Forecasting Commercial Building Heating Loads:  
A Comparison Using Real and SyntheticDatasets

Kapetanakis D.S.#1, Mangina E.#2, Ridouane E.H.*3, Kouramas K.*4, Finn D.P.#5

#University College Dublin
Belfield, Dublin 4, Republic of Ireland

1dimitrios.kapetanakis@ucdconnect.ie, 2eleni.mangina@ucd.ie, 5donal.finn@ucd.ie

*United Technologies Research Center
Penrose Business Center, Cork City, Republic of Ireland
3ridouae@utrc.utc.com, 4kouramac@utrc.utc.com

Abstract
This paper is concerned with the development of data-driven predictive models to forecast heating loads in commercial buildings. These predictions can be used for enhanced control of heating ventilation and air-conditioning systems with the potential for more efficient operation and decreased energy consumption. A commercial building is employed in order to investigate the implications a lack of completeness and reliability in a real dataset can have on the accuracy of predictive models. The real dataset is obtained from the building energy management system installed in the building. A detailed representation of the building in EnergyPlus is used to generate a synthetic dataset, thereby providing a benchmark for this investigation. Data analysis of the real and synthetic datasets is performed to detect relationships between variables and select the input variables for the predictive models. Numerous predictive models are examined for their ability to forecast the building heating load when using the real and synthetic datasets. The most suitable model is selected by comparing the root mean square error of the predictions. The results indicate that the predictive models using the real dataset, in contrast with the synthetic dataset, are unable to generate highly accurate predictions.

Keywords – predictive models; heating load; real and synthetic data

1. Introduction

A maximum saving of 8% of the energy consumption in the EU is expected to be generated with the use of efficient energy management systems in buildings [1]. The enhancement of heating ventilation and air-conditioning (HVAC) control, particularly in commercial buildings, is one of the most cost-effective approaches to decrease the energy usage and increase compliance with the European Directives on the energy performance of buildings [2]. An accurate and easily implemented methodology for assessing building thermal loads is a useful and increasingly important first step, before enhanced control strategies can be considered. The prediction of thermal load of HVAC systems
is important for energy management especially during peak energy demand hours [3].

Estimations of building heating load can be generated using appropriate simulation software [4] when detailed data such as building geometry, occupancy as well as environmental variables are available. In reality, such data are often not readily available, especially for older buildings, where uncertainty arising from parameter and occupancy estimation can lead to significant additional modelling efforts [5]. An alternative way to forecast these loads is to take advantage of building energy management (BEM) systems recorded data. These data records include underlying information regarding building thermal response and can be introduced to data-mining models, which utilise extensive assessment of input and output variables, in order to produce accurate predictions [6].

The main objectives of the current paper are to describe an approach of selecting useful information from BEM systems data in order to be used as inputs in predictive models, as well as the development of predictive models of heating loads for commercial buildings, using both real and synthetic datasets. The input variables under examination should be recorded and easily obtained from the BEM systems or any other source for a long period of time. Hence, weather and indoor conditioning data are investigated as possible input variables. A commercial building, located in Cork, Ireland is utilised as a testbed building for the development stage of the predictive models. Data are extracted on a 15-minute time step basis in order to analyse them for the existence of possible correlations between variables. In practice, incomplete datasets are common when extracting data from BEM systems, which leads to challenges when developing predictive models. Pre-analysis of BEM systems datasets in order to evaluate their completeness, is a prerequisite in the context of developing data-driven models. Rates of less than 1% missing data are generally considered trivial, 1-5% manageable. However, 5-15% requires sophisticated methods to handle, and more than 15% may severely impact any kind of interpretation [7]. In order to investigate the influence that missing data have on the accuracy of the predictive models, a synthetic dataset with no missing values is implemented. This dataset is obtained from an EnergyPlus [8] building energy simulation model of the testbed building.

Forecasting building energy performance for a city or even at a national scale, leads to long processing time when developing such models, therefore reducing model complexity is an important issue. The purpose of the data mining process is to assist with building heating load prediction and in particular with the issue of selecting input variables. This selection is critical for the construction of the predictive model, since redundant input variables introduce unnecessary increases in the complexity during the development of the models. Following the identification of the input variables, various predictive models are developed by implementing both regression-based and machine learning methods, using both real and synthetic datasets.
2. Background

Regression and artificial neural networks (ANN) are amongst the most common methods used in the literature to achieve forecasting of building thermal loads without the use of simulation software. However, during the development of these data-driven 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.

Catalina et al. [9] worked on the development of regression models to predict monthly heating demand for the single-family residential sector. The inputs for the regression models were building shape factor, envelope U-value, window to floor area ratio, building time constant and climate. The average error was 2% between the predicted and simulated values. Aranda et al. [10] used linear regression models to predict the annual energy consumption in the Spanish banking sector. The energy consumption of a single building was predicted as a function of its construction characteristics, climatic area and energy performance.

ANNs have been applied to analyse various types of building energy consumption, as well as heating loads. Gonzalez and Zamarreno [11] used an ANN approach to predict the hourly energy consumption in buildings. The inputs of the network were current and forecasted values of temperature, the current load, the hour and the day. Kalogirou et al. [12, 13] implemented ANN at an early design stage to predict the required heating load of buildings. Input data included the areas of windows, walls and floors, the type of windows and walls, roof classification and the room temperature. The relative error of the network was 3.5%.

The research field related to building thermal loads forecasting has been very active, involving various regression and data mining techniques. Energy prediction of buildings has been the primary focus of numerous journal and conference publications, as well as of various competitions organised by the American society of heating, refrigeration and air-conditioning engineers (ASHRAE). Nevertheless, it is clear from the literature that little attention has been given to the justification of the selection of the input variables to their predictive models. Hence, the investigation of which variables and why they should be considered as inputs to such a model should be a priority towards the development stage of the model.

3. Methodology

The methodology developed in the current work is based on the selection of input variables and determination of predictive models able to forecast heating loads of a commercial testbed building, utilising both real and synthetic datasets. In order to achieve this, the sequence presented below is followed:
i. Acquisition of real dataset with measurements from the testbed building
ii. Generation of synthetic dataset using an EnergyPlus model of the testbed Building
iii. Investigation of linear correlation between variables by calculating the Pearson correlation coefficient
iv. Investigation of monotonic correlation between variables by calculating the Spearman correlation coefficient
v. Selection of input variables
vi. Development of predictive models.

3.1 Testbed Building Description

The NIMBUS building [14] is a two storey quadrangle-type office building, used for research purposes, hosting researchers and students from Cork Institute of Technology (CIT) located in Cork, Ireland.

The heating system of the building incorporates a combined heat and power (CHP) unit, two gas boilers and a water calorifier. An extensive network of meters and sensors has been deployed to facilitate measurement and necessary data collection for the control and monitoring of HVAC systems. These measurements together with relevant information about gas and electricity power consumption measurements, as well as heating/electrical loads and weather forecast, are available from the BEM systems.

3.2 Real and Synthetic Datasets

The initial task of the acquisition of real dataset process is the selection of BEM systems variables to be assessed. All of the variables are selected based on the sensors already installed at the testbed building. These variables are divided into two categories, input and output variables. Inputs are the ones introduced to the predictive model and output is the heating load, which will be forecasted from the models. The chosen as possible input variables are grouped according to weather data and indoor variables, as given in Table 1. The actual measurements from the BEM system where recorded at 1-minute intervals.

<table>
<thead>
<tr>
<th>Weather Data</th>
<th>Indoor Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambient Temperature</td>
<td>Zone Air Temperature</td>
</tr>
<tr>
<td>Ambient Relative Humidity</td>
<td></td>
</tr>
<tr>
<td>Wind Speed</td>
<td></td>
</tr>
<tr>
<td>Solar Radiation</td>
<td></td>
</tr>
</tbody>
</table>

Data analysis is performed to real data recorded from the BEM system and to synthetic data obtained by the EnergyPlus model of the testbed building. The EnergyPlus model provides one year of simulated data at 15-minute intervals. Therefore, real data are aggregated every 15 minutes to maintain the
resolution of real and synthetic datasets at the same level. The implementation of widely-known statistical techniques, which reveals the existence and importance of correlations with the output variables, is used for validating the selection of input variables.

### 3.3 Selection of Input Variables

The procedure followed for analysing the data and examining the existence of a correlation between input and output variables is based on statistical techniques. Initially, the existence of a linear correlation between input and output variables is investigated by performing a Pearson correlation analysis [15], which measures the linearity between paired data. In a sample of data, the Pearson coefficient is denoted by $r_P$ and is constrained between -1 and +1, illustrating the relationship between two continuous variables. A correlation coefficient of 0 indicates that there is no relationship, either positive or negative. A correlation coefficient of +1 denotes that there is a perfect positive correlation. In the case of +1, as one variable increases or decreases, the second variable increases or decreases, respectively, in exactly the same proportion. A correlation coefficient of -1 reveals that there is a perfect negative correlation [16].

The correlation coefficient does not indicate the relative gradient beyond sharing its positive or negative sign. The strength of correlation can be verbally described using the following guide for the absolute value of $r_P$ where:

- 0.00 - 0.19 “very weak”
- 0.20 - 0.39 “weak”
- 0.40 - 0.59 “moderate”
- 0.60 - 0.79 “strong”
- 0.80 - 1.00 “very strong”

Furthermore, the existence of monotonic relationships can be identified as well through the Spearman correlation coefficient [15]. A monotonic relationship is one that the dependent variable either never increases or never decreases as the independent variable increases. The Spearman correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data and is denoted by $r_S$. It is also robust to outliers, unlike Pearson correlation coefficient. The principles of Spearman correlation are the same as the Pearson correlation coefficient.

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. To avoid unnecessary increases in the complexity during the development of the models it is important to select the input variables with caution.
3.3 Predictive Models Development

The development of the predictive models for the testbed building is initiated with the investigation of regression models followed by the examination of ANN models. To investigate all possible scenarios, different types of regression and ANN models are developed using real and synthetic datasets, different partitioning methods as well as different input variables, as given in Table 2.

Three types of regression models, multiple linear, multiple non-linear and generalized linear are implemented in the development stage of the regression models. Moreover, two types of ANN methods, multilayer perceptron and radial basis are considered for the development of machine learning models.

Furthermore, the real and synthetic datasets acquired from the testbed building are divided into training and testing partitions. The training partition is used to develop and train the predictive models, while the testing partition is utilised for the evaluation of the accuracy of the models. The generated predictions are compared with the actual values of the testing partition. Two ways of separating the datasets are examined. The first is that the training partition includes data only from October, November and December and the second is that training consists of all data from the first fifteen days of each month of the heating season.

The input variables used in multiple combinations are Ambient Temperature, Ambient Relative Humidity, Solar Radiation and Zone Air Temperature.

Once the models are developed the accuracy of each one was measured by calculating the root mean square error (RMSE), as shown in Equation 1, where \( y \) are the actual values and \( \hat{y} \) are the predicted values.:

\[
RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}
\]  

4. Results and Discussion

The primary steps of the methodology were the acquisition of real and generation of a synthetic dataset from the testbed building. A complete year of measured data on a 15-minute time step basis, from May 2014 to May 2015,
was extracted from the BEM system of the testbed building. Utilising the EnergyPlus model one year of simulated data at 15-minute intervals was generated.

Considering that the main objective is the prediction of heating loads, both real and synthetic datasets were reduced in order to include only the heating season of the building. Data analysis of the obtained real dataset indicated that the heating season is from 7\textsuperscript{th} of October 2014 to 8\textsuperscript{th} of April 2015. Furthermore, due to the office usage of the building, weekends and public holidays were excluded from the dataset. In addition, the data analysis was performed only for the hours when the HVAC system provided heating to the building, hence from 07.15 am to 17.00 pm. The obtained real dataset contained periods of time where the measurements were interrupted, either due to maintenance of the system or malfunction. Calculation of the percentage of the missing data was necessary, in order to determine an appropriate way to deal this missingness [7]. The missing data of the real dataset for the heating season were identified to be 0.94\% of the dataset, therefore completely excluded from the analysis. The same limitations for the heating season, the working HVAC system hours and the exclusion of weekends and public holidays were applied at the synthetic dataset as well to keep consistent between real and synthetic data.

4.1 Input Variables Selection

The next step was the examination of linear correlation between input and output variables using the Pearson correlation analysis and the results are shown in Table 3. According to the Pearson analysis, Ambient Temperature and Zone Air Temperature were the only input variables that are correlated with Heating Load, both for real and synthetic datasets. The Pearson coefficients indicate that the correlation between Ambient Relative Humidity and Solar Radiation with the Heating Load was “very weak” for the real dataset and “weak” for the synthetic one. Moreover, the correlation between Heating Load and Wind Speed was “very weak” when examining both datasets.

Table 3. Pearson correlation results for Real and Synthetic data

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Real Data</th>
<th>Synthetic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambient Temperature</td>
<td>Heating Load</td>
<td>moderate</td>
<td>moderate</td>
</tr>
<tr>
<td>Ambient Rel. Humidity</td>
<td>Heating Load</td>
<td>very weak</td>
<td>weak</td>
</tr>
<tr>
<td>Solar Radiation</td>
<td>Heating Load</td>
<td>very weak</td>
<td>weak</td>
</tr>
<tr>
<td>Wind Speed</td>
<td>Heating Load</td>
<td>very weak</td>
<td>very weak</td>
</tr>
<tr>
<td>Zone Air Temperature</td>
<td>Heating Load</td>
<td>weak to moderate</td>
<td>strong to very strong</td>
</tr>
</tbody>
</table>

Subsequently, a Spearman correlation analysis was performed to investigate the monotonic relationships between input and output variables. The outcome of the analysis is displayed in Table 4. Based on the Spearman
analysis, the existence of correlation between variables that had been discovered with Pearson coefficient is verified. Table 4 reveals that the correlation between Heating Load and Ambient Temperature can be better described as a monotonic one, since results yield a “strong” correlation instead of “moderate” when using the synthetic dataset.

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Real Data</th>
<th>Synthetic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambient Temperature</td>
<td>Heating Load</td>
<td>moderate</td>
<td>strong</td>
</tr>
<tr>
<td>Ambient Rel. Humidity</td>
<td>Heating Load</td>
<td>very weak</td>
<td>weak</td>
</tr>
<tr>
<td>Solar Radiation</td>
<td>Heating Load</td>
<td>very weak</td>
<td>weak</td>
</tr>
<tr>
<td>Wind Speed</td>
<td>Heating Load</td>
<td>very weak</td>
<td>very weak</td>
</tr>
<tr>
<td>Zone Air Temperature</td>
<td>Heating Load</td>
<td>weak to moderate</td>
<td>strong to very strong</td>
</tr>
</tbody>
</table>

Table 4. Spearman correlation results for Real and Synthetic data

Through this analysis, two input variables that are correlated with the Heating Load were identified. These are Ambient Temperature and Zone Air Temperature, both when real or synthetic data were used. Therefore, a configuration of the predictive models using as inputs only the Ambient and the Zone Air Temperature is introduced in order to verify the importance of the input variables selection process.

4.2 Predictive Models

4.2.1 General Results

Combination of all possible scenarios in Table 2 led to the development of 60 regression and 40 ANN models. The average RMSE of the predictive models using real data is 23.999 and of the models using synthetic data 9.881.

The most accurate model using real data was found to have a RMSE value of 19.389 when generating predictions for the heating load of the NIMBUS building, while the least accurate had 30.645. Regarding the developed models using synthetic data, the RMSE value for the most accurate was 3.812 and the least accurate 20.520. The most accurate model using the real dataset, used as inputs the Ambient Temperature, the Ambient Relative Humidity and the Zone Air Temperature, while using the synthetic dataset the inputs where Ambient and Zone Air Temperature. Furthermore, the ANN multilayer perceptron method and the first fifteen days of each month of the heating season as training partition were implemented for the development of the most accurate predictive models using real and synthetic data.

4.2.2 Regression VS ANN Models

The results obtained from the predictive models were assessed to examine possible influences from the different regression based and machine learning methods as well as training partitions used. The assessment indicated that ANN based predictive models reduce by almost 24% the RMSE of the
predictions when using synthetic data, as depicted in Figure 1. In addition, the average RMSE remains at the same level when using real data regardless of the type of the predictive method applied.

![Fig. 1 Average RMSE of Real and Synthetic datasets models using Regression and ANN](image1)

![Fig. 2 Average RMSE of Real and Synthetic datasets models using different partitions](image2)

### 4.2.3 Different Training Partitions

The average RMSE of the predictive models decreases, both for real and synthetic datasets, when the training partition consists from the first fifteen days of each month of the heating season, as displayed in Figure 2.

### 4.2.4 Different Input Variables

The importance of the input variables selection method was investigated in order to verify its impact to the accuracy of the predictive models. The results are presented in Figure 3. The RMSE of the models which use real data, with Ambient Temperature and Zone Air Temperature as input variables, is 23.87. For all other combinations of input variables, the average RMSE is 24.03. Similarly, the RMSE of the models based on synthetic data with input variables only Ambient Temperature and Zone Air Temperature is 7.54, whereas with all other combinations of input variables the average RMSE is 10.46.

Another interesting finding during the assessment of the obtained results was the influence of the Zone Air Temperature as an input variable, which is highlighted in Figure 4. The average RMSE of the models using real data without Zone Air Temperature as an input is 23.41, while when it is present as an input the average RMSE reduces to 23. Likewise, the average RMSE of the models using synthetic data without Zone Air Temperature as an input is 18.86, whereas when it is present as an input the average RMSE decreases to 7.63. This captures the importance of the Zone Air Temperature as an input variable, since it provides useful information to the predictive model regarding the indoor conditions and when coupled with weather data significantly improves the accuracy of the predictions.
5. Conclusions

It is clearly shown that the variables that affect the Heating Load, are Ambient Temperature and Zone Air Temperature for this particular testbed building. This is verified both from the data analysis performed for the input variables selection and the results obtained from the developed models. The predictive models using real data were unable to generate highly accurate predictions. Further assessment is required in order to identify the reasons for that. A first assumption regarding the reason could be related to the fact that the testbed building is used for various research purposes, hence the quality of the data might not be equivalent to a typical office building. On the contrary, the accuracy achieved with the predictive models using synthetic data is maintained at a high level. Moreover, regarding the predictive models, results indicate that models using as training partition the first fifteen days of each month of the heating season, can achieve higher accuracy in predictions, compared to the ones using only data from October, November and December.

Additionally, models using as input variables the Ambient and Zone Air Temperature performed better in forecasting the Heating Load. Especially Zone Air Temperature proved to be a critical variable for the generation of more accurate predictions. Future research work includes the application of the methodology regarding the selection of input variables and the predictive models to different types of commercial buildings in various climates. In this way, the existence of a pattern related to the selection will be investigated and the effectiveness of regression based and machine learning techniques in forecasting heating loads of commercial buildings will be further examined.

Acknowledgment

This research has been conducted with financial support from the Irish Research Council, the United Technologies Research Centre (UTRC) and the Electricity Research Centre, University College Dublin.
References


