Input Variable Selection for Thermal Load Predictive Models of Commercial Buildings

Dimitrios-Stavros Kapetanakis[a,*], Eleni Mangina[b], Donal P. Finn[a]

[a] *School of Mechanical and Materials Engineering, University College Dublin (UCD)*
[b] *School of Computer Science and Informatics, University College Dublin (UCD)*

Abstract

Forecasting of commercial building thermal loads can be achieved using data from Building Energy Management (BEM) systems. Experience in building load prediction using historical data has shown that data analysis is a key factor in order to produce accurate results. This paper examines the selection of appropriate input variables, for data-driven predictive models, from wider datasets obtained from BEM systems sensors, as well as from weather data. To address the lack of available complete datasets from actual commercial buildings BEM systems, detailed representation of reference buildings using EnergyPlus were implemented. Different types of commercial buildings in various climates are examined to investigate the existence of patterns in the selection of input variables. Data analysis of the simulated results is used to detect the correlation between thermal loads and possible input variables. The selection process is validated by comparing the performance of predictive models when the full or the pre-selected set of variables is introduced as inputs.

Keywords:
building thermal loads, input selection, predictive model

1. Introduction

In Europe buildings account for approximately 41% of total final energy consumption [1]. Furthermore, European Directives are mandating ambitious reductions in building energy consumption and emissions [2, 3]. In order to decrease the energy usage for heating and cooling, as well as increase compliance with the European Directives on the energy performance of buildings [4], it is of fundamental importance to operate efficiently existing Heating Ventilation and Air-Conditioning (HVAC) systems. Effective operation of HVAC systems is one of the most cost-effective methods for ensuring energy efficiency, reliability and safety in the building sector [5]. Implementation of more effective energy management systems in buildings is anticipated to achieve savings of the order of 5% to 20% of energy consumption [5]. Through the optimisation of the control of HVAC systems, building energy use can be reduced while maintaining or even improving occupant comfort [6]. The combination of low capital cost investment

and the potential for significant cost and energy savings has increased the interest of building energy stakeholders in the forecasting of thermal loads in the commercial building domain.

Considering system optimisation, an accurate and rapid forecast of building thermal loads could be useful, depending on its utilisation. Knowledge of immediate or near future building loads can be used to adjust the operation of the HVAC system according to the forthcoming needs of the building, especially during peak energy demand hours [7]. Different techniques have been proposed in recent years to capture this knowledge in the form of predictions [8, 9]. Building simulation software can simulate the operation of a building and generate predictions of thermal loads when detailed building geometry as well as physical data, such as construction elements, U values, etc., and occupancy profiles are available. In reality, such data are often difficult to obtain or even unknown, especially for older buildings, where uncertainty arising from parameter and occupancy estimation can lead to significant additional modelling efforts and consequential commissioning effort in real-world applications [10].

Forecasting building thermal loads can also be achieved by exploiting the combination of recorded data from Building Energy Management (BEM) systems and predictive modelling techniques. These data records include underlying information regarding preceding building thermal loads and can be introduced to data-driven models, which utilise extensive assessment of input and output variables, in order to produce accurate predictions [9]. Several techniques, such as regression [11–13], artificial neural networks (ANN) [14–16] and support vector machine (SVM) [17–19] have been implemented for this purpose. Moreover, the accuracy of the prediction is a crucial factor to allow confidence with this approach. Commercial buildings incorporate BEM systems to control HVAC and to monitor indoor environment conditions. Therefore, commercial buildings are suitable for the application of energy saving measures by optimising the control of their HVAC systems.

The objective of this paper is to investigate the effect of selecting input variables for data-driven models capable of predicting thermal loads of commercial buildings. This pre-selection process is critical for the development of predictive models, since the use of redundant input variables introduces unnecessary additional complexity during the development and execution of the models. To date, input variables are largely identified by model developers based on a combination of domain expertise and performance heuristics. Review of the existing literature reveals that little previous work has been published concerning the selection of input variables [20]. Robust variable selection could aid in more efficient data collection, cleaning and dissemination from commercial buildings, since the importance of each variable varies [20]. In the current paper, a filter approach is described based on a statistical analysis of historic performance data, which for commercial buildings is typically available from BEM systems and meteorological weather data and is applied to different commercial building types in various climates. The effect of the selection of input variables is captured by comparing the performance and execution time of neural network predictive models, when the full or the pre-selected set of variables is introduced as inputs.

The remainder of this paper is organised as follows. Section 2 provides an overview of related work, highlighting the fact that the selection of input variables for the predictive models has not been a primary focus. Section 3 describes the methodology, which is based on data analysis, used in the selection of input variables. General and specific case

studies results are presented in Section 4. The evaluation of the effect of this selection process on predictive models is also given in Section 4. Finally, conclusions and possible future work are provided in Section 5.

2. Background

Regression, ANN and SVM are the most commonly applied techniques in the literature for generating predictions of building thermal loads without the use of physics-based simulation software. However, towards the development of these predictive models, the selection (or justification) of input variables has not been subjected to the same level of scrutiny as for physics-based whole-building simulation models.

In regard to input variables, physics-based energy models use different inputs in order to produce useful results. Nevertheless, some inputs have more significant influence than others on the modelled energy performance of the building [21]. Numerous studies have used sensitivity analysis to identify which building characteristics have the greatest impact on energy results from building models [21–24]. On the other hand, the variables selected as inputs for data-driven models are usually identified by model developers based on a combination of domain expertise and availability of data. Concerning the data-driven predictive models, various studies have applied regression methods for forecasting building thermal loads, [11, 25, 26], utilising different variables as inputs without implementing any pre-selection process. This lack of selection of input variables has also been noticed in case studies where ANN [14, 15, 27–29] and SVM [17, 18, 30, 31] models were developed.

There are only few examples existing in the literature of researchers that systematically pre-selected the input variables of their predictive models during the development stage. Lei and Hu [32] created a baseline model for office building energy consumption in a hot summer and cold winter region using simple linear regression for the energy consumption with three weather variables considered as inputs. Furthermore, a linear correlation of input variables with the energy consumption was investigated and taken into account for the development of the regression models. Zhao and Magoulès [33] developed SVM predictive models of fifty office buildings, in France, to forecast their hourly electricity consumption using as possible inputs; weather variables, occupancy, internal heat gains and indoor variables. The datasets were generated using EnergyPlus and two input variable selection techniques, correlation coefficient and gradient guided selection, were applied. Results indicated that the selected subset of input variables was valid and provided acceptable predictions. Lately, Massana et al. [34] compared regression, ANN and SVM models predicting the electical load of an office building in Girona, Spain. The available input variables for the models were weather variables, indoor temperature, occupancy and calendar data from which a sub-set was selected after testing empirically several configurations. The SVM model with only temperature and occupancy as input variables provided the best balance of accuracy and computational cost.

A summary of all previous case-studies regarding building types and climate used as well as their input variables is presented in Table 1.

Table 1: Summary of previous case-studies

Reference	Building Type	Climate	Method	Input Variables
[11]	Banking Sector	Spain	Regression	Construction characteristics, climatic location and energy performance
[25]	Single-family Residential	Temperate	Regression	Building shape factor, envelope U-value, window to floor area ratio, building time constant and climate
[26]	17 Residential Buildings	Temperate	Regression	Building heat loss, south equivalent surface and ambient temperature
[32]	11 Commercial Buildings	China	Regression	Outdoor temperature, relative humidity and solar radiation
[27]	250 Residential Buildings	Cyprus	ANN	Window area, wall area, floor area and type of windows and walls
[28]	9 Residential Buildings	Cyprus	ANN	Window area, wall area, floor area, type of windows and walls, roof classification and room temperature
[15]	N/A (Contest data used)	N/A	ANN	Current and forecasted temperatures, current load, hour and day
[29]	2 Commercial Buildings	Canada	ANN	Outdoor wet-bulb temperature and exit water temperature of the chiller
[14]	3 Sample Buildings	Turkey	ANN	Orientation, insulation thickness and transparency ratio
[17]	4 Commercial Buildings	Tropical	SVM	Mean outdoor dry-bulb temperature, relative humidity and solar radiation
[30]	1 Commercial Building	China	SVM	Outdoor dry-bulb temperature and solar radiation intensity
[31]	1 Commercial Building	China	SVM	Outdoor temperature, humidity and solar radiation
[18]	1 Commercial Building	China	SVM	Temperature and humidity of air provided to the HVAC system
[33]	4 Commercial Buildings	France	SVM	Outdoor temperature, humidity, solar radiation, wind speed, ground temperature, occupancy, internal heat gains, indoor temperature and infiltration

Continued on next page

Table 1: Summary of previous case-studies (Continued)

Reference	Building Type	Climate	Method	Input Variables
[34]	1 Commercial Building	Spain	Regression, ANN, SVM	Outdoor temperature, humidity, solar radiation, indoor temperature, occupancy and calendar data

The research field related to building thermal loads forecasting has been very active, involving various regression and data mining techniques. Nevertheless, it is clear from the literature that the justification of the selection of the input variables to the predictive models has not been the primary focus. Hence, the investigation of which variables and why they should be considered as inputs should be a priority towards the development stage of the model. This investigation is crucial for the predictive models, considering that excessive input variables introduces unnecessary complexity during the development and execution of the models. Moreover, the selection process of input variables should be achieved utilising a systematic approach, to maintain consistency, while being applied to different building types in various climates, in order to cover as many case studies as possible.

In general, input variable selection methods can be distinguished into three major categories as follows: filter, wrapper and embedded methods [33]. The filter method is based on ranking the input variables with a correlation or mutual information criteria and the selection is based on the highest ranking. This method can be considered as a pre-processing step since it takes place before the development of the predictive model [33]. The wrapper method identifies and evaluates the subsets of input variables according to the magnitude of accuracy they contribute to a given output variable. Similarly to wrapper method, the embedded one assesses in the same way the benefit of input variable sets, but the selection occurs directly in the training process. In this way multiple training for each candidate subset is avoided [35].

3. Methodology

The methodology developed in the current research is based on the filter method to detect interrelationships between variables in order to select the input variables for predictive models capable of forecasting thermal loads of commercial buildings. This methodology is part of the overall research methodology followed by the authors as previously published in [36]. The process of data analysis consists of the investigation of linear and monotonic correlations, to identify the intra-variable relationships and the relative importance therein. To achieve this, the sequence presented below is followed:

i A synthetic dataset is generated for a representative group of commercial buildings in various climates using EnergyPlus

ii Data analysis for the investigation of linear and monotonic correlations between variables by calculating the Pearson and Spearman correlation coefficient, respectively

iii Selection of input variables

iv Evaluation of the input variables selection on the accuracy of ANN thermal loads predictive models

3.1. Synthetic Database Generation

To avoid the implications of missing data a synthetic database is created utilising representative commercial buildings in different climate types. Taking into account the fact that is infeasible to model every commercial building, or even to represent every building sub-sector [37], a small number of prevalent building types is selected. Six reference commercial building models created by the Department of Energy (D.O.E.) of the U.S. [38] are selected to acquire the synthetic database. The repository of D.O.E. covers building types that directly characterize more than 60% of the commercial building stock and are very similar to other commercial building types. The six reference buildings implemented in this research are summarised in Table 2 and their EnergyPlus models have been obtained from D.O.E. [38]. The EnergyPlus models, when executed, provide one year simulated data at 15-minute intervals. Data analysis is based on simulated data in order to preclude the possibility of dealing with incomplete data sets that are often acquired when extracting recorded data in BEM systems.

Investigation of all possible climate conditions is a prerequisite in order to identify the existence of a pattern in the selection of input variables for a predictive model of thermal loads of commercial buildings. Thus, sixteen representative cities were selected to capture all type of climates. The selection was based on the American National Standards Institute (ANSI) and the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) distinction of climate zones [39]. Table 3 displays the representative cities of each zone and their climates, while their associated EnergyPlus weather files (EPW) were also obtained from D.O.E. [38]. The geometry and the types of HVAC systems of the reference buildings are kept the same for all climate zones. Furthermore, all buildings are mechanically ventilated and their windows are double-glazed. On the contrary, each reference building utilised has different construction characteristics (U-values, etc) and HVAC system sizing for different climate zones, to be compliant with the established building regulations.

Table 2: Reference Buildings [38]

Building Type	**Floor Area (m^2)**	**Floors No.**	**Heating System**	**Cooling System**	**Window to Wall Ratio (%)**	**Building Index**
Hospital	22,420	5	Gas boiler	Water cooled chillers	15.85	B1
Large Hotel	11,345	6	Gas boiler	Air cooled chiller	30.16	B2
Large Office	46,320	12	Gas boiler	Water cooled chillers	40.00	B3
Secondary School	19,590	2	Gas boiler and furnace	Air conditioner	35.00	B4
Strip Mall	2,090	1	Gas boiler	Water cooled chillers	10.50	B5
Supermarket	4,180	1	Gas furnace	Unitary DX	10.86	B6

Table 3: Representative Cities and Climates [39]

Zone No.	Representative City	Climate
1A	Miami	Very Hot, Humid
1B	Riyadh	Very Hot, Dry
2A	Houston	Hot, Humid
2B	Phoenix	Hot, Dry
3A	Memphis	Warm, Humid
3B	El Paso	Warm, Dry
3C	San Francisco	Warm, Marine
4A	Baltimore	Mild, Humid
4B	Albuquerque	Mild, Dry
4C	Salem	Mild, Marine
5A	Chicago	Cold, Humid
5B	Boise	Cold, Dry
5C	Vancouver	Cold, Marine
6A	Burlington	Cold, Humid
6B	Helena	Cold, Dry
7	Duluth	Very Cold
8	Fairbanks	Subarctic

Table 4: Input Variables Under Examination

Weather Data	Indoor Data
Ambient Temperature (^{o}C)	Zone Air Temperature (^{o}C)
Ambient Relative Humidity (%)	Zone Relative Humidity (%)
Wind Speed (m/s)	
Solar Radiation (W/m^2)	

The combination of the selected reference buildings and the different climate zones produces a sample of 102 case studies that are simulated with EnergyPlus. The generated results of the simulations form the synthetic database that is utilised for the data analysis process. EnergyPlus has the advantage of reporting a variety of input and output variable data, while BEM systems typically record a limited range of variables. In order to generate a realistic synthetic database, the most commonly measured variables in BEM systems of commercial buildings are chosen as possible input variables and are grouped according to weather data and indoor variables, as given in Table 4. Heating and cooling loads (in kWh) of the representative commercial buildings are the chosen output variables. It is noted that building characteristics such as height, floor and window area, wall and insulation thickness are crucial elements when calculating the thermal loads of a building. However, for each particular case study these parameters remain

constant over the analysis period of the buildings, hence they have no contribution to predictive model training [33]. Furthermore, the accuracy of the predictive models could benefit from the introduction of building internal heat gains as an input variable, but these kind of data are rarely measured in conventional BEM systems, thus they were not considered as an input variable in the present research.

3.2. Data Analysis

The procedure followed for analysing the data and examining the existence of a correlation between input and output variables was based on statistical techniques. The data analysis was performed only for the hours when the HVAC systems provided heating or cooling to each building case study. Initially, the existence of a linear correlation between input and output variables was investigated by performing a Pearson correlation [40], which measures the linearity between paired data. Furthermore, the existence of monotonic relationships was explored as well through the Spearman correlation coefficient [40]. Both Pearson and Spearman correlation coefficients are constrained between -1 and +1, illustrating the relationship between two continuous variables. The statistical analysis was carried out using the IBM SPSS Statistics 20 software [41].

3.3. Selection of Input Variables

The selection process of the input variables that will be introduced to the predictive models is based on the calculated Pearson and Spearman correlations between input and output variables. Absolute values of the calculated Pearson and Spearman coefficients are used to simplify this process. Only moderate, strong and very strong relationships are of interest, hence the threshold value for introducing an input variable to the predictive model is 0.5, based on the guide of the absolute values, as suggested by Evans in [42].

3.4. Evaluation of Input Variables Selection

The effect of the input variables selection on thermal load predictive models is evaluated with the comparison of the performance of discrete ANN models, developed with all possible variables or only the proposed selected ones as inputs. Moreover, individual predictive models are developed for each case study for the heating and cooling loads of the building. Data from the first six months of the year (January to June) are used as the training partition during the development of the models while the remainder of the data (July to December) are used as the testing partition.

The accuracy of each predictive model is calculated based on the root mean square error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|^2}{n}} \tag{1}$$

where, y are the actual values, $\hat{y}$ are the predicted values of the heating and cooling loads and n is the total number of timesteps summed up at the testing partition period.

The IBM SPSS Modeler 14.2 software [43] was used for the development of the predictive models using a computer with Intel Core i7-3630QM processor and 8 GB of DDR3 RAM.

4. Results

The initial task of the methodology was the creation of a synthetic database. This was achieved with the simulation of the selected reference buildings using the weather data from the climate zones described in Table 3. The output and input variables of each case were recorded. Once the database was created, the next step was the examination of linear and monotonic correlation between input and output variables using the Pearson and Spearman coefficient analysis, respectively.

4.1. Overall Results

To assess the obtained results efficiently, a colour coded table was generated. Firstly, the reference commercial buildings were assigned a predefined building index, as given in Table 2. The colour code utilised for the visualisation of the results is outlined in Table 5. Green was used to indicate that both heating and cooling loads are correlated with the selected input variable. Red denotes the existence of correlation between the selected input variable and the heating load, while the blue represents the correlation between the selected variable and the cooling load. In case that there was no correlation between the selected input variable and heating or cooling loads, white was employed.

The results from the Pearson correlation analysis are presented in Figure 1. The Pearson coefficient between each possible input variable and heating and cooling loads was calculated and the results were interpreted using the colour coding, as per Table 5. All possible input variables were examined for all reference buildings using all the different climate zones, as per Table 3.

The Pearson correlation analysis reveals strong correlations between well-established input output variable pairs, such as Ambient Temperature with heating and cooling loads, that can be observed to be “strong” in almost all cases. The other possible input variables for the predictive models seem to be correlated variously with the thermal loads, depending on the type of building and climate zone of each case. Wind Speed was found not to be correlated linearly with heating and cooling loads of the reference buildings in any climate zone, hence was omitted from Figure 1. This was most likely due to the implementation of mechanical ventilation and the usage of double-glazed windows for all reference buildings. The Pearson correlation coefficients between Wind Speed and heating and cooling loads varied from 0.0003 to 0.43 and from 0.001 to 0.38, respectively, for all case studies. In particular, the highest Pearson correlation coefficients between Wind Speed and heating loads were detected for all reference building types in climate

Table 5: Colour Coding Explanation

Colour	Correlation
	Heating AND Cooling
	Heating Only
	Cooling Only
	No Correlation

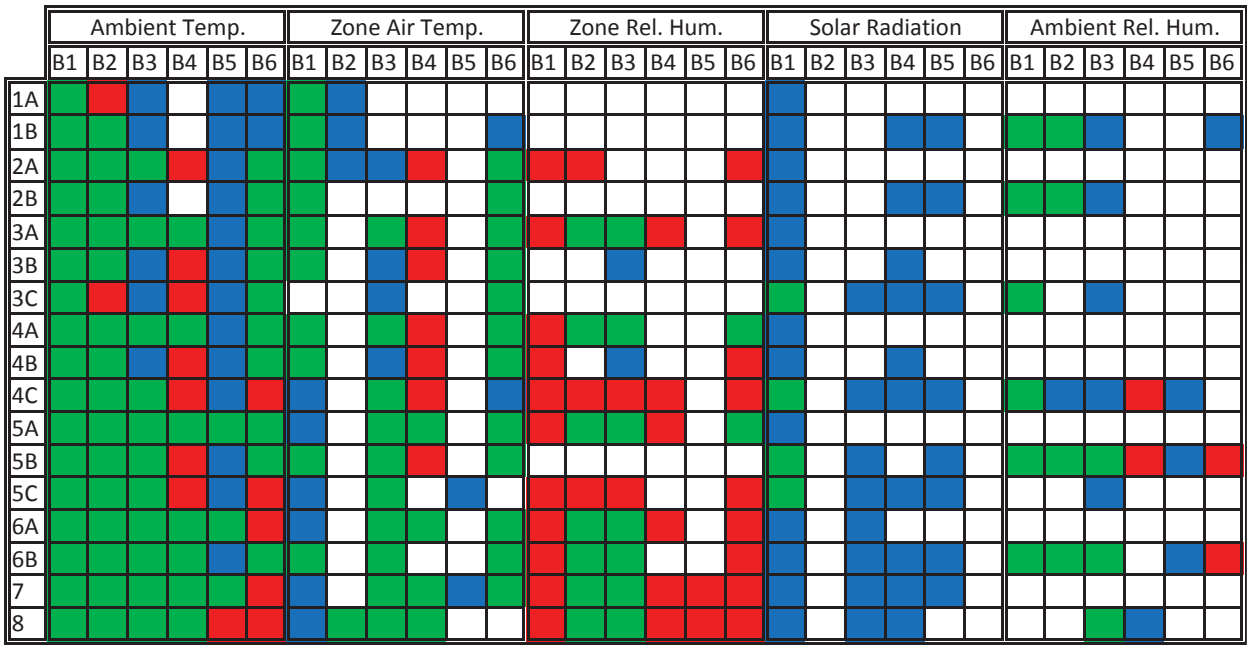


Figure 1: Pearson Correlation Results.

zone 8, where the climate is subarctic, but none of them was found to be higher than the threshold value of 0.5. A similar lack of correlation between Wind Speed and energy consumption was found in [33], while using correlation coefficient and gradient guided selection methods.

Proceeding with the methodology, a Spearman correlation analysis was performed. The outcome of the examination for monotonic relationships between input and output variables is displayed in Figure 2. Based on the Spearman analysis, the existence of correlation between variables that had been discovered with Pearson coefficient is verified. Figure 2 reveals that the correlation between some input and output variable pairs can be better described as a monotonic one, since cases that no linear correlation was detected have resulted a monotonic correlation. A monotonic relationship is one that the dependent variable either never increases or never decreases as the independent variable increases. The strongest monotonic correlation was detected once again between the Ambient Temperature and the thermal loads of the reference buildings. Furthermore, from the Spearman correlation analysis, a useful finding is the correlation between Zone Air Temperature and buildings thermal loads, which can be observed in more cases than using the Pearson correlation analysis. The results from linear and monotonic correlation analysis are different due to the fact that Spearman correlation coefficient is also robust to outliers, unlike Pearson correlation coefficient. Solar Radiation, Zone Relative Humidity and Ambient Relative Humidity were correlated diversely with heating and cooling loads, depending on the type of building and the climate zone of each case. Another interesting finding is that once again Wind Speed was not correlated monotonically with heating and cooling loads of the reference buildings in any climate zone. The usage of mechanical ventilation and double-glazed windows for all reference buildings are potential reasons for not detecting any monotonic correlation between Wind Speed and thermal loads. More specifically,

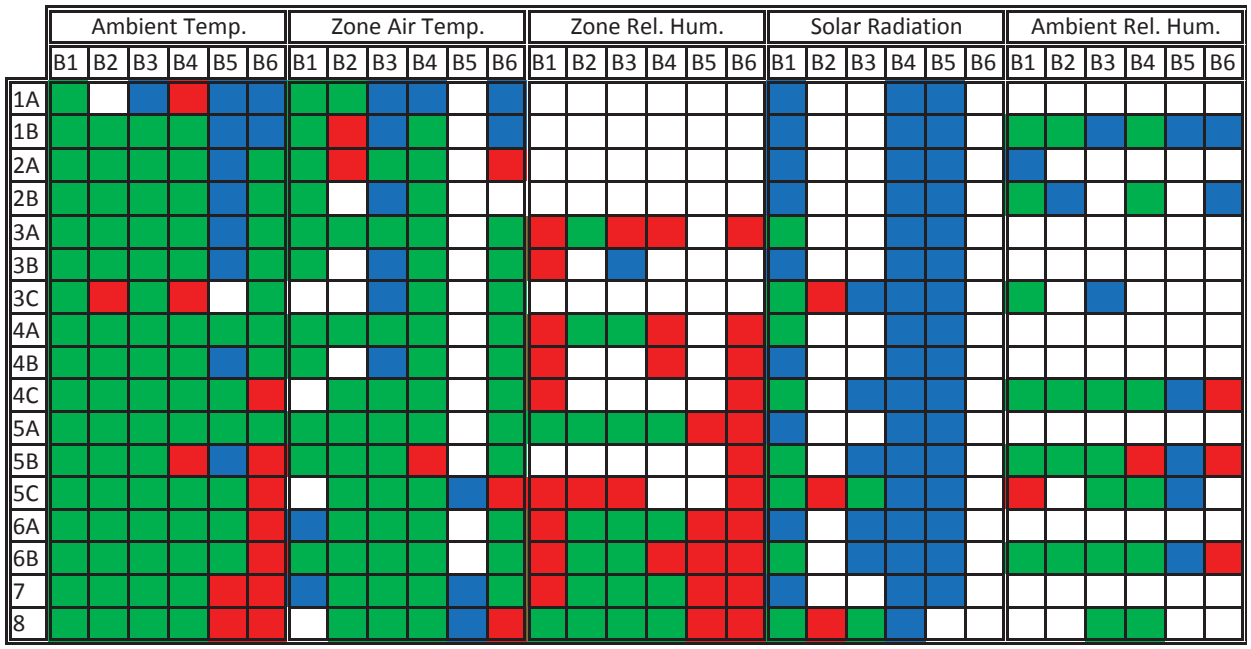


Figure 2: Spearman Correlation Results.

Spearman correlation coefficients between Wind Speed and heating and cooling loads varied from 0.009 to 0.48 and from 0.009 to 0.4, respectively, for all case studies, but similarly to Pearson correlation coefficients none of them was found to be higher than the threshold value of 0.5.

Closer examination of Figures 1 and 2 revealed that 12 and 3 cases, out of the 102 case studies in total, were found to detect no statistical correlation between the input variables and the heating and cooling load, respectively. These case studies are presented in Table 6, denoted by their building type and the climate zone. Furthermore, they were excluded from the assessment of the effect of the selection of input variables. Attempting to interpret the absence of correlation between the input variables and the thermal loads for the case studies at Table 6, an interesting observation was noticed. For the case studies, identified not to have a correlation between the heating load and the input variables, 9 out of 12 cases referred to the Stripmall (B5) reference building type. This was an indication that the existence of no correlation might be linked to the HVAC operation schedule of this particular building type. The other 3 case studies with no correlation between the heating load and the input variables were more straightforward to be explained, since they were detected in very hot climate zones (1A and 1B), where the requirements for heating were limited or even non existent. On the contrary, for the case studies where no correlation between the cooling load and the input variables was identified, 2 out of 3 refer to the Supermarket (B6) in cold climate zones (5C and 8), where the requirements for cooling were limited. Lastly, the third case study, that no correlation was captured between the cooling load and the input variables, referred to the Large Hotel (B2) and might also be linked to the HVAC operation schedule.

Table 6: Case Studies with No Correlation between Input Variables and Heating or Cooling Loads

Heating Load	**Cooling Load**
B3_1A	B2_3C
B5_1A	B6_5C
B5_1B	B6_8
B5_2A	
B5_2B	
B5_3A	
B5_3B	
B5_3C	
B5_4B	
B5_5B	
B6_1A	
B6_1B	

4.2. Specific Case Studies

An in-depth examination of specific case studies illustrates the robustness of the proposed methodology and its ability to capture the nature of the correlation, between input and output variables, when the same reference building is analysed in different climates. The case studies subject to additional examination are the reference buildings Large Hotel (B2) and Large Office (B3) in zones 1B and 5A, where the climate is Very Hot and Dry and Cold and Humid, respectively. The selection of these two types of commercial buildings is based on the likelihood that hotels and large offices could be the building categories most amenable to integrate predictive models in their operation. Moreover, climate zones 1B and 5A are selected to demonstrate the performance of the methodology for an extreme transition from a hot and dry to a cold and humid climate.

The correlation of the variables is illustrated implementing colour coded scatter plots of the available data from the synthetic database. As earlier, the white background at the scatter plot denotes that there is no correlation between the plotted variables. Gradient hues of orange were used for the background, to indicate the existence of ”moderate” (light orange), ”strong” (orange) and ”very strong” (dark orange) correlations.

4.2.1. Large Hotel (B2)

Figures 3 and 4 depict the results from the Pearson and Spearman correlation analysis for the Large Hotel (B2) in climate zone 1B, respectively. The correlation between the thermal loads of the building with Ambient Temperature and Ambient Relative Humidity is captured using Pearson and Spearman analyses. Moreover, Zone Air Temperature is correlated with Cooling Load when using the Pearson analysis. However, the existence of ”strong” correlation between both Heating and Cooling Load and Zone Air Temperature is revealed from the Spearman analysis. Hence,

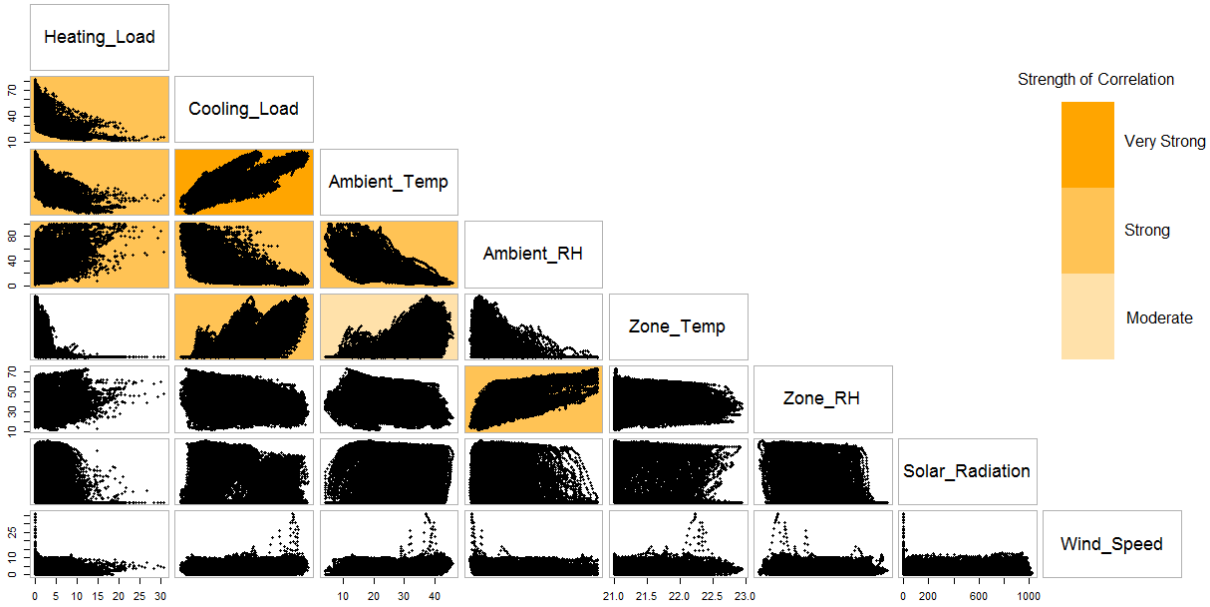


Figure 3: Linear Correlation Between Variables for the Large Hotel (B2) in Climate Zone 1B.

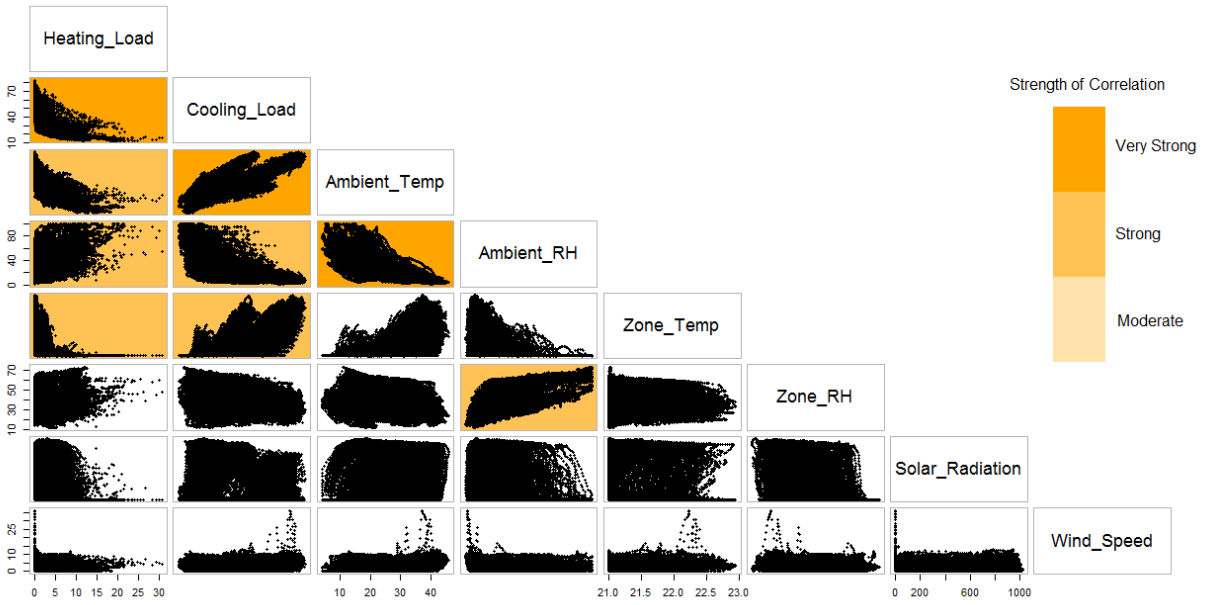


Figure 4: Monotonic Correlation Between Variables for the Large Hotel (B2) in Climate Zone 1B.

based on the developed methodology, the input variables for the Large Hotel in climate zone 1B should be Ambient Temperature, Ambient Relative Humidity and Zone Air Temperature.

Linear and monotonic correlation analysis for the same building in climate zone 5A are displayed in Figures 5

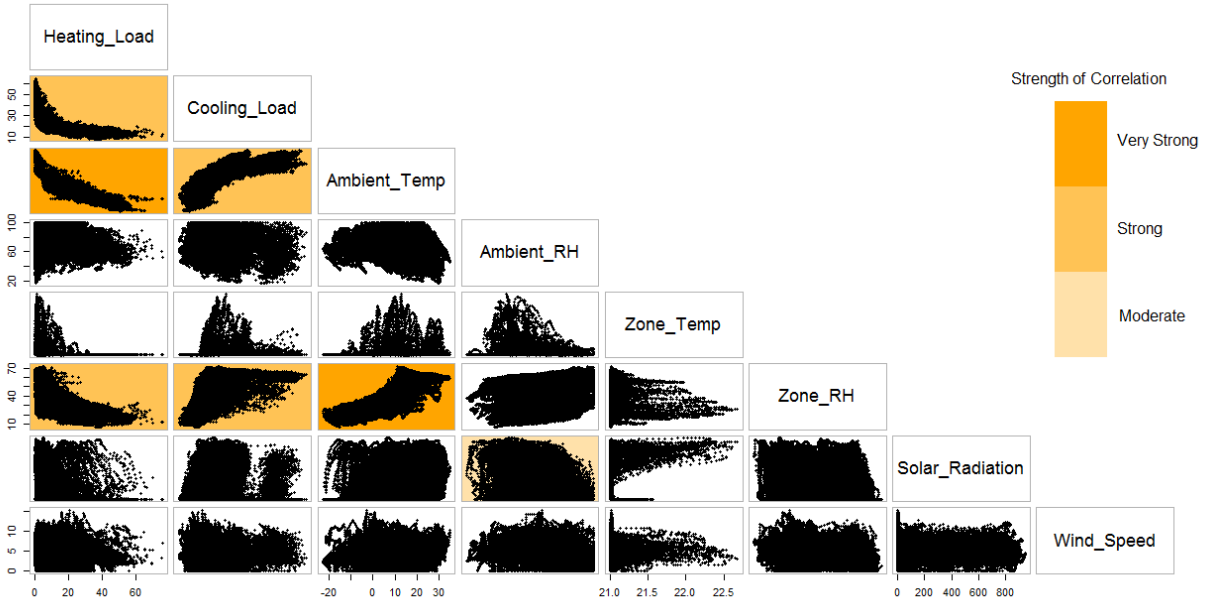


Figure 5: Linear Correlation Between Variables for the Large Hotel (B2) in Climate Zone 5A.

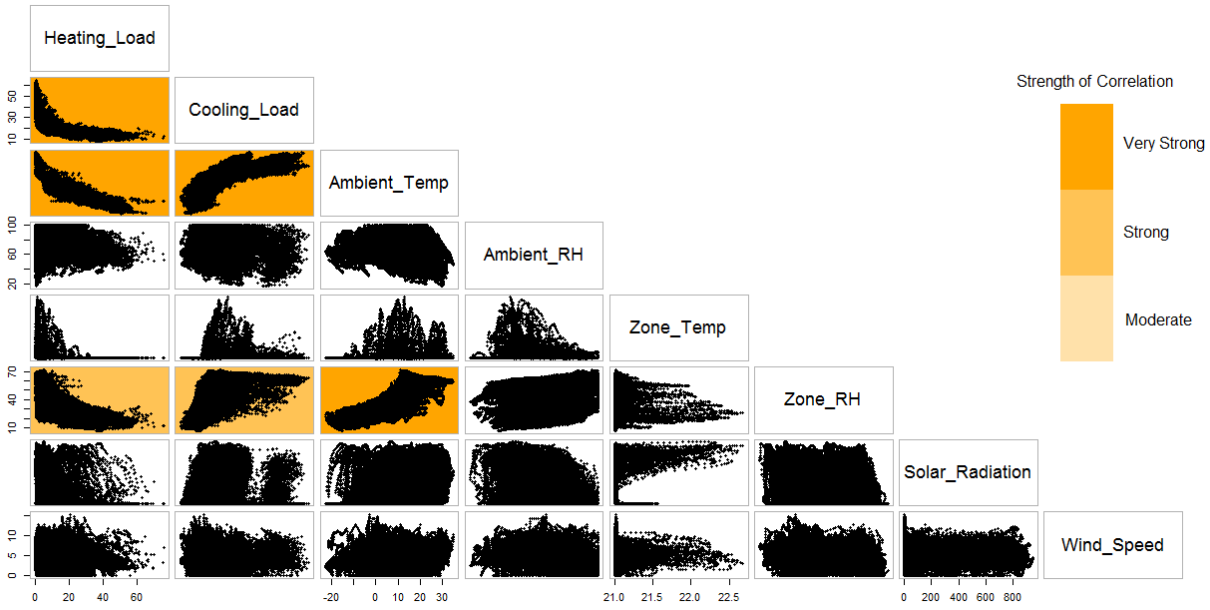


Figure 6: Monotonic Correlation Between Variables for the Large Hotel (B2) in Climate Zone 5A.

and 6, respectively. Linear correlation analysis indicates that Ambient Temperature has ”very strong” correlation with Heating Load and ”strong” correlation with Cooling Load, while Zone Relative Humidity has ”strong” correlation with both thermal loads. The monotonic correlation analysis verified the results from the linear analysis and unveiled

that the correlation between Ambient Temperature and Cooling Load is ”very strong”. In this particular case study results from both analyses yield that the input variables of a predictive model in climate zone 5A should be Ambient Temperature and Zone Relative Humidity.

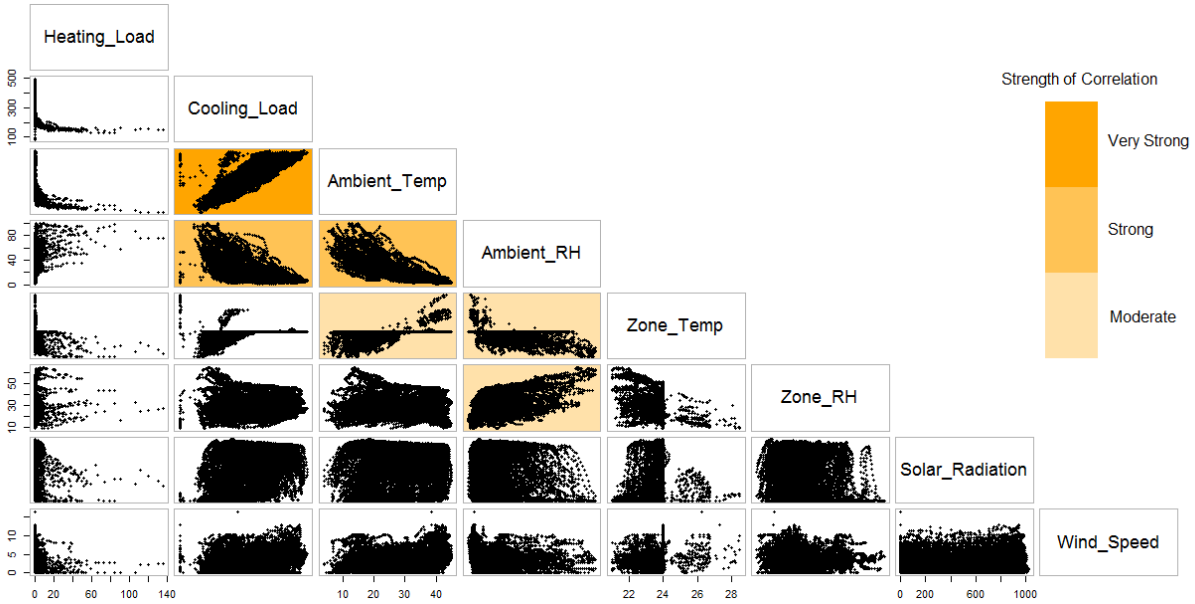


Figure 7: Linear Correlation Between Variables for the Large Office (B3) in Climate Zone 1B.

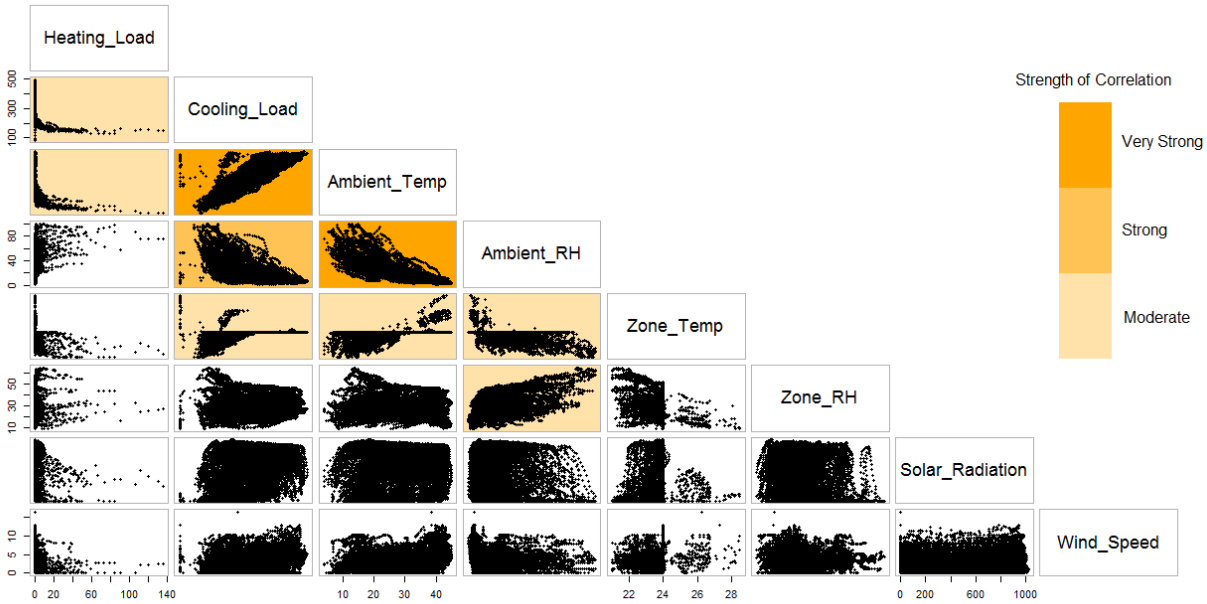


Figure 8: Monotonic Correlation Between Variables for the Large Office (B3) in Climate Zone 1B.

4.2.2. *Large Office (B3)*

The examination of Pearson and Spearman correlation analysis was repeated for the Large Office (B3) in the same climate zones. Figures 7 and 8 present the results from the linear and monotonic correlation analysis in climate zone 1B, respectively. Cooling Load is correlated with Ambient Temperature and Ambient Relative Humidity, based on the linear correlation analysis. Interestingly, Heating Load is not correlated to any of the possible input variables using Pearson analysis. Once again, the monotonic correlation analysis reveals input output correlated pairs that are not captured using the linear one. In this case, the correlation between Heating Load and Ambient Temperature, as well as between Cooling Load and Zone Air Temperature was discovered.

To predict the Heating Load for the Large Office in climate zone 1B, Ambient Temperature should be the only input variable. In regard to the prediction of Cooling Load, the input variables selected are Ambient Temperature, Ambient Relative Humidity and Zone Air Temperature.

The results for the same building in climate zone 5A are displayed in Figures 9 and 10. Pearson and Spearman correlation analysis indicate that Heating and Cooling Load are correlated with Ambient Temperature, Zone Air Temperature and Zone Relative Humidity. Therefore, Ambient and Zone Air Temperature along Zone Relative Humidity were identified as the input variables for predicting the thermal loads of the building, based on both types of analysis.

4.3. *Evaluation Results*

The performance of the ANN predictive models for the thermal loads of the specific case studies buildings was evaluated when all possible variables or only the selected ones were introduced as inputs. The accuracy of the ANN

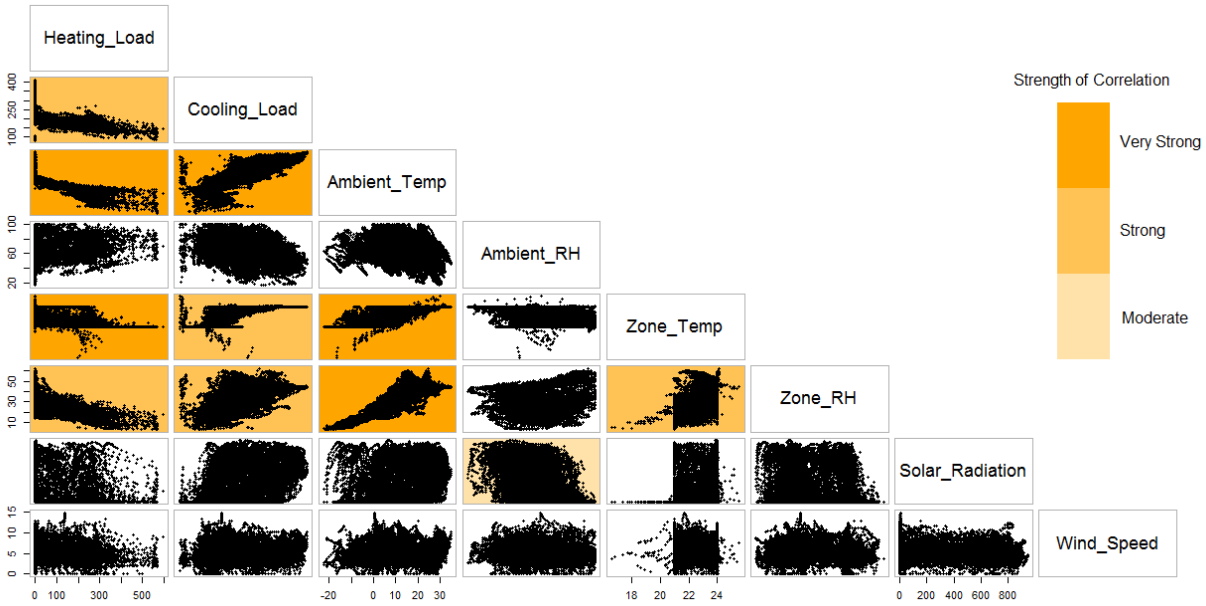


Figure 9: Linear Correlation Between Variables for the Large Office (B3) in Climate Zone 5A.

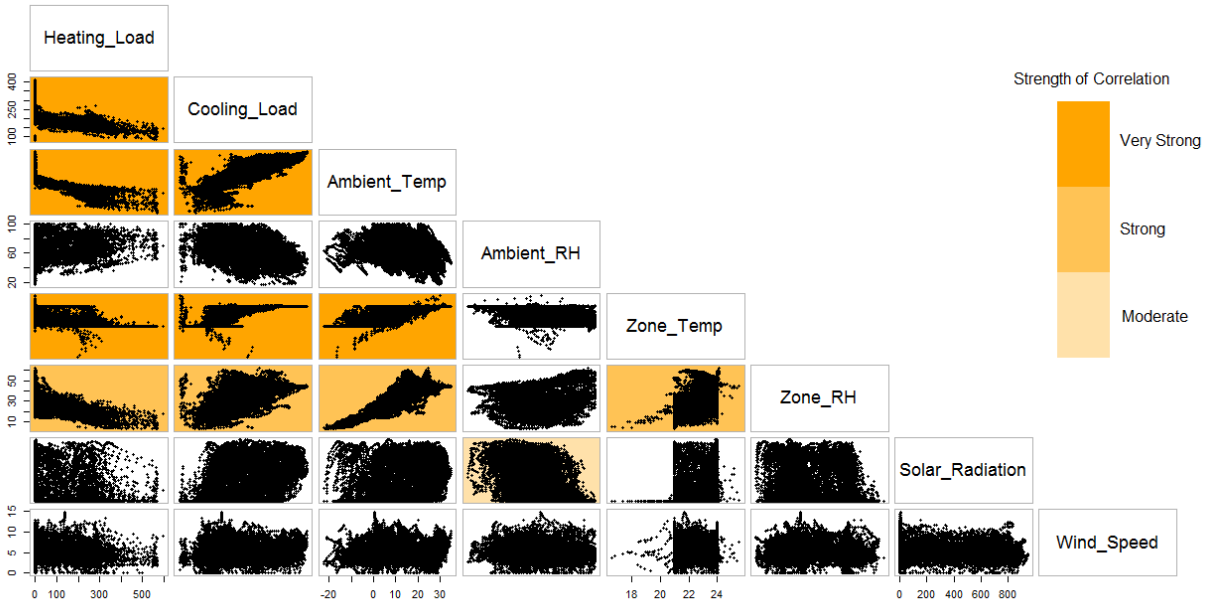


Figure 10: Monotonic Correlation Between Variables for the Large Office (B3) in Climate Zone 5A.

Table 7: RMSE (kWh) of ANN Models with All Possible Variables as Inputs

	Heating Load						**Cooling Load**					
	B1	B2	B3	B4	B5	B6	B1	B2	B3	B4	B5	B6
1A	0.91	0.41	0.87	0.43	0.06	1.14	1.06	1.70	8.42	3.40	0.86	0.61
1B	1.97	0.85	1.95	0.34	0.16	1.79	3.40	2.07	15.41	4.55	1.58	0.64
2A	1.34	0.81	7.02	3.48	0.41	2.00	1.68	1.73	11.37	4.30	1.27	0.87
2B	1.87	0.62	2.06	1.80	0.29	2.87	3.32	2.34	17.43	5.29	1.88	1.52
3A	1.68	1.29	9.72	3.99	0.49	2.42	2.16	1.65	13.09	5.15	1.04	0.48
3B	2.32	0.81	4.38	1.46	0.26	2.06	2.91	1.74	20.03	7.84	1.24	0.76
3C	1.94	1.21	1.63	3.00	0.12	1.52	1.09	1.27	9.46	11.02	0.93	0.33
4A	3.51	1.32	19.82	4.48	0.55	2.34	1.70	1.76	14.45	8.41	1.42	0.71
4B	2.82	0.73	3.27	1.76	0.30	2.21	2.59	1.84	24.73	6.63	1.24	0.87
4C	3.46	1.40	6.27	2.87	0.69	2.65	4.54	1.54	17.94	9.92	1.10	0.46
5A	4.74	1.77	14.48	7.23	0.90	2.94	1.66	1.59	17.34	6.59	1.13	0.79
5B	2.23	1.04	4.40	6.31	0.65	2.43	1.63	1.96	15.48	9.73	1.22	0.48
5C	4.79	1.22	5.51	6.16	0.73	2.35	2.02	1.28	12.97	9.18	0.86	0.43
6A	4.93	1.50	18.41	8.01	0.89	2.93	1.82	1.77	17.13	8.73	1.06	0.54
6B	4.82	1.41	18.49	6.80	0.83	2.72	2.31	1.69	17.19	9.91	1.07	0.78
7	6.42	2.06	18.43	10.28	1.31	2.79	1.67	1.58	15.90	8.66	1.02	0.64
8	11.41	2.90	20.05	17.92	1.98	3.22	1.82	2.19	13.89	8.83	0.91	0.55

predictive models when all possible variables were introduced as inputs is given in Table 7. Evaluation of the effect on the model accuracy was accomplished by calculating the difference between RMSE values obtained from the

predictive models developed with all or reduced input variables. In order to visualise this effect, a new colour code was implemented. Red and its gradients were used to capture the case studies in which the predictive models developed with all input variables had lower RMSE values than the ones developed with the reduced subset of the input variables, hence the predictive accuracy of the models was reduced. White indicates that the developed predictive models with all or a reduced subset of the input variables perform equally well. Green and its gradients were used to reflect the case studies in which the predictive models developed with the reduced subset of the input variables had lower RMSE values than the ones developed with all input variables, hence the predictive accuracy of the models was increased. Results from the comparison of the RMSE of the developed predictive models are presented in Figure 11.

The difference of the obtained RMSE values for all the case studies, between the predictive models developed with all input variables and the ones with the reduced subset, is plotted in Figure 12 for heating and cooling load. It is observed that the difference of the RMSE of the predictive models for 93% and 75% of the case studies, for heating and cooling load, respectively, was less than -1 kWh or even positive, when the reduced subset was used instead of all the input variables. As expected, there were some case studies where the RMSE of the predictive models was substantially increased when the reduced subset instead of all the input variables is utilised.

Table 8 summarises the case studies where the biggest differences of RMSE values were captured, along with the

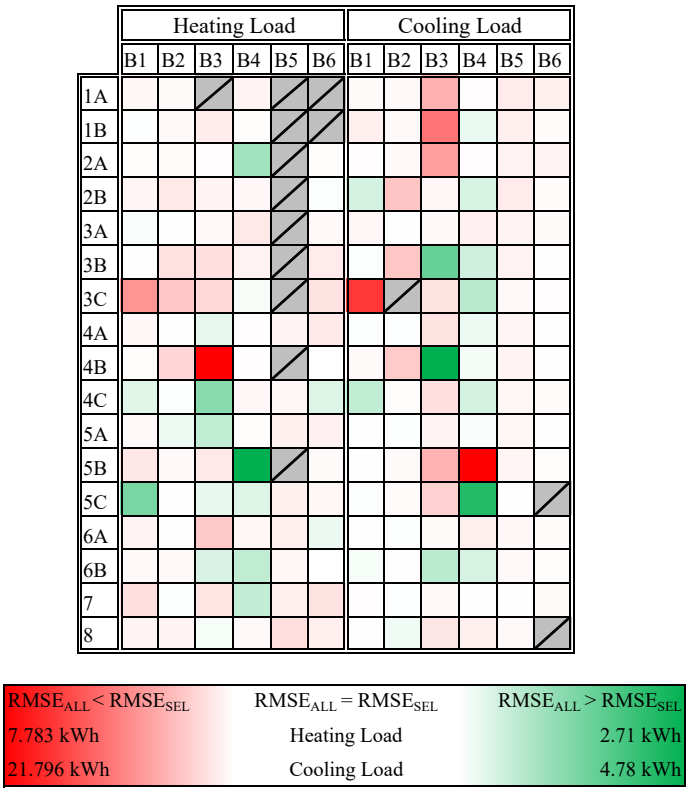
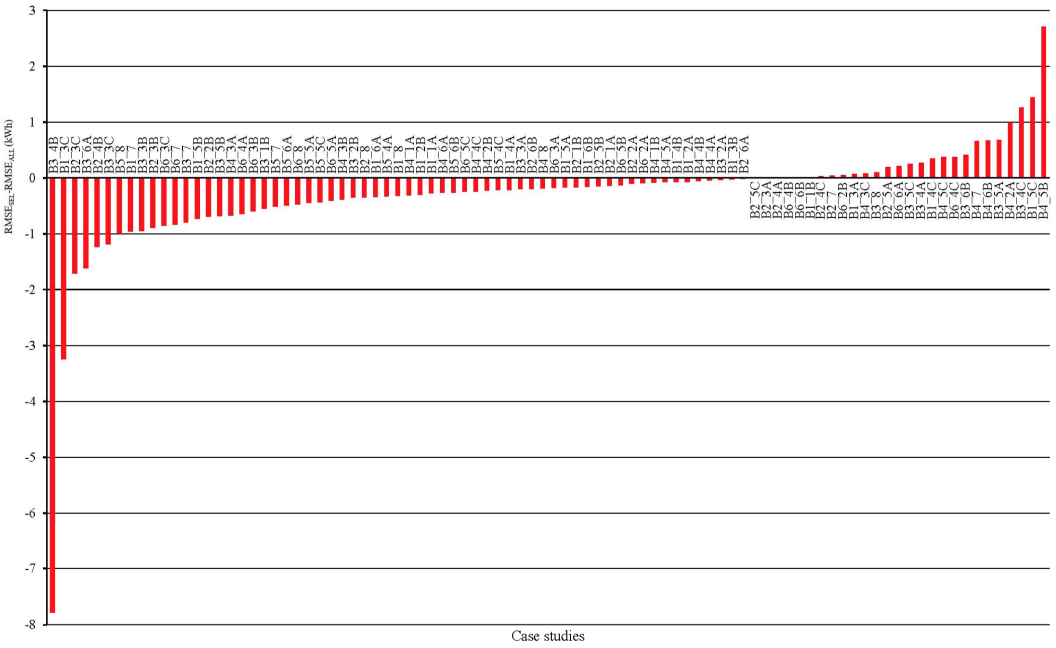
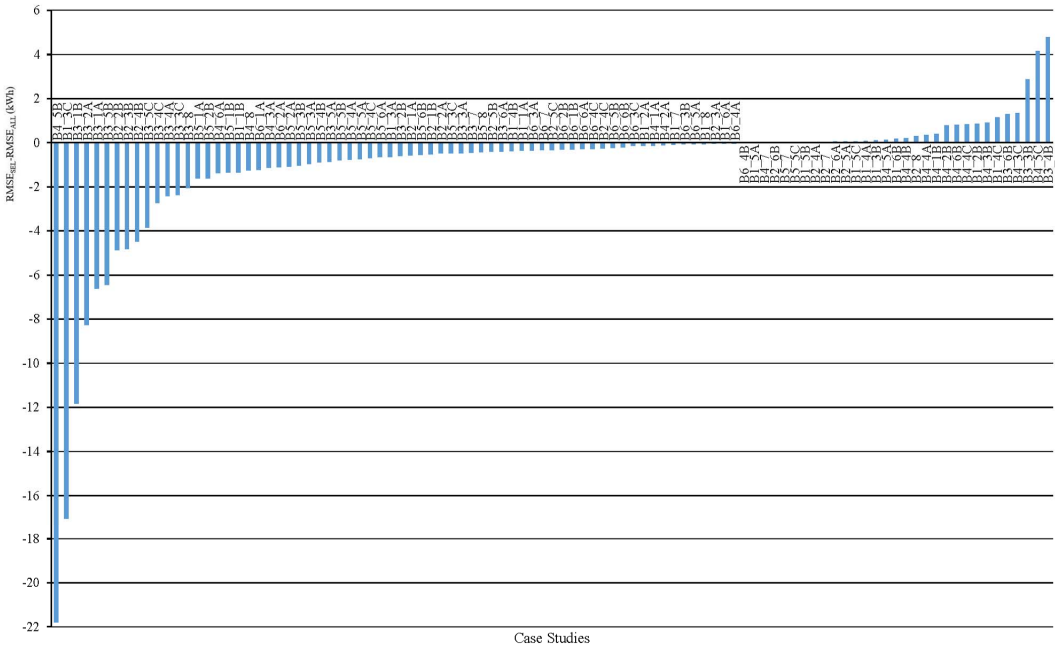


Figure 11: RMSE Comparison of All Against Selected Input Variables.



(a) Heating Load



(b) Cooling Load

Figure 12: Difference of RMSE between All and Selected Input Variables.

minimum, maximum and average thermal load, as well as their standard deviation for the 15-minute intervals of the testing partition (from July to December) of the synthetic database. The biggest increase of the RMSE (7.783 kWh)

Table 8: Case Studies with the Biggest Difference of RMSE

	Case Study	$RMSE_{ALL}$	$RMSE_{SEL}$	Minimum Load (kWh)	Maximum Load (kWh)	Average Load (kWh)	Standard Deviation
Heating Load	B3_4B	3.27	11.05	0	222.57	4.41	20.01
	B4_5B	6.31	3.60	0	59.58	1.73	7.08
Cooling Load	B4_5B	9.73	31.52	0	124.07	62.12	34.28
	B3_4B	24.73	19.95	79.30	418.83	295.24	70.65

was observed for the large office (B3) in climate zone 4B for heating, where the RMSE increased from 3.27 to 11.053 kWh when the reduced subset was used as input variables. For the models forecasting the cooling load, the biggest increase of the RMSE value (21.796 kWh) was captured for the secondary school (B4) in climate zone 5B, which increased from 9.732 to 31.528 kWh. On the contrary, there were also some case studies where the RMSE of the predictive models was decreased when the reduced subset of input variables was used, which are also shown in Table 8 (building B4 in climate zone 5B and building B3 in climate zone 4B).

In addition to the accuracy of the predictive models for each specific case study, the time to execute the ANN models was recorded as an indicator of their complexity. Figure 13, illustrates the difference in execution time of the

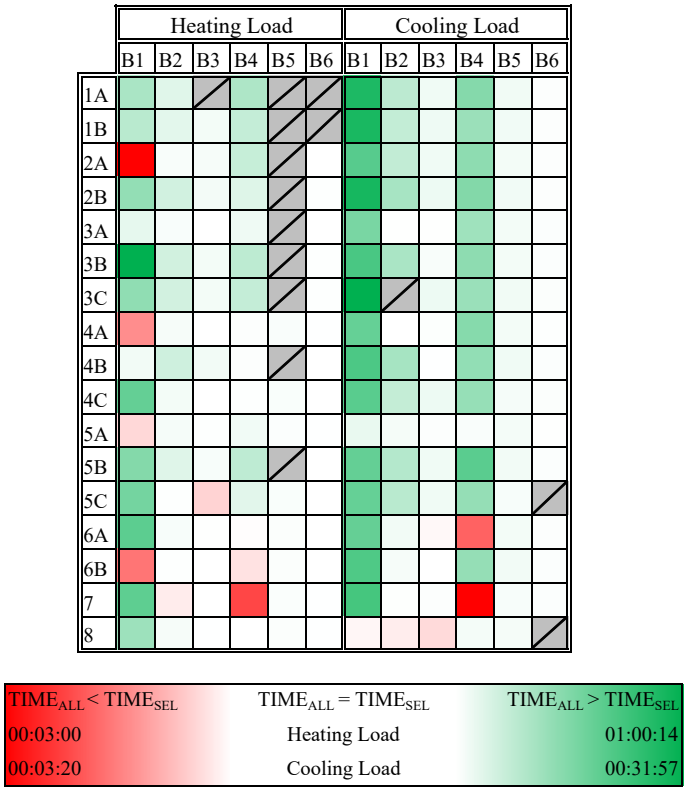


Figure 13: Execution Time Comparison of All Against Selected Input Variables.

predictive models when all possible variables or only the selected ones were introduced as inputs. A decrease at the execution time when the reduced subset was used instead of all input variables, was observed for 90% and 94% of the case studies, for the heating and cooling load, respectively, indicating a reduction in the complexity of the models.

The maximum increase of execution time was around three minutes, both for heating and cooling load, for the case studies where the predictive models developed with all input variables were executed faster than the ones developed with the reduced subset of the input variables. In particular, the execution time of the heating load predictive model for the hotel (B1) in climate zone 2A was increased from 33 to 36 minutes, while for the cooling load predictive model for the secondary school (B4) in climate zone 7 was increased from 14 minutes and 31 seconds to 17 minutes and 51 seconds. On the contrary, the biggest reductions of execution time were noticed for the hotel (B1), both for heating and cooling load, for the case studies where the predictive models developed with the reduced subset of the input variables were executed faster than the ones developed with all input variables. More specifically, the reduction of the execution time reached up to one hour, from one hour and 33 minutes to 33 minutes (building B1 in climate zone 3B) for heating and almost 32 minutes, from 32 minutes and 11 seconds to 14 seconds (building B1 in climate zone 3C) for cooling load.

4.4. Discussion

It can be seen that the input variables used in the data-driven predictive models correlate differently with thermal loads depending on building type and climate zone. It is highlighted that each case study is unique. However, the variation of the correlation with the climate zone could be interpreted for some of the environmental variables as a result of their almost constant nature. For instance, in all humid climates (climate zones 1A, 2A, 3A, 4A, 5A and 6A) it is noticed that Ambient Relative Humidity is not correlated with Heating Load for any of the case studies examined. This is an indication that since the Ambient Relative Humidity does not vary significantly and is maintained all year around at high levels in these climates, Heating Load is not affected by changes of this variable. Hence, it should not be introduced as an input to the predictive models because it does not offer any insight during their training process. Moreover, the variation of the correlation depending on the building type results from the different usage, inertia and HVAC operation of each reference building.

The relationship between Ambient Temperature and thermal loads was the most consistent one, since based on the Pearson and Spearman analysis for all 102 case studies, it was found to be present for the majority of the cases. Furthermore, the correlation between Zone Air Temperature and building thermal load was observed in more cases, when using the Spearman correlation analysis rather than the Pearson correlation analysis. Solar Radiation, Zone and Ambient Relative Humidity were correlated variously with heating and cooling loads, depending on each case. On the other hand, Wind Speed was not found to be correlated either linearly or monotonically with heating and cooling loads of the reference buildings in any climate zone. The variations of the correlation between input and output variables are illustrated with the additional examination of specific case studies. The diversity on the selection of input variables was emphasized using the Large Hotel and the Large Office buildings in very hot and cold climate zone.

Evaluating the influence of the selection of input variables implementing ANN predictive models, in order to forecast thermal loads of the specific case studies, it was evident that the majority of the models maintain the same level of prediction accuracy. In addition, it is captured the fact that the level of complexity is reduced when the selected variables are used as inputs due to less time required for the execution of the models.

5. Conclusions

The importance of performing both linear and monotonic correlation analysis has been illustrated both in the overall results (Section 4.1) and with the use of the specific case studies (Section 4.2). It is essential to select the input variables after examining both linear and monotonic correlation analysis results, to prevent the exclusion of some important input variables to the predictive models. Based on the assessment performed regarding the effect of the selection of input variables on the accuracy and complexity of the predictive models, it can be concluded that for the majority of the case studies, the accuracy is maintained at the same level and the complexity is reduced when the input variables are selected based on their correlation with the thermal loads.

To summarise, the selection of input variables for the predictive models should be made carefully and based on data analysis in order to avoid unnecessary additional complexity during the execution of the models. The proposed methodology can be easily implemented as a pre-processing step to help with the selection process of input variables prior to the development of the predictive models.

Future research work includes the application of the methodology to various types of predictive models, such as regression and SVM models. In this way, this methodology of selecting input variables will be investigated regarding its effectiveness on machine learning techniques in forecasting thermal loads of commercial buildings. The same reference building types in all possible climate zones will be used as case studies for the predictive models in order to maintain the consistency of the approach.

Aknowledgements

This research has been conducted with financial support from the Irish Research Council and United Technologies Research Centre (UTRC). The research was also financially assisted by the Electricity Research Centre, University College Dublin, which is supported by the Commission for Energy Regulation, Bord Gis Energy, Bord na Mona Energy, Cylon Controls, EirGrid, Electric Ireland, Energia, EPRI, ESB International, ESB Networks, Gaelectric, Intel and SSE Renewables.

References

[1] International Energy Agency, Transition to Sustainable Buildings, IEA, Paris, 2013.

[2] European Commission, Directive 2009/125/EC of the European Parliament and of the Council establishing a framework for the setting of ecodesign requirements for energy-related products (recast), Tech. rep., European Commission, Brussels (2009).

[3] European Commission, Directive 2010/30/EU of the European Parliament and of the Council on the indication by labelling and standard product information of the consumption of energy and other resources by energy-related products (recast), Tech. rep., European Commission, Brussels (2010).

[4] European Commission, Directive 2010/31/EU of the European Parliament and the Council on the Energy Performance of Buildings (recast), Tech. rep., European Commission, Brussels (2010).

[5] U.S. D.O.E., Advanced Energy Retrofit Guides, Office Buildings, Tech. rep., Pacific Northwest National Laboratory (2010).

[6] U.S. D.O.E., Operations & Maintenance Best Practices: A Guide to Achieving Operational Efficiency, Tech. rep., Pacific Northwest National Laboratory (2011).

[7] A. Kusiak, M. Li, Z. Zhang, A data-driven approach for steam load prediction in buildings, Applied Energy 87 (3) (2010) 925–933. doi:10.1016/j.apenergy.2009.09.004.

[8] ASHRAE, ASHRAE Handbook-Fundamentals, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, 2009, Ch. Energy Estimating and Modeling Methods, pp. 32.1–32.33.

[9] A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: A review, Renewable and Sustainable Energy Reviews 23 (2013) 272–288. doi:10.1016/j.rser.2013.03.004.

[10] S. S. Kwok, E. W. Lee, A study of the importance of occupancy to building cooling load in prediction by intelligent approach, Energy Conversion and Management 52 (7) (2011) 2555–2564. doi:10.1016/j.enconman.2011.02.002.

[11] A. Aranda, G. Ferreira, M. Mainar-Toledo, S. Scarpellini, E. L. Sastresa, Multiple regression models to predict the annual energy consumption in the Spanish banking sector, Energy and Buildings 49 (2012) 380–387. doi:10.1016/j.enbuild.2012.02.040.

[12] P. Ferreira, A. Ruano, S. Silva, E. Conceição, Neural networks based predictive control for thermal comfort and energy savings in public buildings, Energy and Buildings 55 (2012) 238–251, cool Roofs, Cool Pavements, Cool Cities, and Cool World. doi:10.1016/j.enbuild.2012.08.002.

[13] J. C. Lam, K. K. Wan, D. Liu, C. Tsang, Multiple regression models for energy use in air-conditioned office buildings in different climates, Energy Conversion and Management 51 (12) (2010) 2692–2697. doi:10.1016/j.enconman.2010.06.004.

[14] B. B. Ekici, U. T. Aksoy, Prediction of building energy consumption by using artificial neural networks, Advances in Engineering Software 40 (5) (2009) 356–362. doi:10.1016/j.advengsoft.2008.05.003.

[15] P. A. González, J. M. Zamarreño, Prediction of hourly energy consumption in buildings based on a feedback artificial neural network, Energy and Buildings 37 (6) (2005) 595–601. doi:10.1016/j.enbuild.2004.09.006.

[16] S. A. Kalogirou, Applications of artificial neural-networks for energy systems, Applied Energy 67 (12) (2000) 17–35. doi:10.1016/S0306-2619(00)00005-2.

[17] B. Dong, C. Cao, S. E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy and Buildings 37 (5) (2005) 545–553. doi:10.1016/j.enbuild.2004.09.009.

[18] Z. Hou, Z. Lian, An Application of Support Vector Machines in Cooling Load Prediction, in: Intelligent Systems and Applications, 2009. ISA 2009. International Workshop on, International Workshop on Intelligent Systems and Applications, Wuhan,, 2009, pp. 1–4. doi:10.1109/IWISA.2009.5072707.

[19] H. C. Jung, J. S. Kim, H. Heo, Prediction of building energy consumption using an improved real coded genetic algorithm based least squares support vector machine approach, Energy and Buildings 90 (2015) 76 – 84. doi:http://dx.doi.org/10.1016/j.enbuild.2014.12.029.

[20] D. Hsu, Identifying key variables and interactions in statistical models ofbuilding energy consumption using regularization, Energy 83 (2015) 144 – 155. doi:http://dx.doi.org/10.1016/j.energy.2015.02.008.

[21] W. J. Turner, O. Kinnane, B. Basu, Demand-side Characterization of the Smart City for Energy Modelling, Energy Procedia 62 (2014) 160–169, 6th International Conference on Sustainability in Energy and Buildings, SEB-14. doi:10.1016/j.egypro.2014.12.377.

[22] J. C. Lam, S. C. Hui, Sensitivity analysis of energy performance of office buildings, Building and Environment 31 (1) (1996) 27–39. doi:10.1016/0360-1323(95)00031-3.

[23] I. A. Macdonald, Quantifying the Effects of Uncertainty in Building Simulation, Ph.d. thesis, Department of Mechanical Engineering, Glasgow, UK (2002).

[24] J. Heller, M. Heater, M. Frankel, Sensitivity Analysis: Comparing the Impact of Design, Operation, and Tenant Behavior on Building Energy Performance, Tech. rep., Ecotype and the New Buildings Institute, Vancouver, (2011).

[25] T. Catalina, J. Virgone, E. Blanco, Development and validation of regression models to predict monthly heating demand for residential buildings, Energy and Buildings 40 (10) (2008) 1825–1832. doi:10.1016/j.enbuild.2008.04.001.

[26] T. Catalina, V. Iordache, B. Caracaleanu, Multiple regression model for fast prediction of the heating energy demand, Energy and Buildings 57 (2013) 302–312. doi:10.1016/j.enbuild.2012.11.010.

[27] S. A. Kalogirou, C. C. Neocleous, C. N. Schizas, Building heating load estimation using artificial neural networks, in: 17th International Conference on parallel architectures and compilation techniques, San Fransisco, California, US, 1997, pp. 1–8.

[28] S. Kalogirou, G. Florides, C. Neocleous, C. Schizas, Estimation of Daily Heating and Cooling Loads Using Artificial Neural Networks, in: 7th REHVA World Congress, Clima 2000, Napoli, 2001.

[29] J. Yang, H. Rivard, R. Zmeureanu, On-line building energy prediction using adaptive artificial neural networks, Energy and Buildings 37 (12) (2005) 1250–1259. doi:10.1016/j.enbuild.2005.02.005.

[30] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Applying support vector machine to predict hourly cooling load in the building, Applied Energy 86 (10) (2009) 2249–2256. doi:10.1016/j.apenergy.2008.11.035.

[31] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks, Energy Conversion and Management 50 (1) (2009) 90–96. doi:10.1016/j.enconman.2008.08.033.

[32] F. Lei, P. Hu, A Baseline Model for Office Building Energy Consumption in Hot Summer and Cold Winter Region, in: International Conference on Management and Service Science, 2009. MASS '09., IEEE, Wuhan,, 2009, pp. 1–4. doi:10.1109/ICMSS.2009.5301031.

[33] H.-X. Zhao, F. Magoulès, Feature selection for predicting building energy consumption based on statistical learning method, Journal of Algorithms & Computational Technology 6 (1) (2012) 59–78.

[34] J. Massana, C. Pous, L. Burgas, J. Melendez, J. Colomer, Short-term load forecasting in a non-residential building contrasting models and attributes, Energy and Buildings 92 (2015) 322 – 330. doi:http://dx.doi.org/10.1016/j.enbuild.2015.02.007.

[35] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[36] D. S. Kapetanakis, E. Mangina, D. P. Finn, Methodology for Commercial Buildings Thermal Loads Predictive Models Based on Simulation Performance, in: Proceedings of the 8th International Conference on Simulation Tools and Techniques, SIMUTools '15, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 2015, pp. 256–260. doi:10.4108/eai.24-8-2015.2261063.

[37] M. Deru, K. Field, D. Studer, K. Benne, B. Griffith, P. Torcellini, B. Liu, M. Halverson, D. Winiarski, M. Rosenberg, M. Yazdanian, J. Huang, D. Crawley, United States Department of Energy Commercial Reference Building Models of the National Building Stock, National Renewable Energy Laboratory, 2011.

[38] U.S. D.O.E., Commercial Prototype Building Models, `https://www.energycodes.gov/development/commercial/90.1_models`, accessed on 20.04.15 (2014).

[39] ANSI / ASHRAE, Energy Standard for Buildings Except Low-Rise Residential Buildings, ANSI/ASHRAE Standard 90.1-2007, ASHRAE, Atlanta, GA, USA (2007).

[40] M. Spiegel, J. Schiller, A. Srinivasan, Probability and Statistics, 4th Edition, McGraw Hill Professional, 2013.

[41] IBM Corp., IBM SPSS Statistics for Windows, `http://www-01.ibm.com/software/analytics/spss/products/statistics/` (2013).

[42] J. D. Evans, Straightforward statistics for the behavioral sciences, CA: Brooks/Cole Publishing, Editor: Pacific Grove, 1996.

[43] IBM Corp., IBM SPSS Modeler 14.2, `http://www-01.ibm.com/software/analytics/spss/products/modeler` (2011).