figures from the Graphical Analysis

A.4.1 – Daily Plots

The graphical analysis was done on the 1st of October 2015, for discovering plots of similar kinds as seen regarding the months and the working and non-working hours. These have been presented below.

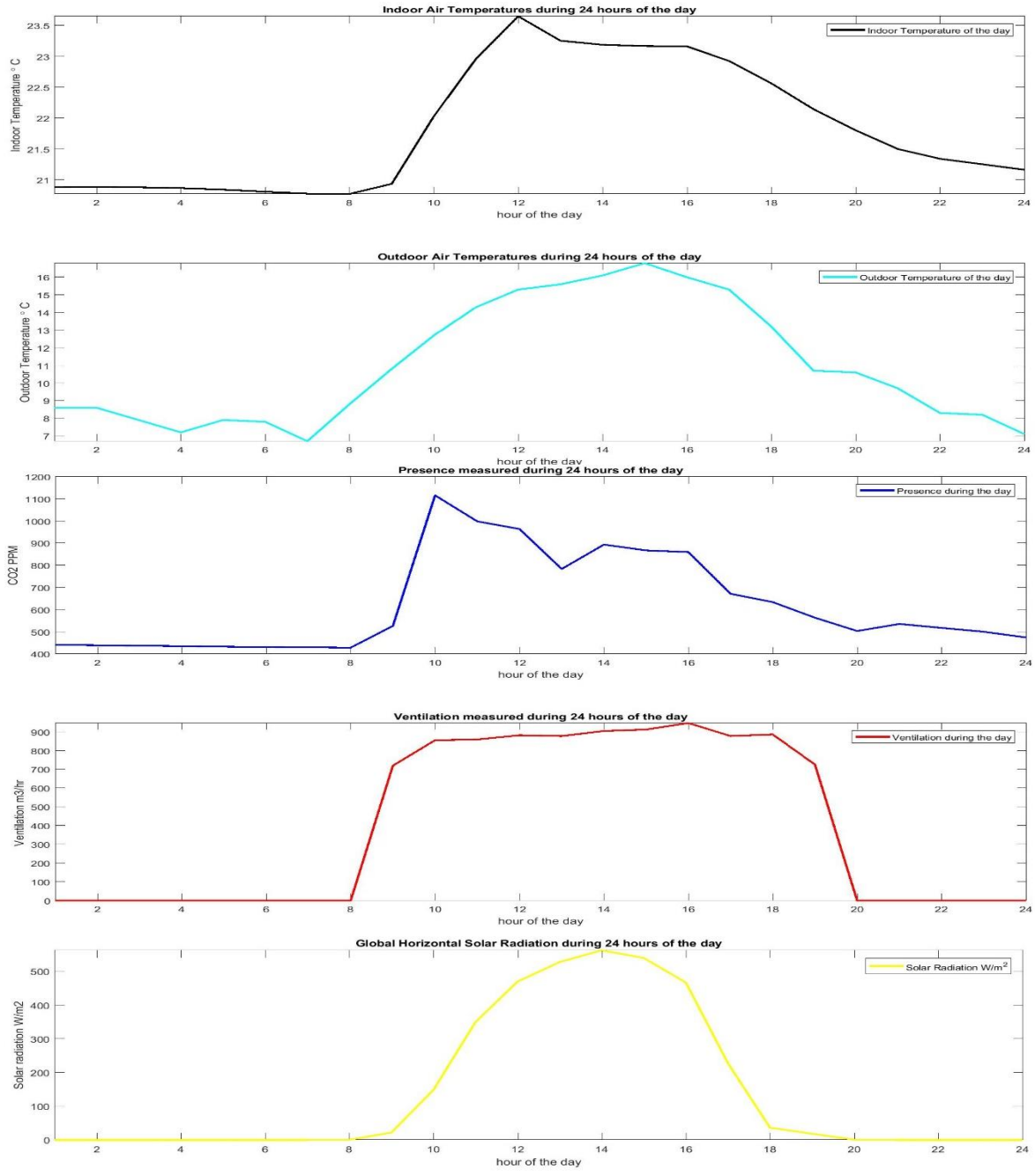


Figure 44 Graphical analysis plots of October 1st 2015.

A.4.2 – Seasonal Plots

The seasonal plots for spring and winter have been shown below.

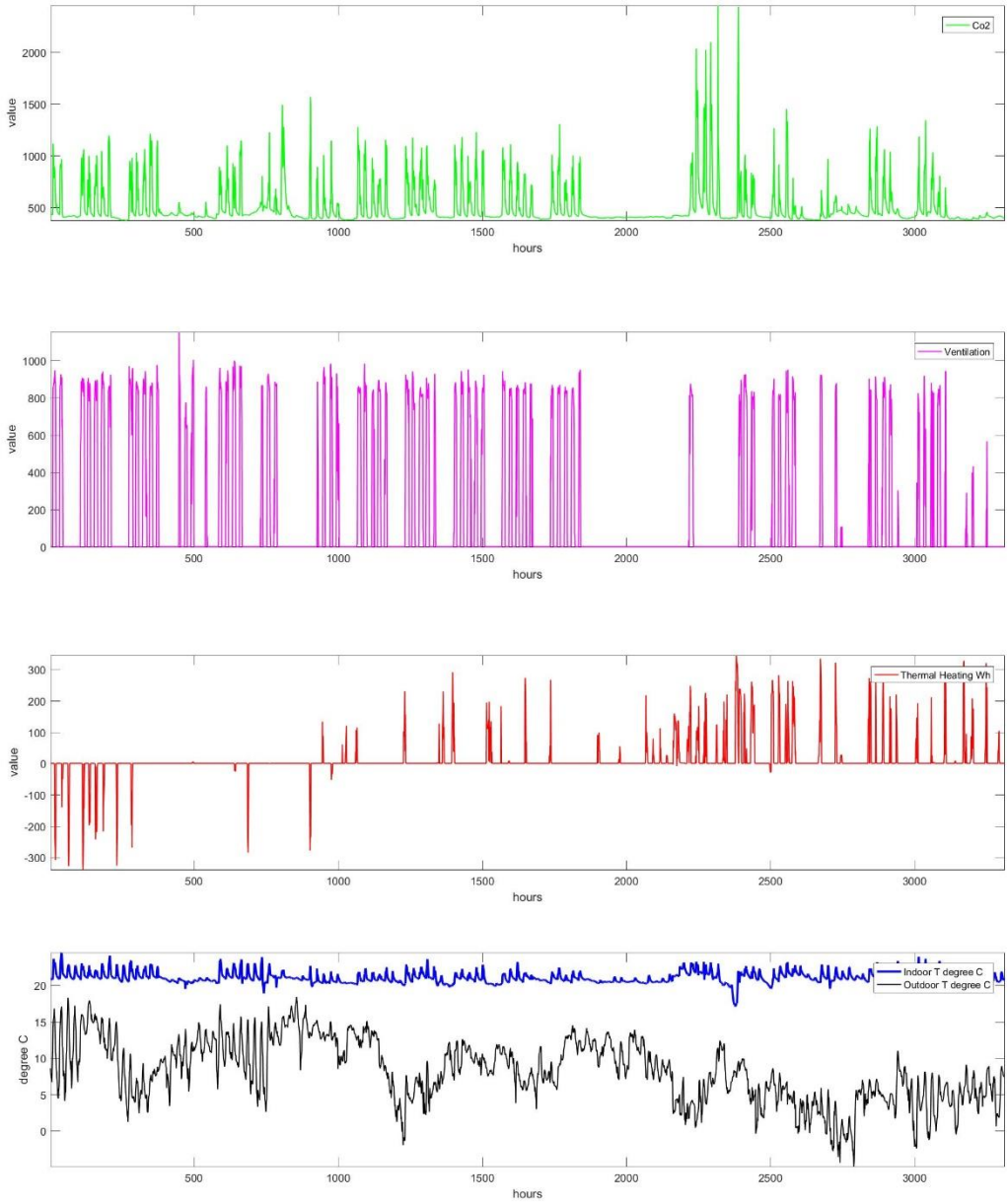


Figure 45 shows the various graphs explaining the readings of presence (CO2 PPM,) Ventilation air flow rate, and the thermal energy utilization of the class room during the *winter* period.

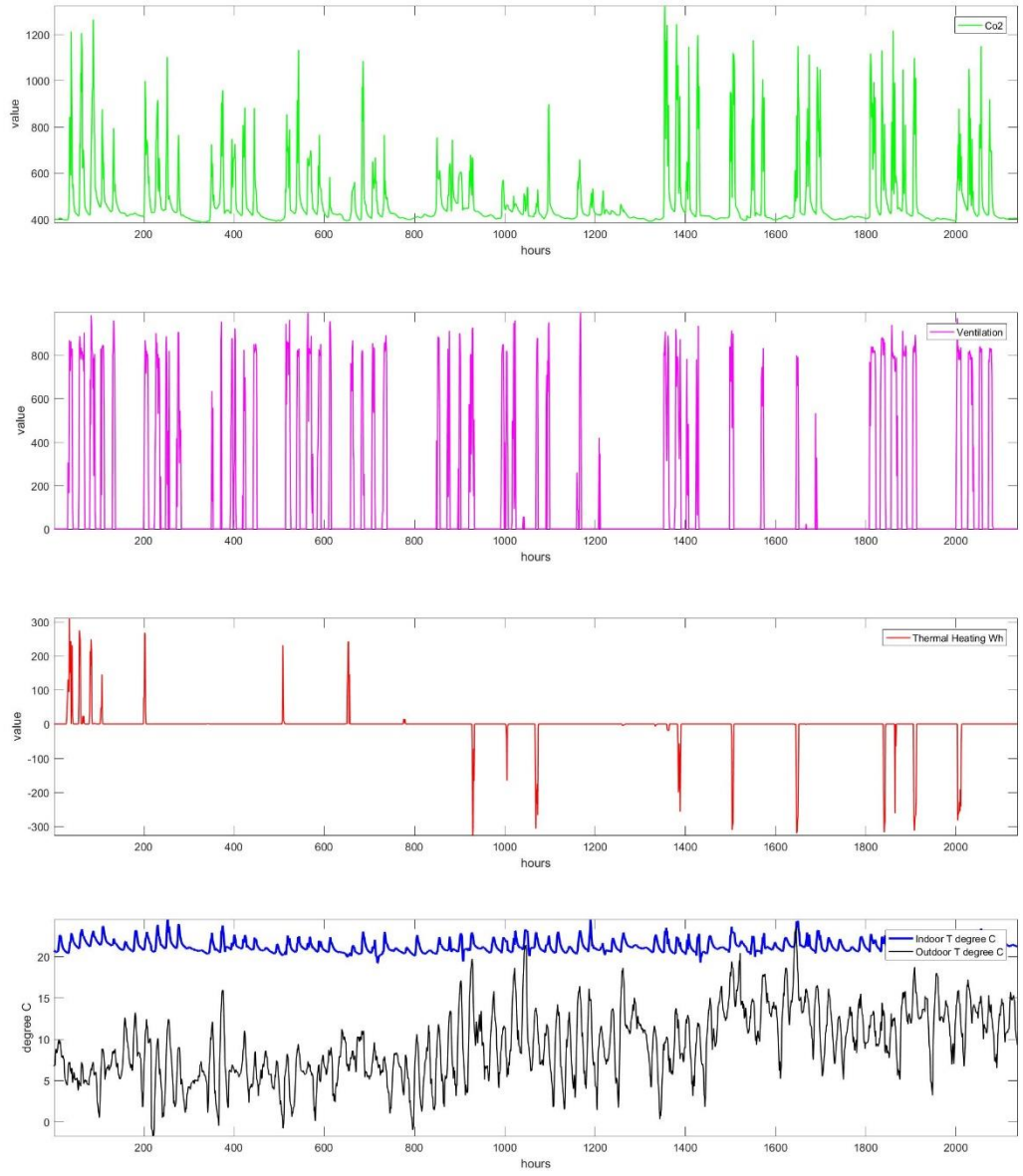


Figure 46 shows the various graphs explaining the readings of presence (CO_2 PPM,) Ventilation air flow rate, and the thermal energy utilization of the class room during the *spring* period

Appendix 5 – Correlation Coefficients

A.5.1 – Regarding Multiple and Partial Correlation Coefficients for only heating and only cooling Hours.

The model is automated for the user to choose from Net thermal energy demand (result shown in section 7.3), only heating hours and only cooling hours. The graphs and tabulated results for heating and cooling have been shown below.

Heating Only

Table 24 shows the obtained Multiple and Partial correlation coefficient (R-Value) and the P-values for the various parameters as stated when correlated against only Heating Hours.

Category	Internal Heat	Wall Temperature	Indoor Air Temperature	Floor Surface Temperature	Presence	Ventilation	Wind Speed	Outdoor Air Temperature	Solar Radiation
Multiple Correlation Coefficient	0.158	-0.050	-0.081	0.353	0.089	0.145	0.153	-0.168	-0.056
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Partial Correlation Coefficients	0.04	0.05	-0.07	0.29	-0.02	0.09	0.10	-0.01	0.03
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

The figure 47 shows the Multiple and Partial correlations of net thermal demand with the dynamic variables for a period of all hours with a heating demand. Naturally the most correlated is the floor temperature as it is the main source of thermal energy.

Cooling Hours only

Table 25 shows the obtained Multiple and Partial correlation coefficient (R-Value) and the P-values for the various parameters as stated when correlated against only Cooling Hours.

Category	Internal Heat	Wall Temperature	Indoor Air Temperature	Floor Surface Temperature	Presence	Ventilation	Wind Speed	Outdoor Air Temperature	Solar Radiation
Multiple Correlation Coefficient	0.215	0.393	0.432	-0.708	0.221	0.289	0.012	0.360	0.430
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Partial Correlation Coefficients	0.008	-0.055	0.145	-0.668	-0.002	0.099	0.103	-0.192	0.189
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

The figure 48 shows the Multiple and Partial correlations of net thermal demand with the dynamic variables for a period of all hours with a cooling demand. Naturally the most correlated is the floor temperature as it is the main source of thermal energy.

Appendix 6 – Sample Dataset

Below table xx shows a sample of dataset after organizing and cleaning that was used for the MLR and Neural Network models developed during this research. The data is for one day, 1.1.2015, and shows values over an hourly average.

Table 26 shows the values obtained from hourly averages of sensor recordings from the room 1075. The data belongs to 1.1.2015.

Timestamp	Appliance electrical average J	Lighting electrical average J	Heating average J	Cooling average J	Wall temperature average [C]	Air temperature average	Supply Air temperature average	Co2 average	Airflow average	Pipe1in average temperature	Pipe1out average temperature	Pipe2in average temperature	Pipe2out average temperature	Pipe3in average temperature	Pipe3out temperature average
1/1 0:00	18170	65780	81928	9426	20.866	21.372	22.515	422.9	0	22.5807	22.52665	22.6653	22.61275	22.53425	22.56565
1/1 1:00	18975	65665	0	128038	20.733	21.301	22.4919	423	0	22.30035	22.38935	22.3897	22.432	22.40915	22.5024
1/1 2:00	18860	64860	2705565	36015	20.649	21.221	22.4678	422.85	0	24.26565	22.81715	24.20015	22.93365	22.39275	22.4502
1/1 3:00	18975	64630	1794037	32578	20.589	21.162	22.4435	422.5	0	23.6402	22.78655	23.6305	22.87765	22.42235	22.4123
1/1 4:00	19205	65665	4425503	0	20.595	21.153	22.4269	424	0	25.80805	23.4152	25.875	23.50245	22.44985	22.42455
1/1 5:00	19205	64630	5724921	0	20.697	21.173	22.4149	423.75	0	27.11735	23.97555	27.16625	24.04405	22.81015	22.52615
1/1 6:00	19320	62905	5439934	0	20.785	21.184	22.40875	423.9	0	27.2334	24.1683	27.26425	24.25735	23.0259	22.6231
1/1 7:00	19320	65435	5468523	0	20.867	21.228	22.4111	424.95	0	27.1721	24.26025	27.19995	24.35365	23.0229	22.68695
1/1 8:00	19780	64917.5	5338584	0	20.895	21.241	22.4187	425.9	0	27.16505	24.3309	27.2552	24.42615	23.06055	22.711
1/1 9:00	19550	63537.5	4426713	0	20.928	21.26	22.4298	426.3	0	26.90085	24.3535	26.94465	24.4476	23.05305	22.74005

Timestamp	Appliance electrical average J	Lighting electrical average J	Heating average J	Cooling average J	Wall temperature average [C]	Air temperature average	Supply Air temperature average	Co2 average	Airflow average	Pipe1in average temperature	Pipe1out average temperature	Pipe2in average temperature	Pipe2out average temperature	Pipe3in average temperature	Pipe3out temperature average
1/1 10:00	19780	66470	3807074	0	20.963	21.293	22.4488	425.85	0	26.2062	24.1167	26.3966	24.18635	22.9416	22.7367
1/1 11:00	19665	64975	4834788	0	21.009	21.325	22.47035	426.45	0	27.09945	24.34565	27.16655	24.45035	22.9968	22.7551
1/1 12:00	19895	64975	4490496	0	21.131	21.384	22.5013	426.8	0	26.7839	24.40535	26.7691	24.5099	23.0516	22.796
1/1 13:00	19780	66585	2650348	0	21.161	21.452	22.5371	426.2	0	25.4743	23.9114	25.56975	24.0001	22.9788	22.8083
1/1 14:00	19665	66930	5844	298412	21.094	21.476	22.5657	425.9	0	22.97335	23.10255	23.0258	23.1952	22.87215	22.7744
1/1 15:00	19665	65550	1445406	235567	20.999	21.45	22.5826	425.45	0	23.37145	22.9837	23.5521	23.02555	22.74595	22.71625
1/1 16:00	19780	66700	1275388	209241	20.922	21.384	22.5852	423.975	0	24.21025	23.47755	24.33345	23.5148	22.77995	22.67975
1/1 17:00	19780	66930	656312	137602	20.81	21.333	22.58845	422.5	0	23.1546	22.92345	23.08675	22.9859	22.5238	22.6198
1/1 18:00	19550	65090	4797290	0	20.713	21.271	22.58	421.2	0	26.13055	23.646	26.13475	23.73205	22.51165	22.59185
1/1 19:00	20010	66815	4939354	0	20.779	21.257	22.5755	421.5	0	26.979	24.14725	26.98375	24.22845	22.9172	22.699
1/1 20:00	20010	66815	4800428	0	20.856	21.274	22.5807	422.5	0	26.94065	24.2466	26.97385	24.33285	23.02485	22.77055
1/1 21:00	20125	63825	2821662	0	20.882	21.276	22.58235	422.5	0	25.54725	24.0082	25.66425	24.0917	23.0094	22.8006
1/1 22:00	19895	66700	1322771	30726	20.804	21.272	22.5865	422.6	0	24.10935	23.4252	24.1809	23.47925	22.7093	22.7382
1/1 23:00	884005	3217700	203281410	698931	20.727	21.239	22.58565	421.4	0	25.67225	23.3731	25.9149	23.4669	22.73145	22.7109

Appendix 7 – Multivariate Linear Regression Model

Appendix 7.1 – User Interface of the Stepwise fit model developed for an entire year of data.

The screenshot of the user interface of a stepwise fit model has been shown below. Here the varying p-value and r-squared values can be seen as the model adds or removes certain input parameters. The RMSE drops with increasing number of statistically valid parameters which can be seen from the curve beneath. One important aspect of this tool is that the user can choose to allow a certain parameter from entering or leaving the MLR equation, thus witnessing the effect of changing RMSE or R^2 by the said parameter.

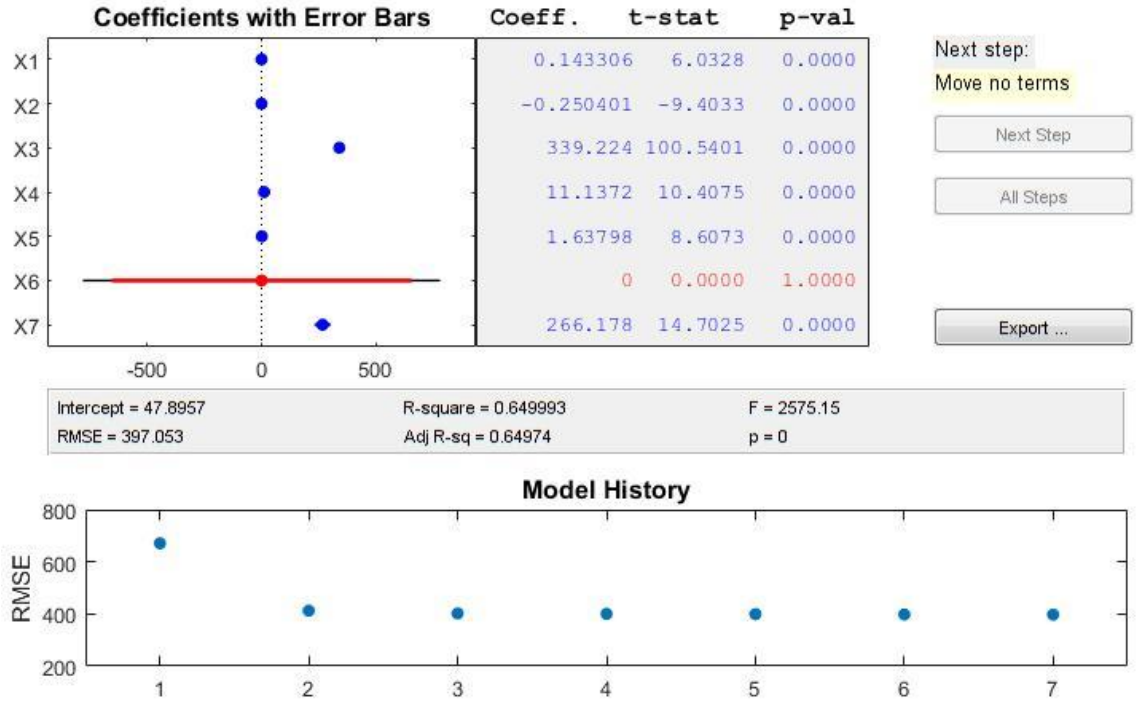


Figure 49 shows the screenshot of a stepwise fit user interface. The values of RMSE and R-squared can be seen along with the coefficients estimated for each input variable. The X6 – which is infiltration by wind can be seen to be removed, due to an excessive P-value.

Appendix 7.2 – Results from Stepwise and FITLM Functions.

Appendix 7.2.1 – Regarding the Statistical Validations of MLR models.

1.FITLM Residuals and their normal distribution

The training is performed over three timesteps, full year's data, working hours and non-working hours of the year. The residuals have a normal distribution with 95% probability as can be seen by the images xx a, b and c below.

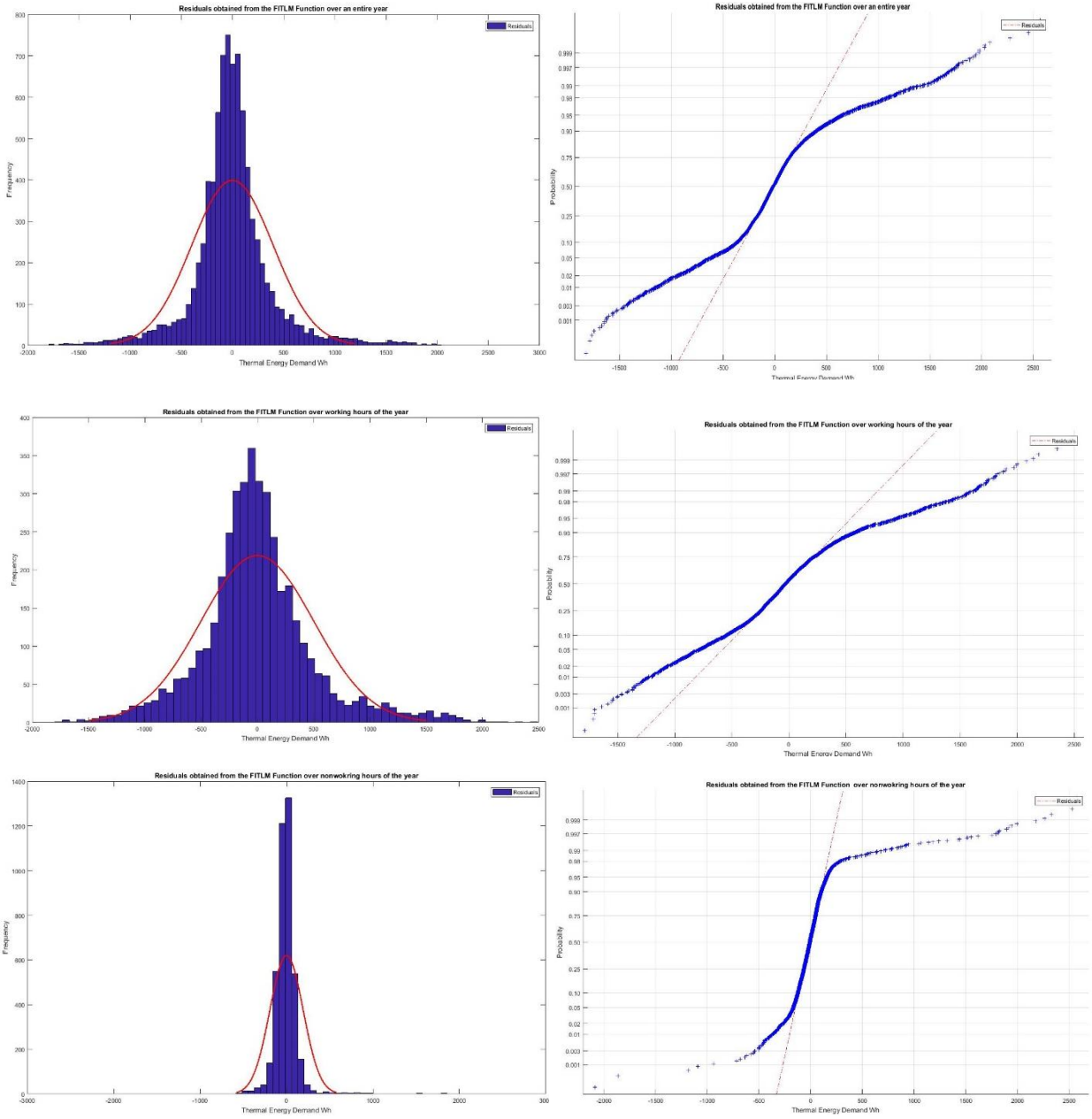


Figure 50 a, b and c show the normally distributed residuals obtained from the FITLM function for datasets belonging to the full year (a) working hours (b) and nonworking hours(c).

2.Stepwise Residuals and their normal distribution

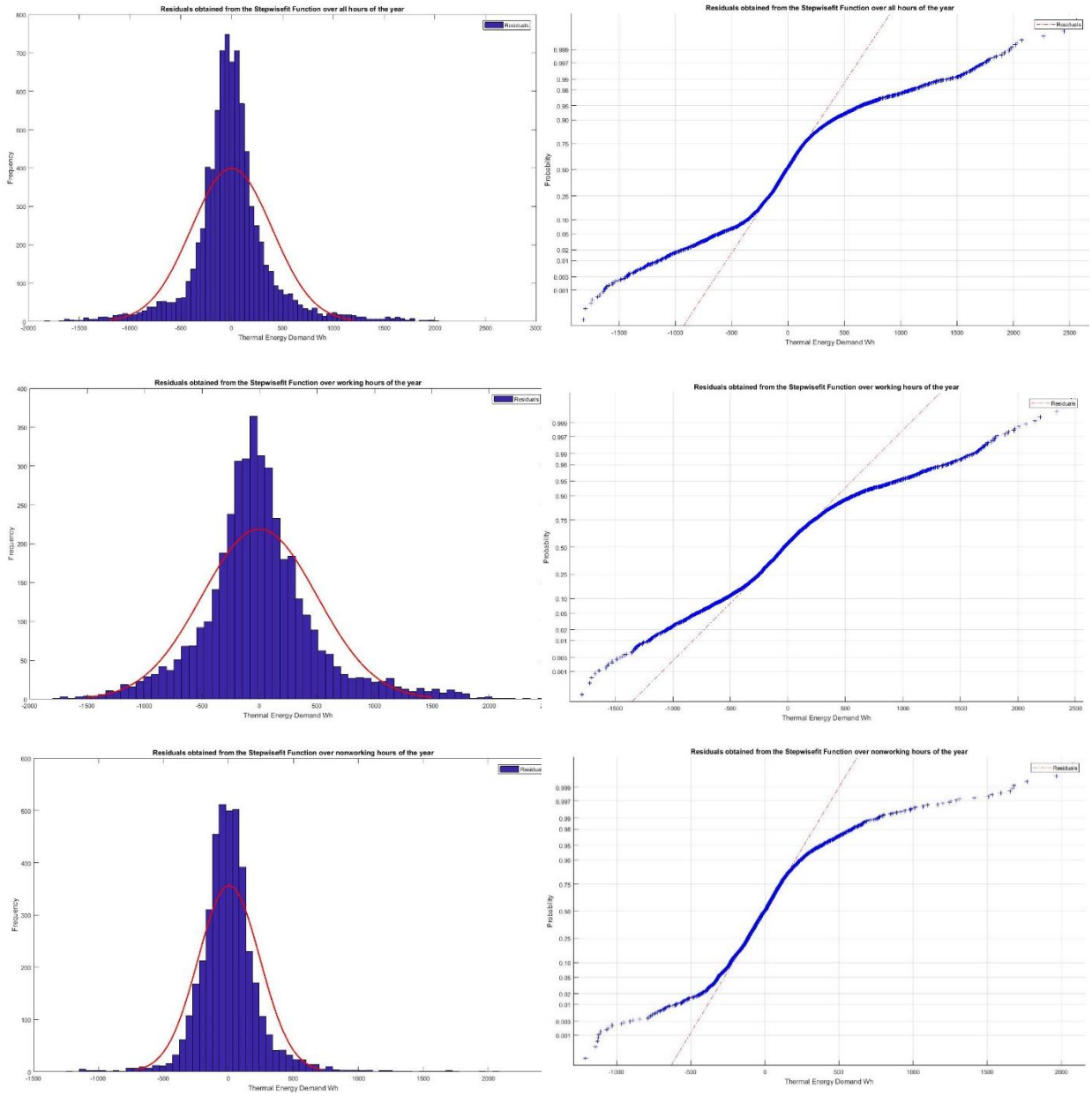


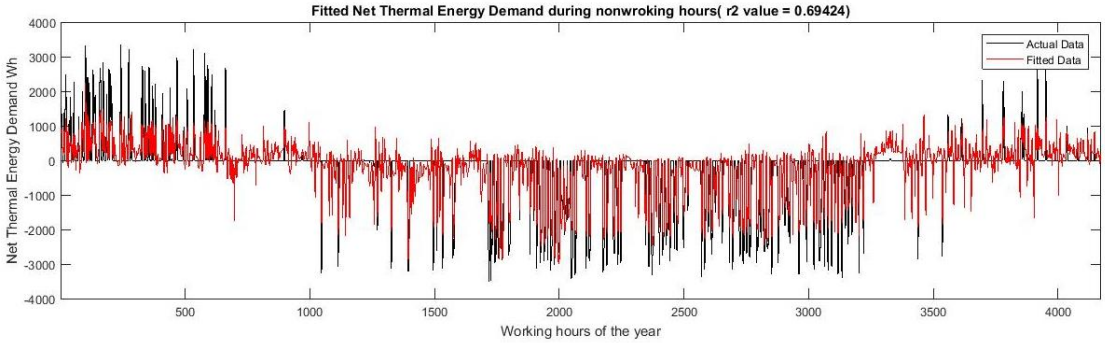
Figure 51 a, b and c show the normally distributed residuals obtained from the Stepwise fit function for datasets belonging to the full year (a) working hours (b) and nonworking hours(c).

Appendix 7.2.2 – Regarding the working and non-working hour graphs.

The fitted data has been graphed over the working and non-working hours as shown below.

1. FITLM

    - *Working hours graph*



    - *Non-working hours graph*

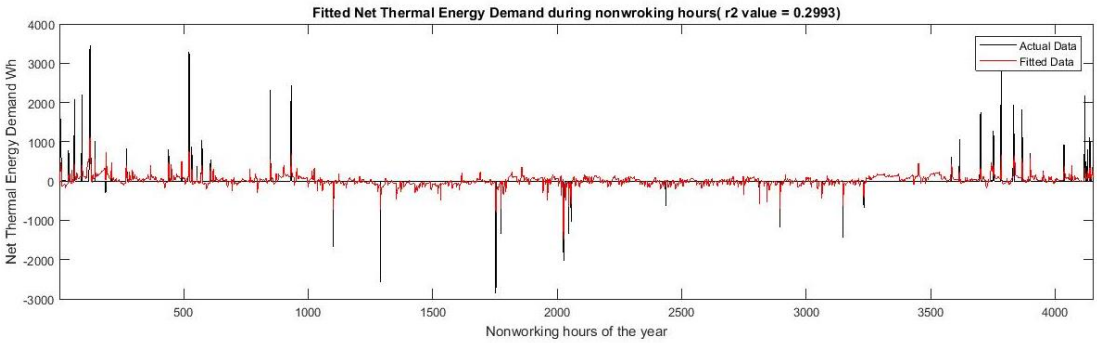
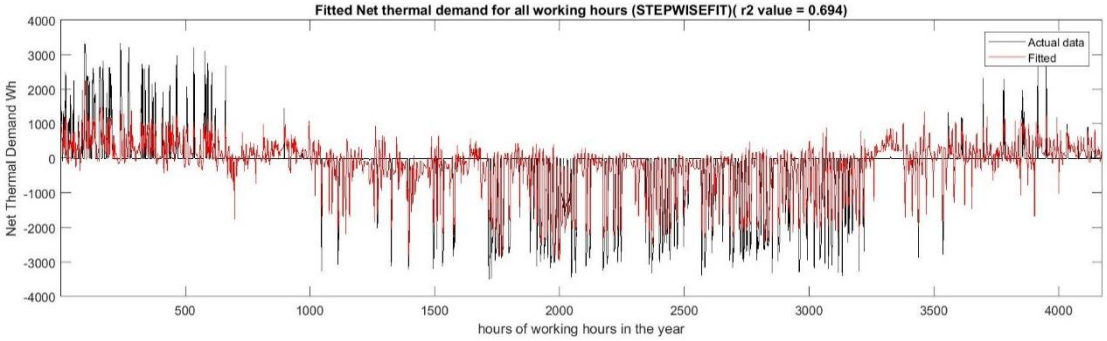


Figure 53 shows the fitted graphical representation of the net thermal demand over the working and nonworking hours of the year using FITLM.

2. Stepwise fit

    - *Working hours graph*



- *Non-working hours graph*

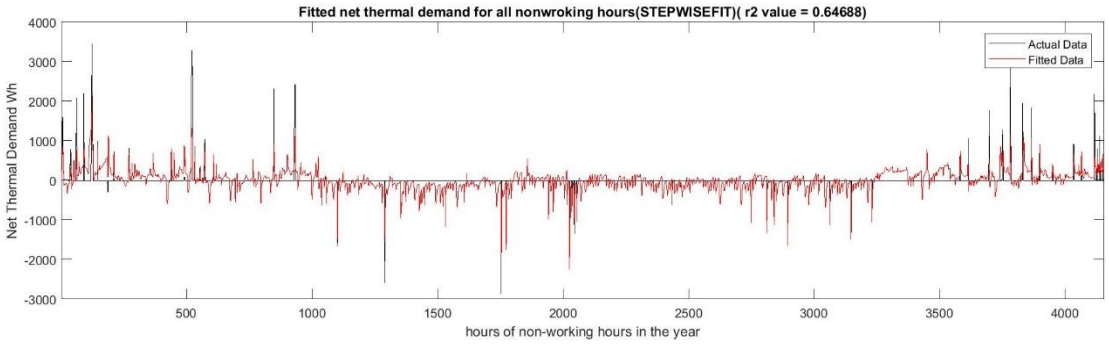


Figure 54 shows the fitted graphical representation of the net thermal demand over the working and nonworking hours of the year using Stepwise fit.

Appendix 7.3 – Comparative Analysis of FITLM and Stepwise Fit functions

The two main functions adopted from MATLAB for the development of MLR models were the FITLM and stepwise fit functions. Both these functions have shown similar results and consumed approximately the same amount of time for training. However, the most important distinctions of the models have been summarized below.

With regards to efficiency **–** It was seen from the results that the overall efficiencies of MLR models using Stepwise fit were much higher than the ones used FITLM. The stepwise fit uses a feed forward method, wherein the parameters are added one by one to the model, in order to improve the fit. This way parameters having a slightly negative effect on the RMSE or R-squared value can be immediately removed. This is probably the main reason for a higher efficiency in stepwise fit models.

With regards to ease of use – The stepwise fit function has a built-in user interface wherein each parameter can be added or removed from the MLR equation. The effect of each of these parameters on the overall R^2 value and RMSE of the model can be observed and taken into consideration. On the other hand, the FITLM function has a backlog wherein, the model runs and presents the results of the fit and prediction. The removal or addition of parameters from the MLR equation must be done manually and the model must be run each time to notice a change in the efficiencies.

With regards to automation – The models developed under stepwise fit are more automated in addition or removal of parameters. It is simpler to use as compared to the FITLM, with higher efficiency and allows for a user based control over the interface.

Appendix 7.4 – Weekly Coefficient estimations using Stepwise Fit.

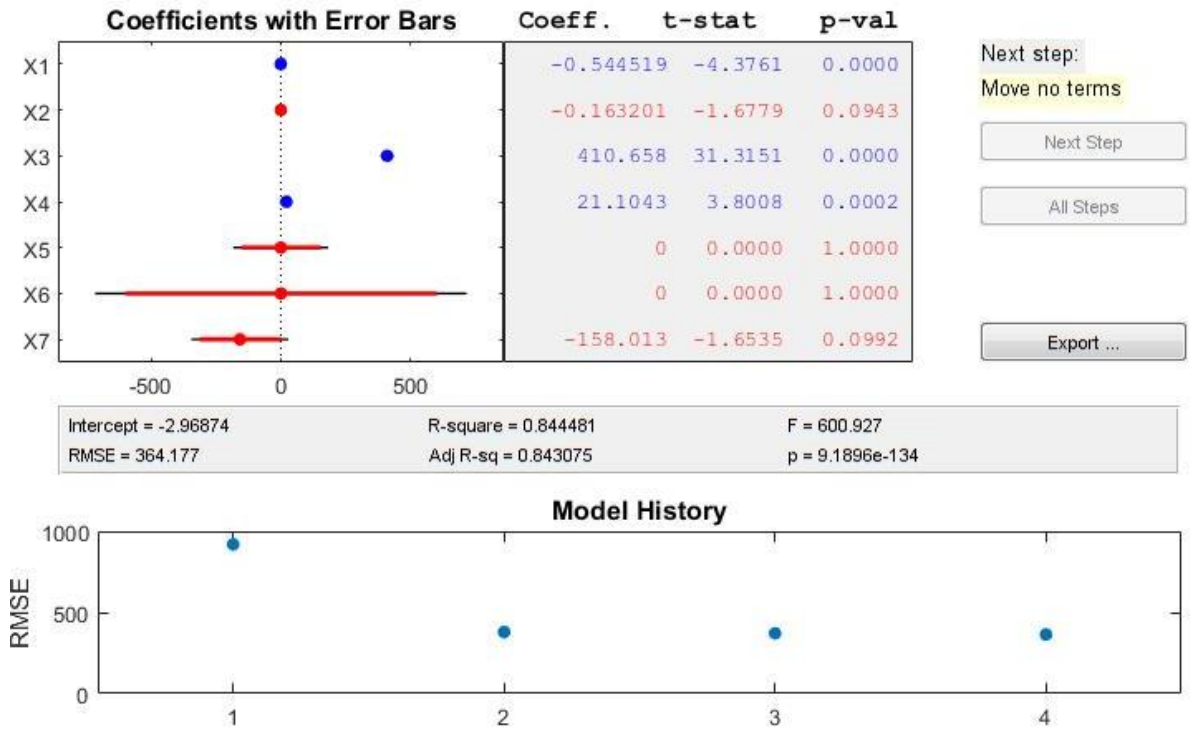


Figure 55 shows the obtained stepwise fit training over 2 weeks of data, 24tha and 25th week in the year.

During a weekly timestep, the most effective and statistically valid parameters contributing to the training are the flux due to internal heating $Q_{internal\ heating}$, Q_{floor} , $Q_{envelope}$**.** These three parameters provide for 84% of the training with an RMSE of 364. With smaller timesteps, the number of parameters effectively contributing towards training has been seen to decrease.

Appendix 7.5- MLR Model Improvements

Accounting for the delay in solar radiation – The graphs obtained for the three sub-datasets with a delay of 1 hour in the data have been shown in figure 56 a, b, and c.

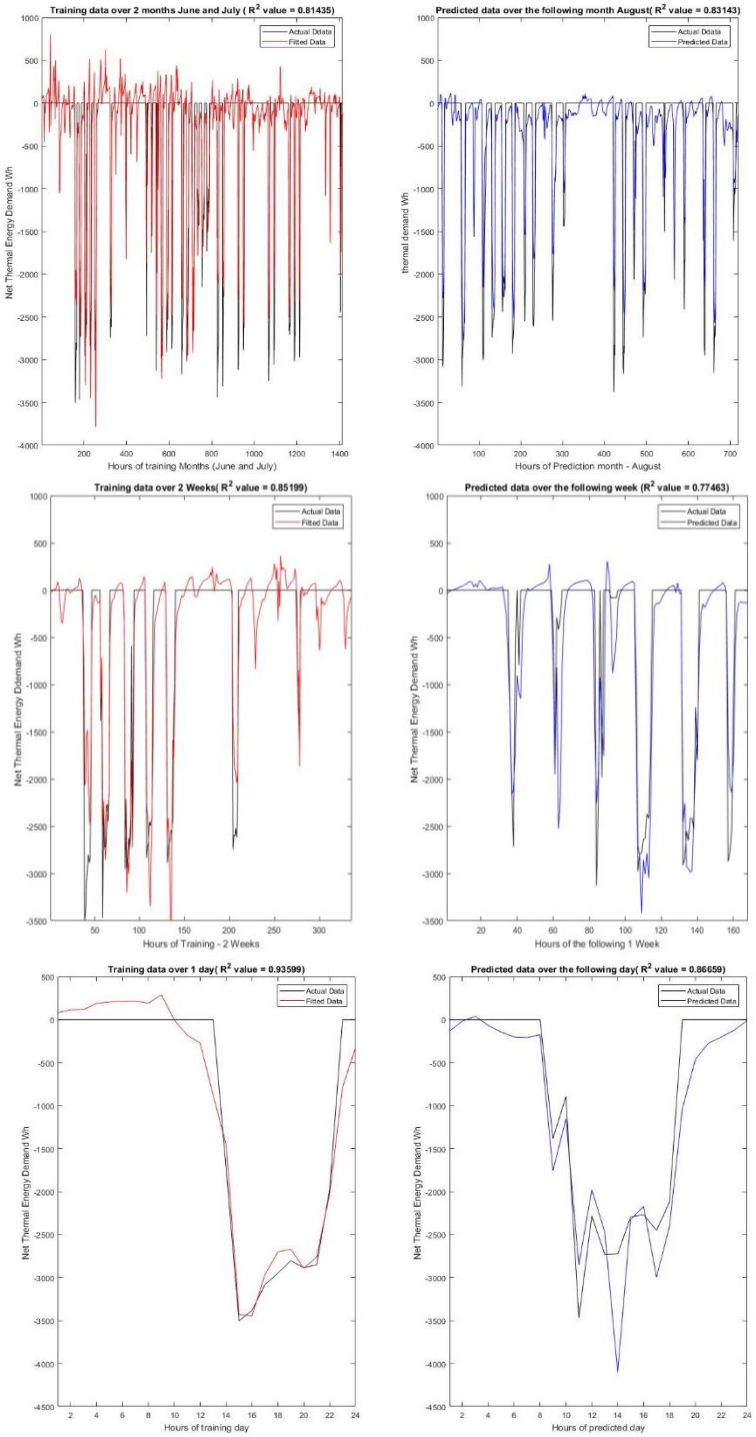


Figure 56 a, b and c – Graphical Images representing the fitted and predicted data over the monthly, weekly and daily sub-datasets, on accounting for the lag in solar radiation.

Using the indoor air temperatures of the side rooms to evaluate their effect on the room in question – Since the heating and cooling of the rooms is done via a priority based system (see section 6.2.4) Perhaps accounting for the adjacent rooms may help improvise the model. This would be taking into consideration if the side rooms need heating or cooling demand, thus enabling a better fit.

In general, models developed on data at the building level, incorporate the use of outdoor climatic conditions, as they impact the thermal demand. This is seen at the room level too. However, an individual room unlike is surrounded by other rooms and the indoor climatic condition of these rooms may or may not influence the thermal demand of the principal room.

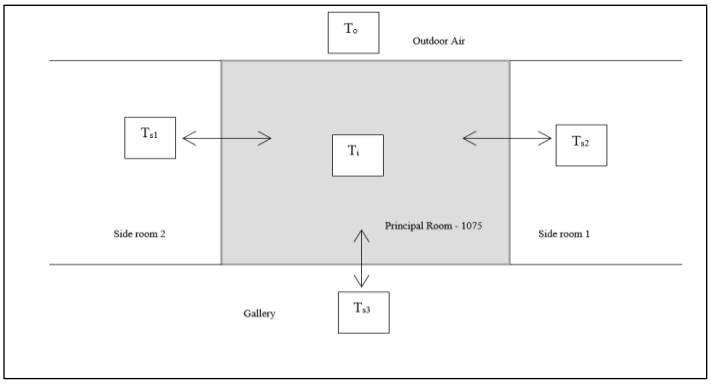


Figure 57 - descriptive diagram of the room 107 and the rooms around it, including the gallery, and the outdoor.

This was added as another heat flux under the thermal flux to the room from the side walls as follows;
Since this building is heavily temperature controlled, and due to lack of information on the temperature in the gallery (see figure xx above) an average temperature of the side rooms has been considered as the secondary outdoor air temperature.

$Q_{envelope}$ (outdoor) = $\sum_i \boldsymbol{U}_{\boldsymbol{o}}^{\boldsymbol{i}} . \boldsymbol{A}_{\boldsymbol{o}}^{\boldsymbol{i}} . (\boldsymbol{T}_{\boldsymbol{o}} - \boldsymbol{T}_{\boldsymbol{i}}) - Existing\ Flux$

$Q_{envelope}$ (side rooms) = $\sum_i \boldsymbol{U}_{\boldsymbol{indoor}}^{\boldsymbol{i}} . \boldsymbol{A}_{\boldsymbol{indoor}}^{\boldsymbol{i}} . (\bar{\boldsymbol{T}}_{\boldsymbol{s}} - \boldsymbol{T}_{\boldsymbol{i}})$ – *Addition Flux*

Where $\boldsymbol{U}_{\boldsymbol{indoor}}^{\boldsymbol{i}}$ – The Heat capacitance (W/m^{2}K) of walls facing the gallery and side rooms and,

$\boldsymbol{A}_{\boldsymbol{indoor}}^{\boldsymbol{i}}$ – The area of the walls facing the gallery and side rooms.

$\bar{\boldsymbol{T}}_{\boldsymbol{s}}$ – Combined average indoor air temperature of the side rooms

The same months and weeks were chosen once again for the MLR model, with an additional variable in terms of ($\bar{\boldsymbol{T}}_{\boldsymbol{s}} - \boldsymbol{T}_{\boldsymbol{i}}$).

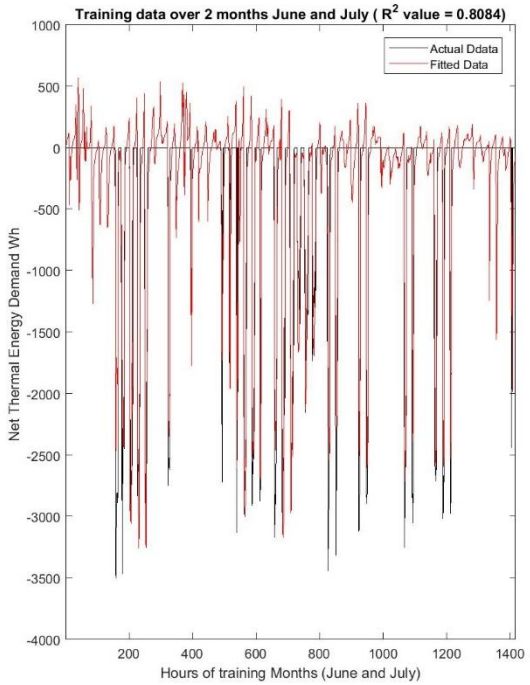
It was seen that there was a degradation in the overall predictive power of the model for the principal room 1075. This could suggest that the side-room temperatures do not play a major role in determining the thermal energy demand at

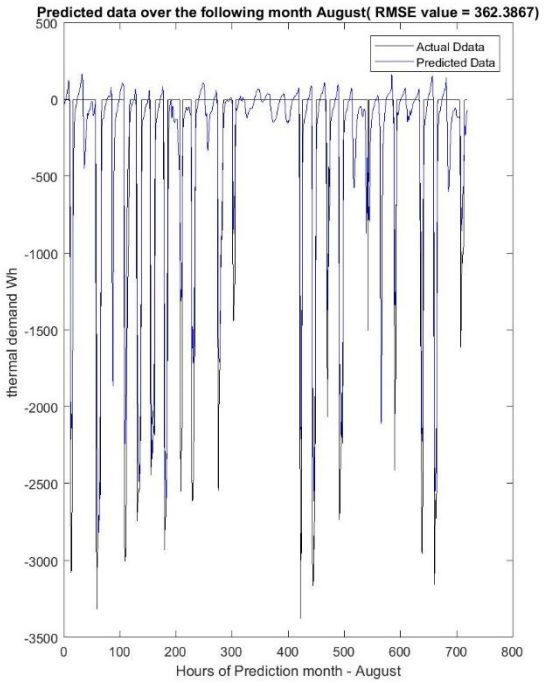
the room level, and form an unnecessary increment in the complexity of the model, leading to a poor predictive capacity. See the annex A.7.xx for graphical representation of the data.

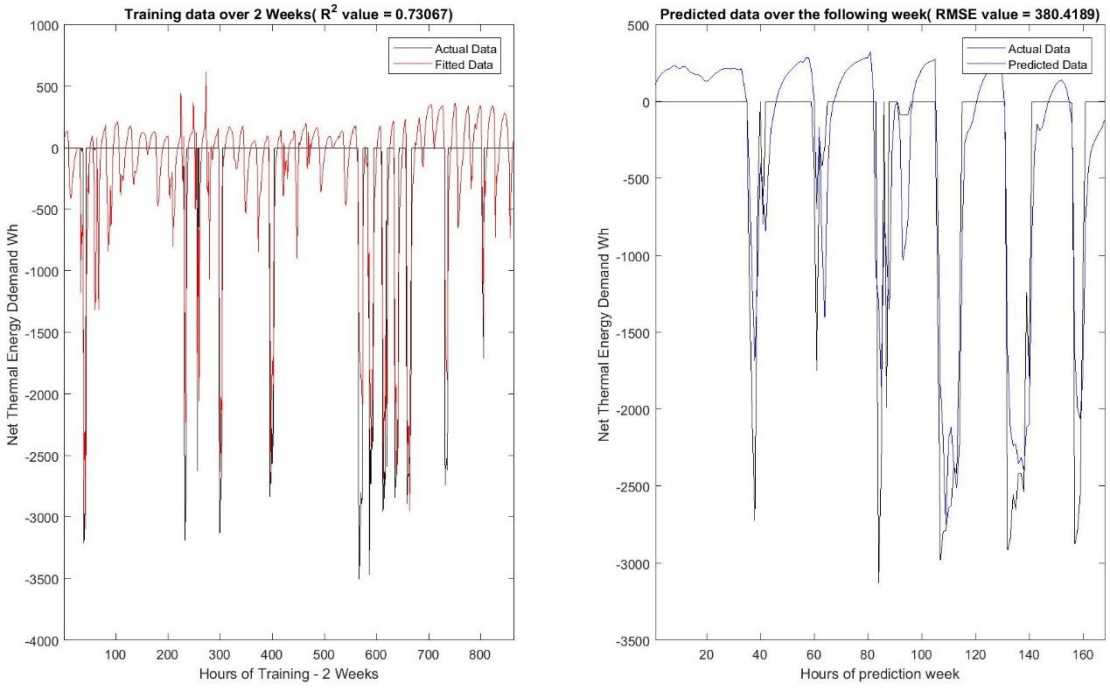
Table 27 Obtained values of R^2 and RMSE for the monthly, and weekly estimates of the room, both with and without the side room temperature.

	Monthly Sub-dataset (Training June and July Prediction Aug)		**Weekly sub datasets- (Week No. 24-26 2015)**	
	Without $((\bar{T}_s - T_i)$	With $(\bar{T}_s - T_i)$	Without $((\bar{T}_s - T_i)$	With $(\bar{T}_s - T_i)$
R2-fit [%]	**84%**	**80.80%**	**88%**	**73%**
R2-pred [%]	**82%**	**78.20%**	**78%**	**69.4%**
RMSE fit Wh	**331.36**	**362**	**320.39**	**364**
RMSE pred. Wh	**342.62**	**362.5**	**423.98**	**380**

Below the figures obtained from the training and predictive models have been placed.







Figures 58 a and b Effect of introducing the side room temperatures in the MLR equations over fitted and predicted data.

Appendix 8 – Artificial Neural Networks

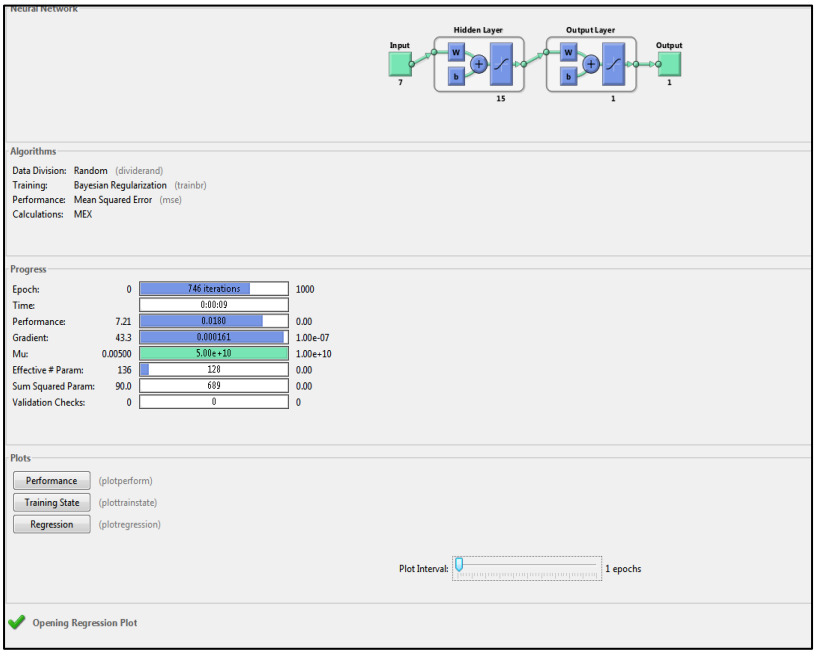


Figure 59 shows the screenshot of the Neural Network toolbox on MATLAB, on completion of a certain model training. It can be seen that the model was trained in 11 seconds, with 748 iterations.

Appendix 8.2- Day-wise ANN models

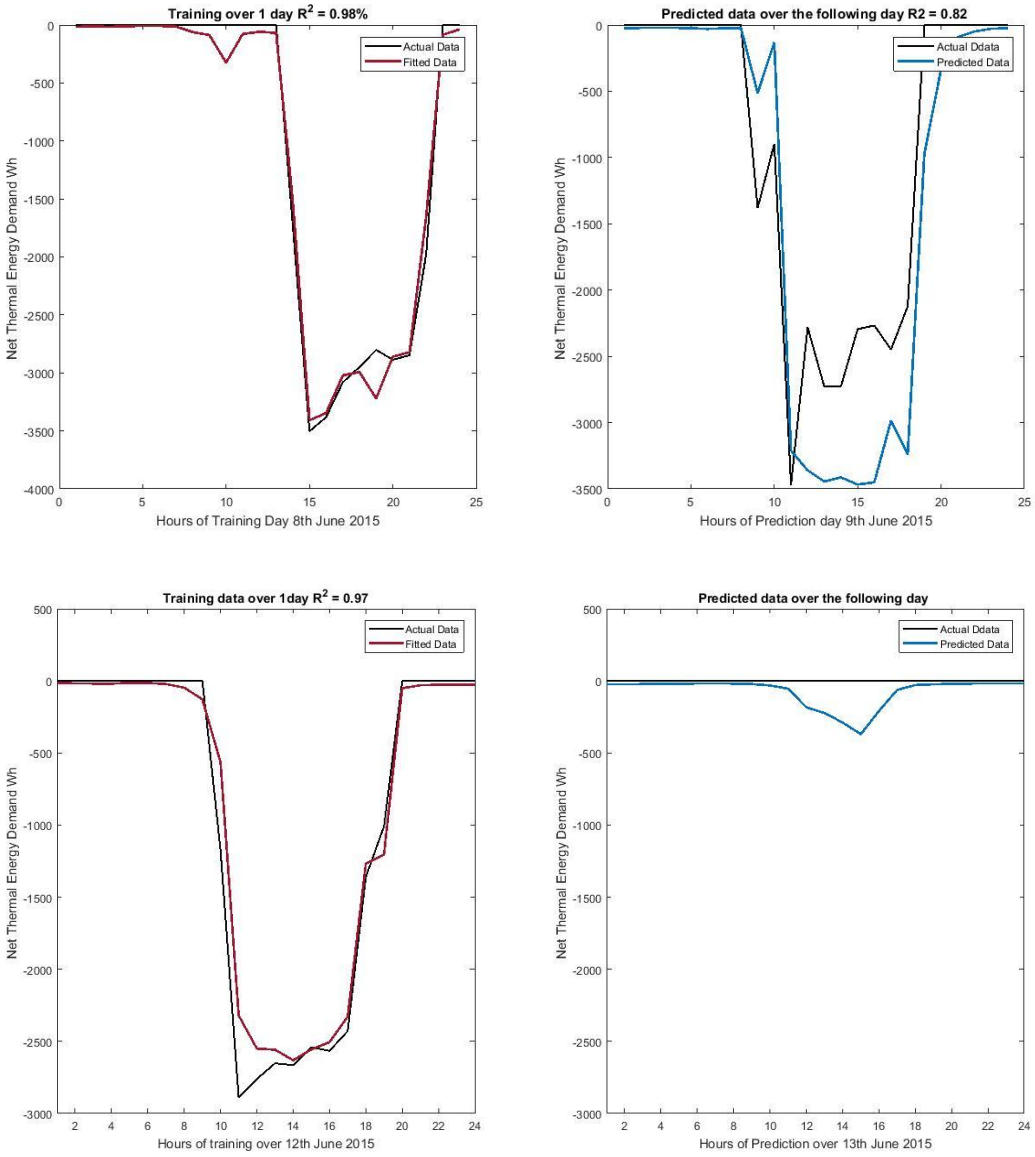


Figure 60 – estimated fitted and predicted data over the day-wise ANN models.

Table 28 Estimated RMSE and R^2 values in thermal energy demand predictions over day-wise sub-datasets for both MLR and ANN models.

Training Day	Adj. R^2 MLR	RMSE MLR [Wh]	Adj. R^2 ANN	RMSE ANN [Wh]	Prediction Day	Adj. R^2 MLR	RMSE MLR [Wh]	Adj. R^2 ANN	RMSE ANN [Wh]
8th June 2015	95%	325.41	98%	229	9th June 2015	87%	437.5	82%	692
12th June 2015	97%	234.45	97.1%	186	13th June 2015	-	278.14	-	122

Table 29 Estimated error in thermal energy demand predictions over the various time-periods chosen in this research for both MLR and ANN models.

Stepwise					
	May	June	Aug	Week	Day
Actual Demand	-226180	-247550	246150	-70724	-22618
Predicted Demand	-120970	-189210	263280	-81342	-27214
Error %	-47%	-24%	7%	15%	20%
R2 pred	49%	50%	82%	78%	87%
ANN					
	May	June	Aug	Week	Day
Actual Demand	-22618	-24755	24615	-70724	-22618
Predicted Demand	-18231	-21275	24912	-72213	-29214
Error %	19%	14%	-1%	-2%	23%
R2 pred	60%	58%	86%	81%	82%