Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

Improving the accuracy of energy baseline models for commercial buildings with occupancy data

Xin Liang[1,2], Tianzhen Hong[2,*], Geoffrey Qiping Shen[3]

Abstract

More than 80% of energy is consumed during operation phase of a building's life cycle, so energy efficiency retrofit for existing buildings is considered a promising way to reduce energy use in buildings. The investment strategies of retrofit depend on the ability to quantify energy savings by "measurement and verification" (M&V), which compares actual energy consumption to how much energy would have been used without retrofit (called the "baseline" of energy use). Although numerous models exist for predicting baseline of energy use, a critical limitation is that occupancy has not been included as a variable. However, occupancy rate is essential for energy consumption and was emphasized by previous studies. This study develops a new baseline model which is built upon the Lawrence Berkeley National Laboratory (LBNL) model but includes the use of building occupancy data. The study also proposes metrics to quantify the accuracy of prediction and the impacts of variables. However, the results show that including occupancy data does not significantly improve the accuracy of the baseline model, especially for HVAC load. The reasons are discussed further. In addition, sensitivity analysis is conducted to show the influence of parameters in baseline models. The results from this study can help us understand the influence of occupancy on energy use, improve energy baseline prediction by including the occupancy factor, reduce risks of M&V and facilitate investment strategies of energy efficiency retrofit.

Keywords: Baseline Model; Occupancy; Building Energy Use; Measurement and Verification; Energy Efficiency Retrofit.

[1] School of International and Public Affairs, Shanghai Jiao Tong University, Shanghai, China. Email: liangxinpku@gmail.com;

[2] Building Technology and Urban Systems Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[3] Department of Building and Real Estate, Hong Kong Polytechnic University, Hong Kong, China.

* Corresponding author. Email: thong@lbl.gov, phone: 1 (510) 486-7082

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

1 Introduction

The buildings sector consumes 40% of the total primary energy in the United States [1], and the consumption has continued to increase, particularly in developing countries [2]. The buildings sector is thus responsible for a quarter of the total global greenhouse gas (GHG) emissions [3], and this proportion can reach around 50% in the United States (U.S.) with adverse impact on global environment, healthcare, and economy [4]. Furthermore, in the life cycle of a building, more than 80% of energy consumption occurs during the actual operation stage, rather than the construction stage [5]. Therefore, improving the energy efficiency of existing buildings is a key issue for reducing the total energy consumption and GHG emissions.

Energy efficiency retrofit for existing buildings is considered a promising method to achieve the target of energy savings [6]. Numerous previous studies indicated energy retrofit can significantly benefit the environment, society, and economy by improving energy efficiency [6,7], reducing emissions [8,9], controlling resource usage [10], enhancing the reputation of building owners [11], improving the health and productivity of occupants [12,13], reducing operation costs [14], increasing rent and occupancy rates [15,16], and creating job opportunities [17].

Owing to the significant benefits on energy conservation and other aspects of society, energy efficiency retrofit has been emphasized around the world. For example, the U.S. government passed the Energy Policy Act (EPA) of 2005 and Executive Order (EO) 13423, which require that 15% of the total number of existing buildings should be retrofitted to improve energy efficiency by 2015 compared with the 2003 baseline. Approximately 30 billion US dollars are assigned to conduct energy efficiency retrofit of existing buildings and facilities [7]. Incented by the policies, the market to provide energy efficiency services through energy service companies (ESCOs) has been blooming in the last decade [8].

Energy performance contracting (EPC), which is a financing package provided by ESCOs, is a commonly used market mechanism to implement energy efficiency retrofit. EPC includes energy savings guarantees and associated design, implementation and operation services [2, 9]. The profit (or the payment to ESCOs) of an EPC is mainly from the amount of energy cost savings after retrofit. The energy savings can be defined as the difference between how much energy the building consumed after retrofit, and how much it *would have* consumed without the retrofit. The former can be obtained from utility meters, and the latter, which is referred to as the energy use "baseline", is not measurable but can only be obtained by prediction. The accuracy of the baseline prediction can significantly impact the energy saving assessment, investment return and payback period. Furtherly, it can likewise impact the investment strategies and development of the building retrofit market.

The whole process of predicting baseline and assessing energy saving is called "measurement and verification" (M&V) [10]. The mechanism of M&V approaches is first monitoring the energy use of buildings, then developing mathematical models trained by observed data, and finally predicting baseline of energy use based on the developed models. Xia and Zhang [10] present a mathematical description of the M&V problem and cast a scientific framework for the basic M&V concepts, propositions, techniques and methodologies. Mathieu, et al. [11] proposed a regression-based model to predict baseline electricity consumption of commercial buildings and industrial facilities. Coughlin, et al. [12] evaluated the performance of three average-based models for baseline. Granderson, et al. [13] proposed an automated M&V method for evaluating model performance. Granderson and Price [14] summarized five baseline models, including both average-based models and regression-based models, and compared the predictive accuracy of these models with several metrics. More complex mathematical models of M&V have been emerged, including multivariate regression models, exponential smoothing models, neural network models, and Fourier series models [15-18].

Uncertainty of M&V models is important, since not only the value of the baseline, but also the accuracy and reliability of the prediction are critical to energy efficiency retrofit. It provides the stakeholders (e.g., ESCOs, building owners, facility managers) the information of investment risk, which is critical in decision making. For example, if the post-retrofit energy use will be 30% lower than the baseline, but the uncertainty exceeds 30%, it is then very risky to invest in this retrofit project. Walter, et al. [8] emphasized the influence of uncertainty and assessed uncertainty of M&V for 17 buildings by calculating the percent differences between predicted baseline and observed data. The results showed there was considerable uncertainty in baseline prediction: 5 out of 17 buildings had more than 20% uncertainty, and in an extreme case it was more than 60%.

The occupancy rate is a key uncertainty factor of M&V. Numerous previous studies indicated that the occupancy rate had significantly positive correlation with the energy use in buildings [19-26]. Occupants in buildings influence energy use in buildings in three major ways [27]: (1) sensible and latent heat gains from occupants, (2) occupants' need of thermal comfort, visual comfort and indoor air quality, and (3) occupant behavior and interactions with building systems and controls [28, 29]. In addition, in commercial buildings, the occupancy rate may increase after energy retrofit, due to lower utility bills, better indoor environment and higher social reputation [30-32]. Miller, et al. [33] indicated the office buildings with green features will have 2-4% occupancy rate premium. Wiley, et al. [34] specified that the office buildings with LEED certification will increase up to 16-18%. Therefore, if the occupancy rate is changed after energy retrofit, the baseline of energy use should be adjusted.

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

Although a number of previous studies emphasized the importance of occupancy factor in predicting the baseline, few studies, if not none, used occupancy factor in baseline prediction models, probably due to the highly stochastic activities and data limitation. Therefore, several research questions related to M&V remain to be answered: Does occupancy rate significantly impact the accuracy of baseline prediction? If yes, how to quantitatively evaluate the impact on prediction accuracy? How is the influence of occupancy on baseline models compared to that of other impact factors (e.g., outdoor air temperature, day of week), stronger, weaker or equal? Is it feasible to improve prediction accuracy of energy baseline by using occupancy data? Nowadays, most commercial buildings have access control system, which can obtain occupancy data in short time intervals. These data provide a new opportunity to deeply analyze the impact of occupancy on the accuracy of baseline prediction.

To address the aforementioned questions, this study proposes a novel method to quantitatively evaluate how accuracy of energy baseline models is improved by including the occupancy factor. Different from previous models, the proposed model of this study considers the occupancy data as independent variables rather than external uncertainty, shown in Figure 1. Although influence of occupancy has been emphasized by numerous previous studies, traditional models have not included occupancy data in the functions of energy prediction. Therefore, in traditional models, occupancy is an external uncertain factor, which can negatively impact the accuracy of energy prediction. Contrarily, in this study, the occupancy data is considered as an independent variable so that the influence of occupancy can be fitted by the function and evaluated by the prediction results. The results of this study showed the accuracy of energy prediction is improved. From theoretical perspective, including occupancy data can improve the prediction accuracy, since the uncertainty of occupancy factor can be controlled and reduced, and less uncertainty can improve the prediction accuracy.

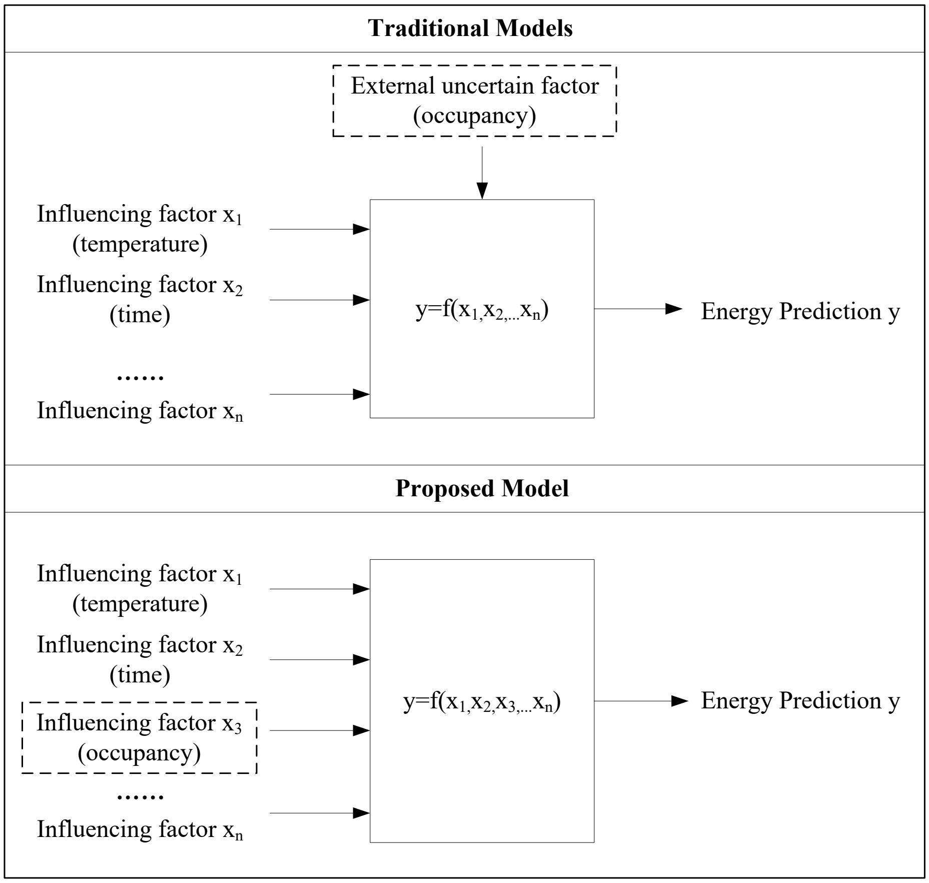
Traditional Models
External uncertain factor (occupancy)
Influencing factor x1 (temperature)
Influencing factor x2 (time)
……
Influencing factor xn
y=f(x1,x2,…xn)
Energy Prediction y
Proposed Model
Influencing factor x1 (temperature)
Influencing factor x2 (time)
Influencing factor x3 (occupancy)
……
Influencing factor xn
y=f(x1,x2,x3,…xn)
Energy Prediction y

Figure 1 Comparison of traditional models and the proposed model

The results of this study reveal that the influence of occupancy on the accuracy of energy prediction. In addition, since the performance of models varies across hours and systems, the proposed method zooms into the hourly performance and different systems (i.e., HVAC, lighting, plug load and total load) of baseline models. Another important feature of this work is it only uses simple algorithm, excluding complex mathematical processing, and the input data is available in most commercial buildings. That means the proposed method is relatively easy to be implemented, and can be well adopted for practical projects. The results of this study can help us understand the quantitative influence of occupancy on energy use and energy baseline models.

2 Methodology

2.1 Framework of evaluating occupancy impact on baseline prediction

The methodology to evaluate occupancy influence on baseline prediction comprises of four steps, illustrated in Figure 2.

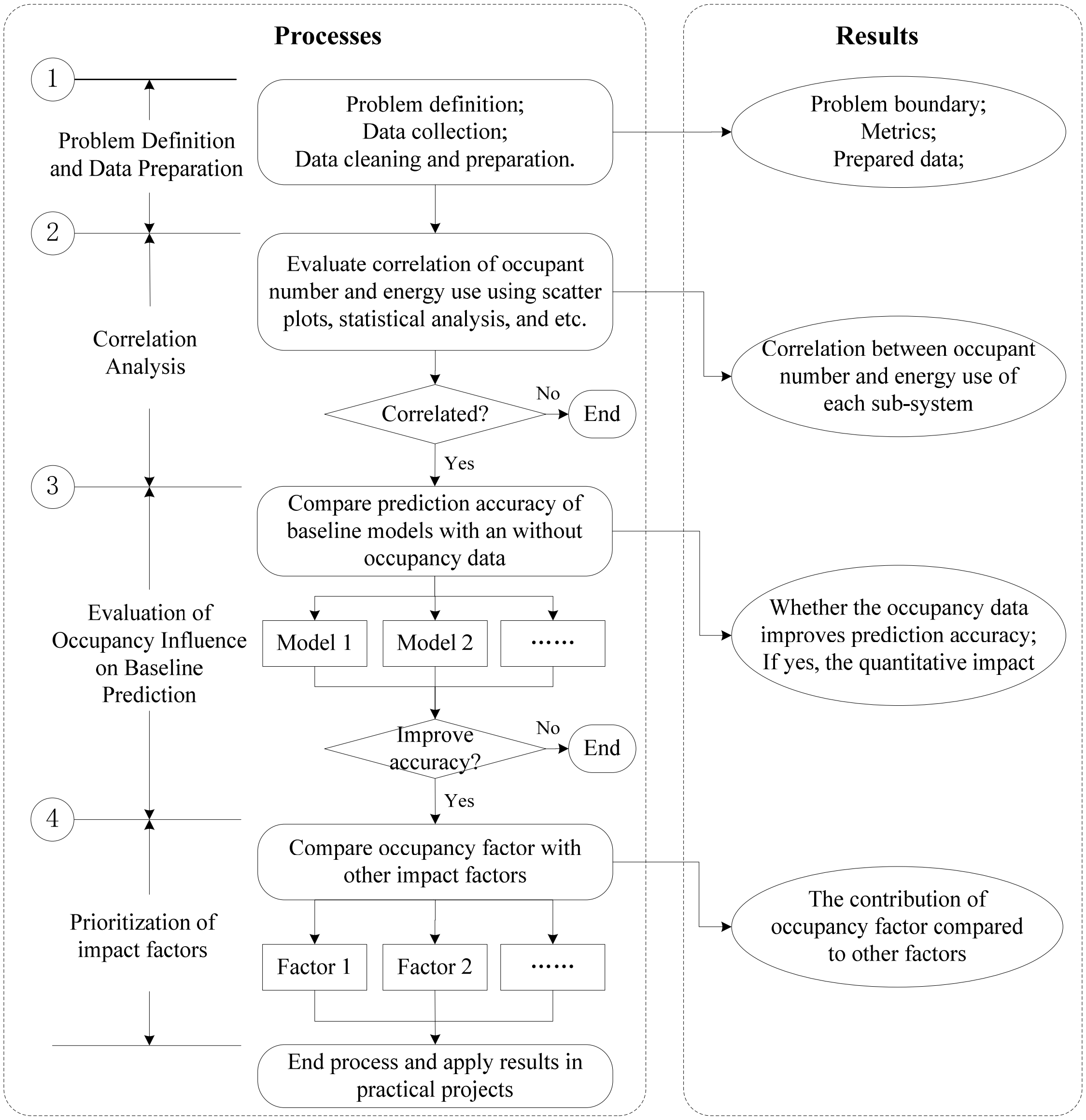
Processes
Results
1
Problem Definition and Data Preparation
Problem definition; Data collection; Data cleaning and preparation.
Problem boundary; Metrics; Prepared data;
2
Correlation Analysis
Evaluate correlation of occupant number and energy use using scatter plots, statistical analysis, and etc.
Correlation between occupant number and energy use of each sub-system
Correlated?
No
End
Yes
3
Evaluation of Occupancy Influence on Baseline Prediction
Compare prediction accuracy of baseline models with an without occupancy data
Whether the occupancy data improves prediction accuracy; If yes, the quantitative impact
Model 1
Model 2
……
Improve accuracy?
No
End
Yes
4
Prioritization of impact factors
Compare occupancy factor with other impact factors
The contribution of occupancy factor compared to other factors
Factor 1
Factor 2
……
End process and apply results in practical projects

Figure 2 Framework of evaluating occupancy impact on baseline prediction

Step 1: Problem Definition and Data Preparation. One aim of this step is to clarify problem definition, boundary, assumption and key metrics of success. The scope of this study focuses on the energy baseline prediction in office buildings. Since there are normally fewer occupants in office buildings on weekends, this study only focuses on the energy use on weekdays. The key metric, which is to assess different models and factors, is the similarity between prediction results and observed data.

The other aim of this step is to prepare data for the analysis in the next steps. It includes acquiring, harmonizing, rescaling, cleaning and formatting data. Due to the failure of sensors, stochastic noise and other interference factors, the raw data set may contain missing data, error data and unstructured data. Before data mining, the raw data should be pre-processed to provide the valid data for further analysis. In this study, three types of data are required (i.e., outdoor air temperature, energy use and occupancy count data). Outdoor air temperature can be obtained from sensors outside buildings or database of weather stations. Energy use data can be obtained from electricity meters in buildings. Occupant number can be obtained from the records of access control system. All these data are recorded with timestamps of short-time intervals, typically at 5 to 15 minutes. Using the short-time "interval data" can significantly reduce the duration of data required in baseline models [8].

Step 2: Correlation Analysis. This step is to verify the correlation between occupancy rate and total energy consumption of buildings. The total energy consumption can be divided into several sub-systems (e.g., HVAC, lighting system and plug load) by using sub-meters. Then, the more occupancy-dependent sub-systems, which have higher correlation with occupancy rate, can be revealed. Scatter plots are applied to visualize correlations qualitatively, while statistical analysis is applied to calculate correlations quantitatively. Correlation coefficients and significance levels are main criteria of correlation test. If occupancy rate and energy use are significantly correlated, the next step will be executed.

Step 3: Evaluation and Comparison of Accuracy of Baseline Models. This step is to quantitatively evaluate the influence of occupancy on the accuracy of baseline models. First, three baseline prediction models are implemented to predict baseline of energy use based on the observed data. Two models, which do not include occupancy factor, are adopted from previous studies. The other one, using occupancy data, is the proposed method in this study. The algorithms of the three models will be illustrated in detail in Section 2.2. Then, the prediction results are compared across the three models. The method and metrics of the evaluation will be introduced in detail in Section 2.3. The results can show whether the occupancy data improves the accuracy of baseline prediction. If the prediction accuracy is significantly improved by occupancy data, the next step will be executed.

Step 4: Prioritization of Impact Factors. Based on horizontal comparison across models in the last step, this step further evaluates influence of occupancy factor by vertical comparison across factors. Besides the number of occupants, there are various uncertain factors which can impact energy consumption of buildings (e.g., outdoor air temperature, facility degradation and climate change). It is important to understand not only the influence of occupancy factor, but also its priority compared to other impact factors. Namely, this step is to identify which factor is more critical to the accuracy of baseline prediction. The results can provide guidance for selecting factors in baseline models. The method and metrics of the factor comparison will be introduced in detail in Section 2.4.

2.2 Baseline prediction models

Three baseline models are implemented to demonstrate the methodology. The first model is a simplistic “native” model, which only depends on one variable, the time of week. It serves as a comparative "floor" of performance [14]. The second model was developed by researchers at LBNL (Lawrence Berkeley National Laboratory), which includes two variables, outdoor air temperature and the time of week [8]. The third one is based on the LBNL model but includes the occupancy variable. The variables included in each model are illustrated in Table 1.

Table 1 The variables included in each baseline model

	Time of week	Outdoor air temperature	Occupancy number
Model 1 (the MW model)	√		
Model 2 (the LBNL model)	√	√	
Model 3 (the new model)	√	√	√

Model 1: the MW (mean-week) model. This model only depends on one variable, the time of week. Consider N observed data points, where data point n is from time t_n including the observed load data L , $n = 1,...,N$. The method is presented by Equation (1), where L_n^p denotes the prediction of baseline at point n, and *time* denotes the time of the week (e.g., 10 am on Monday). For example, the prediction of 10 am on a Monday is the average of historical data for 10 am on all Mondays.

$$L_n^p = L_{n,time}^p = Mean_{time}(L) \tag{1}$$

Model 2: the LBNL model. This is a regression model including the variables of outdoor air temperature and the time of week. The temperature is considered as an important factor of energy use in buildings. The correlation between temperature and energy use is non-linear. In occupied mode, the temperature and energy consumption are normally positively correlated at higher temperature (due to cooling), negatively correlated at lower temperature (due to heating), and relatively un-correlated at moderate temperature (due to no cooling or heating). Therefore, the piecewise-continuous regressions of temperature variable are used in the LBNL model. There are two parts of energy use in this model, one is the time-dependent portion $L_{n,time}^{p}$ and the other one is temperature-dependent portion $L_{n,temp}^{p}$. L_{n}^{p}, the predictive baseline of total energy use at point n, is the sum of these two portions, shown in Equation (2).

$$L_{n}^{p} = L_{n,time}^{p} + L_{n,temp}^{p} \tag{2}$$

The time-dependent portion $L_{n,time}^{p}$ mainly represents the different features of energy consumption among different times. For example, the load is normally lower at night than at working time. $L_{n,time}^{p}$ is modeled by dividing a week into 120 one-hour time slots (24 hours multiply 5 weekdays). An indicator $\tau_{n,i}$ and a coefficient α_i are assigned to each time slot S_i, for $i = 1,...,120$. The whole time-dependent portion $L_{n,time}^{p}$ can be calculated by summing the products of indicators and coefficients of all time slots, shown in Equation (3).

$$L_{n,time}^{p} = \sum_{i=1}^{120} \tau_{n,i}\alpha_i \tag{3}$$

The indicator $\tau_{n,i}$, which is defined in Equation (3), serves to select which coefficient is active. For a given point t_n , only one indicator is one, while other 119 indicators are zero. When $\tau_{n,i} = 0$, the coefficients have no effect.

$$\tau_{n,i} = \begin{cases} 1 & if \;\; t_n \in S_i \\ 0 & if \;\; t_n \notin S_i \end{cases} \tag{4}$$

The temperature-dependent portion $L_{n,temp}^{p}$ mainly represents the different features of energy consumption among different temperatures, which is probably most related to the heating and cooling systems behaviors. As aforementioned, $L_{n,temp}^{p}$ is modeled by a piecewise-linear and continuous function. A number of temperature intervals need to be divided for this piecewise-linear function and a temperature component $\theta_{n,j}$ and a coefficient β_i is assigned to each interval. The temperature T_n is the sum $T_n = \sum_{j=1}^{N_T} \theta_{n,j}$, where N_T is the number of temperature intervals and $\theta_{n,j}$ is the portion of the T_n in interval j. For example, in the case study of [8], four intervals are defined (i.e., 20-40 °F, 40-60 °F, 60-80 °F, 80-100 °F). If the given temperature $T_n = 70$ °F, the values of four components are $\theta_{n,1} = 20$ °F, $\theta_{n,2} = 20$ °F, $\theta_{n,3} = 10$ °F, $\theta_{n,4} = 0$ °F. The whole temperature-dependent portion $L_{n,temp}^{p}$ can be calculated by summing the products of temperature components and coefficients of all intervals, shown in Equation (5).

$$L_{n,temp}^{p} = \sum_{j=1}^{N_T} \beta_j \theta_{n,j} \tag{5}$$

The predictive baseline L_n^p by Equation (2) can be transformed to Equation (6), where the coefficients α_i and β_j can be computed by training with observed data.

$$L_n^p = L_{n,time}^{p} + L_{n,temp}^{p} = \sum_{i=1}^{120} \alpha_i \tau_{n,i} + \sum_{j=1}^{N_T} \beta_j \theta_{n,j} \tag{6}$$

Model 3: the new model. This new model is developed from the LBNL model by including the occupancy variable. Besides the outdoor air temperature and the time variables, the occupancy variable is added in this model. The predictive baseline L_n^p comprises three portions (i.e., the time-dependent portion $L_{n,time}^{p}$, the temperature-dependent portion $L_{n,temp}^{p}$ and the occupancy-dependent portion $L_{n,occ}^{p}$). It is described in Equation (7), where the methods for computing $L_{n,time}^{p}$ and $L_{n,temp}^{p}$ are the same as the LBNL model.

$$L_n^p = L_{n,time}^p + L_{n,temp}^p + L_{n,occ}^p \tag{7}$$

The occupancy-dependent portion $L_{n,occ}^p$ mainly represents the different features of energy consumption among different occupant numbers, which is probably most related to the occupant behaviors (e.g., turning on lights when arriving). Similar to $L_{n,temp}^p$, $L_{n,occ}^p$ can be modeled by a piecewise-linear and continuous function, since the dependence of load on occupant number is not a linear function either. The occupant number and energy consumption are normally positively correlated when buildings are moderate-occupied, but are relatively un-correlated when buildings are heavily-occupied, since no more appliances can be turned on.

A number of occupancy intervals need to be divided for this piecewise-linear function and an occupancy component $\phi_{n,k}$ and a coefficient γ_k is assigned to each interval. The Occupant number O_n is the sum $O_n = \sum_{k=1}^{N_O} \phi_{n,k}$, where N_O is the number of occupancy intervals and $\phi_{n,k}$ is the portion of the O_n in interval k. For example, if occupant intervals are defined (i.e., 0-10, 10-20, 20-50, 50-100, 100-200) and the given number of occupants $O_n = 120$, the values of five components are $\phi_{n,1} = 10$, $\phi_{n,2} = 10$, $\phi_{n,3} = 30$, $\phi_{n,4} = 50$, $\phi_{n,5} = 20$. The whole occupancy-dependent portion $L_{n,occ}^p$ can be calculated by summing the products of occupancy components and coefficients of all intervals, shown in Equation (8).

$$L_{n,occ}^p = \sum_{k=1}^{N_{OI}} \gamma_k \phi_{n,k} \tag{8}$$

The predictive baseline L_n^p by Equation (7) can be transformed to Equation (9), where the coefficients α_i , β_j and γ_k can be computed by regressing with observed data. Then the baseline of energy consumption can be predicted with the obtained coefficients.

$$L_n^p = L_{n,time}^p + L_{n,temp}^p + L_{n,occ}^p = \sum_{i=1}^{120} \alpha_i \tau_{n,i} + \sum_{j=1}^{N_T} \beta_j \theta_{n,j} + \sum_{k=1}^{N_O} \gamma_k \phi_{n,k} \tag{9}$$

In the model of this case study, N_T is set to 2 with the intervals (0-45°F, 45-100°F), and N_O is set to 4 with the intervals (0-10, 10-50, 50-100, 100-200).

2.3 Computing the accuracy of baseline models

The accuracy of baseline models can be quantified by the metric *CVRMSE* (coefficient of variation of the root mean square error) [14]. *CVRMSE* is the root mean square error divided by the mean of the data, which indicates the relative size of error. For example, a value of 0.1 means the difference between prediction and observed data is 10% of observed data. The equation for *CVRMSE* is provided in Equation (10), where L_n^{ob} and L_n^p are the observed data and baseline prediction reprehensively, and N is the size of the data set.

$$CVRMSE = \frac{\sqrt{\frac{\sum_{n=1}^{N} (L_n^{ob} - L_n^p)^2}{N}}}{\frac{\sum_{n=1}^{N} L_n^{ob}}{N}} \tag{10}$$

Cross-validation is applied to facilitate the quantification of the baseline accuracy. The observed data is partitioned into several subsets. Some parts of them are used for model fitting and training, and other parts are used for validating. Then, the process is iterated by changing training set and validation set. In this study, the time interval to partition observed data is one month, since it is normally the utility bill period. First, the model is fitted by the data in one or several intervals (the training length can vary, and the sensitivity analysis of training length will be discussed in Section 3.5), and is validated by the data in the next interval. Then, the training set and validation set are shifted, and the process of training and validating is repeated until the end of data set. The schematic of the cross-validation processes is shown in Figure 3.

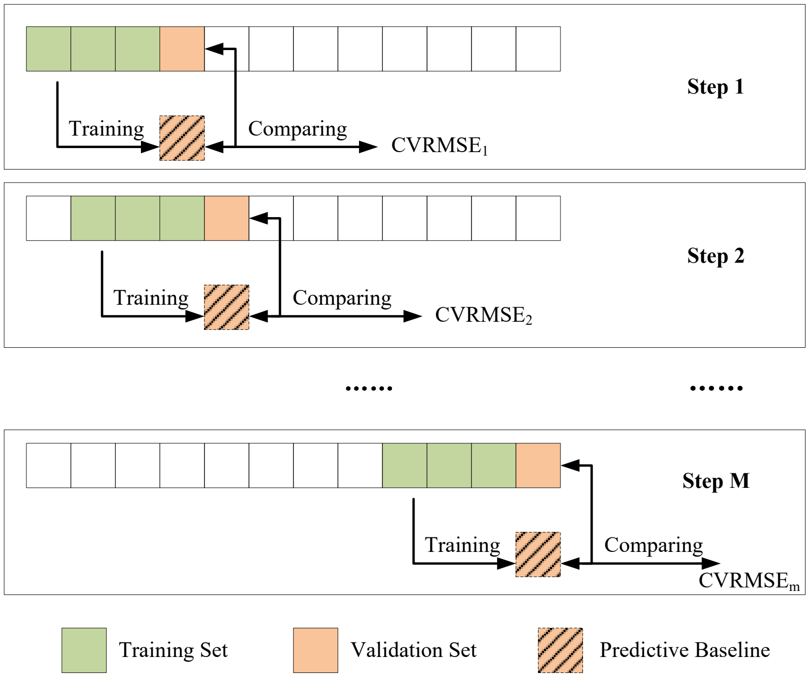

Figure 3 Schematic of the cross-validation processes

In each step, a $CVRMSE$ value can be obtained, and the finial indicator of prediction accuracy is the average of $CVRMSE$, $\overline{CVRMSE}$. The equation for $\overline{CVRMSE}$ is shown in Equation (11), where $CVRMSE_m$ is the $nMAE$ in the *m-th* step, and M is the number of steps.

$$\overline{CVRMSE} = \frac{\sum_{m=1}^{M} CVRMSE_m}{M} \tag{11}$$

2.4 Calculating the influence of variables

Besides comparing the accuracy of models, understanding the impact of each influencing factor is critical in baseline prediction. As shown in Figure 4, Model 1 includes the variable of time while Model 2 includes the variables of time and temperature. If the accuracy of Model 2 is improved, it should be caused by the incremental information of temperature.

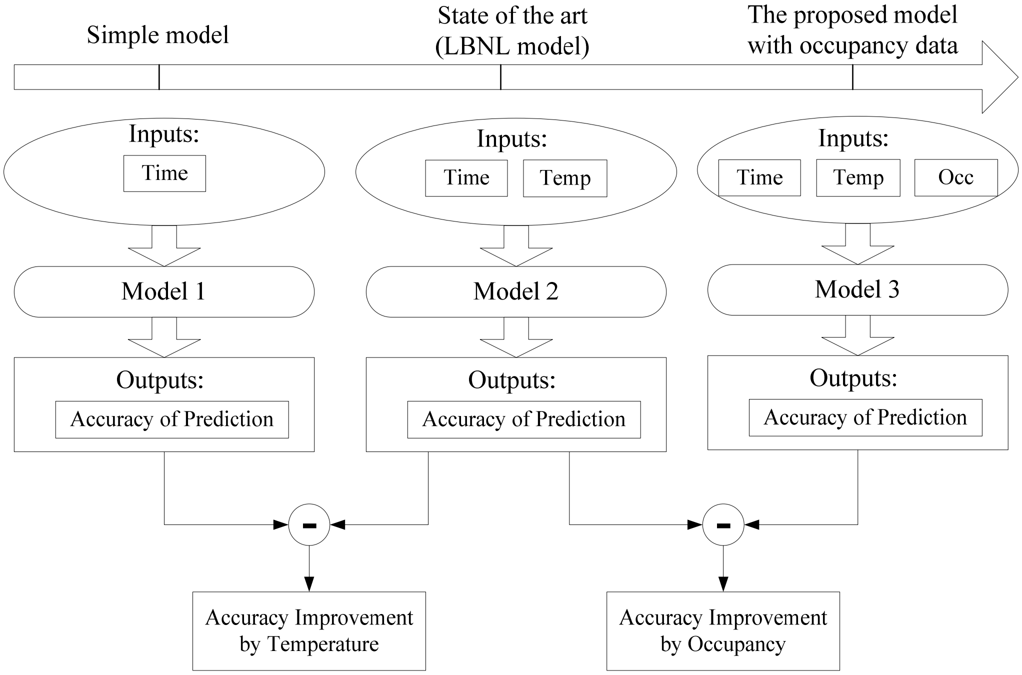

Figure 4 The process of calculating the influence of variables

Since Model 1 and 2 are linear models, the impact of temperature factor $Impact_{temp}$ can be defined as the accuracy improvement of Model 2 compared to Model 1, shown in Equation (12).

$$Impact_{temp}(\%) = \frac{\overline{CVRMSE_{Model2}} - \overline{CVRMSE_{Model1}}}{\overline{CVRMSE_{Model2}}} \times 100\% \qquad (12)$$

Likewise, as shown in Figure 4, the contribution of occupancy factor $Impact_{occ}$ can be defined as the accuracy improvement of Model 3 compared to Model 2, shown in Equation (13), since the only difference between these two models is the variable of occupancy. If there are other impact factors involved in baseline models, the contribution of one factor can be defined with the same method: comparing the accuracy by controlling all other factors and changing the target factor.

$$Impact_{occ}(\%) = \frac{\overline{CVRMSE_{Model3}} - \overline{CVRMSE_{Model2}}}{\overline{CVRMSE_{Model3}}} \times 100\% \qquad (13)$$

3 Results

3.1 Data preparation

A case study was conducted to show how to quantify the availability of occupancy impact on the accuracy of baseline prediction by the proposed method. Building 101 in the Navy Yard, Philadelphia, Pennsylvania U.S. was used in this case study. The building is one of the nation's most highly instrumented office buildings and is the temporary headquarters of the U.S. Department of Energy's Energy Efficient Building Hub (EEB Hub) [35]. Various sensors have been installed by EEB Hub since 2012 to acquire building data of occupants, facilities, energy consumption and environment. The profile of Building 101 is shown in Table 2.

Table 2 The profile of Building 101

Location	Philadelphia, US
Size	6410 m^2
Floor	3 floors
Constructed Year	1911
Building Usage	Office

Four sensors are installed at the gates of the building to record the number of occupants entering and exiting. The sensors are located at the first floor in Building 101, shown in Figure 5. This study uses the data from the year 2014 and the time step is five minutes. The data format of raw sensor records is shown in Table 2. The set ($N_{in1,n}$, $N_{in2,n}$, $N_{in3,n}$, $N_{in4,n}$) denotes the number of entering occupants, while the set ($N_{out1,n}$, $N_{out2,n}$, $N_{out3,n}$, $N_{out4,n}$) denotes the number of exiting occupants at the *n-th* time step. Therefore, the number of total occupants in building can be calculated by Equation (14).

$$N_O = \sum_n (N_{in1,n} - N_{out2,n} + N_{in2,n} - N_{out2,n} + N_{in3,n} - N_{out3,n} + N_{in4,n} - N_{out4,n}) \qquad (14)$$

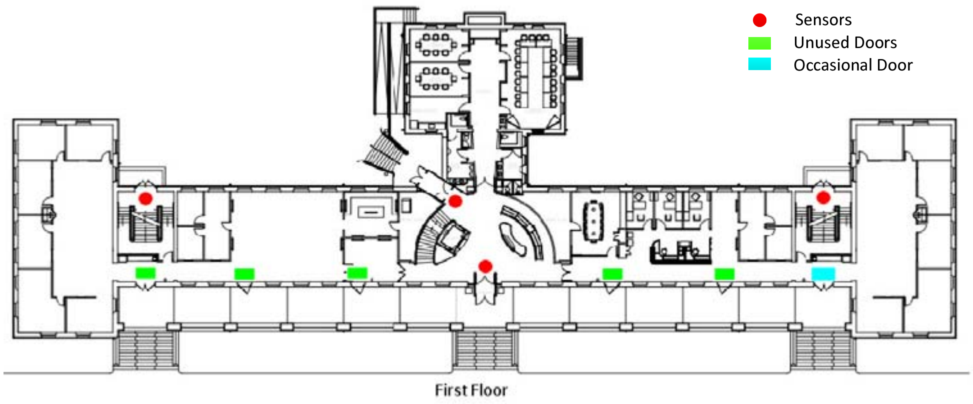

Figure 5 locations of occupancy sensors in Building 101

Table 3 The data format of sensor records

Time step	Sensor1		Sensor2		Sensor3		Sensor4	
	In	Out	In	Out	In	Out	In	Out
1/1/2014 0:00	$N_{in1,1}$	$N_{out1,1}$	$N_{in2,1}$	$N_{out2,1}$	$N_{in3,1}$	$N_{out3,1}$	$N_{in4,1}$	$N_{out4,1}$
1/1/2014 0:05	…..	…..	…..	…..	…..	…..	…..	…..
1/1/2014 0:10	…..	…..	…..	…..	…..	…..	…..	…..
…..	…..	…..	…..	…..	…..	…..	…..	…..
12/31/2014 23:50	…..	…..	…..	…..	…..	…..	…..	…..
12/31/2014 23:55	$N_{in1,n}$	$N_{out1,n}$	$N_{in2,n}$	$N_{out2,n}$	$N_{in3,n}$	$N_{out3,n}$	$N_{in4,n}$	$N_{out4,n}$

The electricity consumption data of the whole building and sub-systems (i.e., lighting, HVAC and plug load) was recorded by sub-meters in 15-minute intervals. Based on this data set, the hourly, daily and monthly energy use of each system can be calculated. The outdoor air temperature was recorded every 15 minutes. Therefore, all the three categories of data (occupancy, temperature and energy use) can be obtained by sensors and meters in Building

101. After harmonizing, rescaling, cleaning and formatting the raw data, it is ready for the further analysis.

3.2 Correlation between occupancy and energy consumption

The correlation between occupancy and energy consumption was investigated with three methods: time series, scatter plots and correlation coefficient tests. Figure 6 illustrates the comparison of the hourly energy consumption and the occupant number. Similar to the ASHRAE 90.1 standard, the occupancy curve during 24 hours represents the dual-peak feature, but the noon-drop is not as deep as that in the ASHRAE 90.1 standard. According to the feature, the occupancy curve can be divided into six periods [36]: (1) the night period (7 pm to 6 am); (2) the going-to-work period (7 am to 9 am); (3) the morning period (10 am to 12 pm); (4) the noon-break period (12 pm to 1 pm); (5) the afternoon period (2 pm to 3 pm); and (6) the going-home period (4 pm to 6 pm). According to the distribution of the boxplot, the higher uncertainties of the occupant number occurred during going-to-work and going-home periods.

The main feature of energy consumption is similar to that of the number of occupants (lowest at night, increasing in the morning and decreasing in the afternoon), but is not quite synchronized. The energy consumption curve can be divided into four periods: (1) the valley period (10 pm to 3 am); (2) the increasing period (4 am to 9 am); (3) the peak period (10 am to 5 pm); (4) the decreasing period (6 pm to 9 pm). The energy consumption rises about three hours earlier than occupants arriving, and falls around two hours later than occupants leaving. It indicates that the operation schedule of building energy systems is around five hours longer than occupied time in this building. In addition, it needs to be noted that the energy consumption does not have dual-peak feature. Namely, the energy consumption keeps the peak value during noon-break, which indicates the lights, HVAC and other plug load equipment are not turned off when occupants leave for lunch.

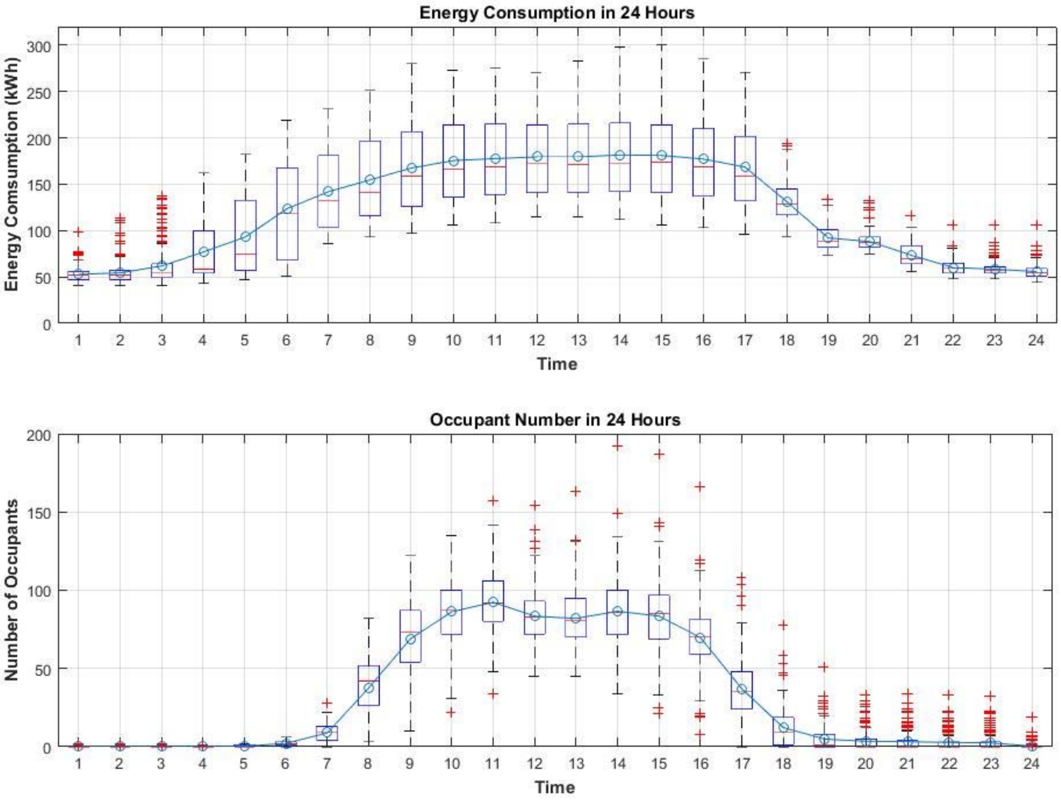

Figure 6 Hourly Energy consumption and occupant number in weekdays. Boxplots show median, quartiles, extreme values, means (blue circles) and outliers (+) of the data set.

Figure 7 shows the correlation between the number of occupants and energy consumption by scatter plots. The color bar indicates the time of the day. The color is closer to red when time is closer to noon, while the color is closer to blue when time is closer to midnight. The total load and occupant number present positive correlation. Although not very significant, the trend can be discovered: the more occupants there are, the higher the total load is. The lighting and plug load systems show more significant positive correlation between energy use and occupant number. Especially in the plug load system, the slope is high, which means a given change of occupant number will cause relatively large change of energy use.

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

Nevertheless, the occupant number does not show significant correlation with energy use in HVAC systems.

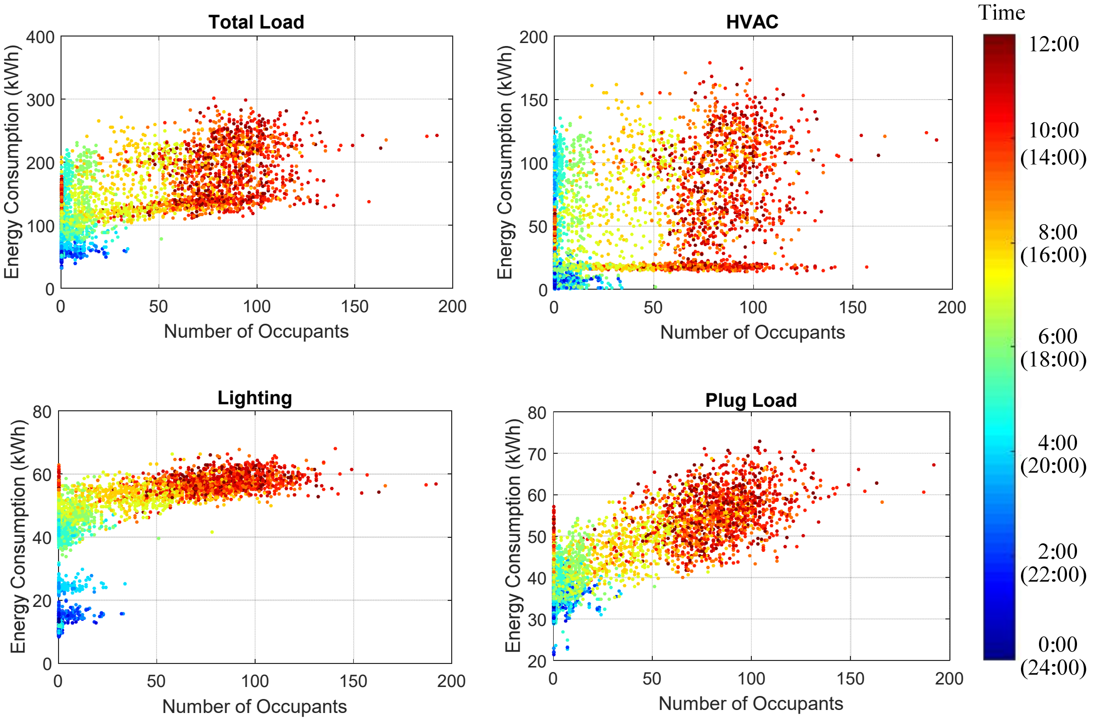

Figure 7 The correlation between the number of occupants and energy consumption

To compare with occupancy, the temperature was likewise analyzed to show the correlation with energy use. As shown in Figure 8, during night (blue dots), the total load is not related to temperature. During daytime (yellow and red dots), there is significantly positive correlation when temperature is higher than 40 °F, otherwise, there is no significant correlation between them. The HVAC system is similar to the total load, but the correlation is more significant. There is no significant correlation in lighting and plug load systems. Since Building 101 uses gas for heating rather than electricity, the total energy use does not rise in lower temperature. However, the energy use in plug load system rises slightly in lower temperature, probably due to personal electric heaters.

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

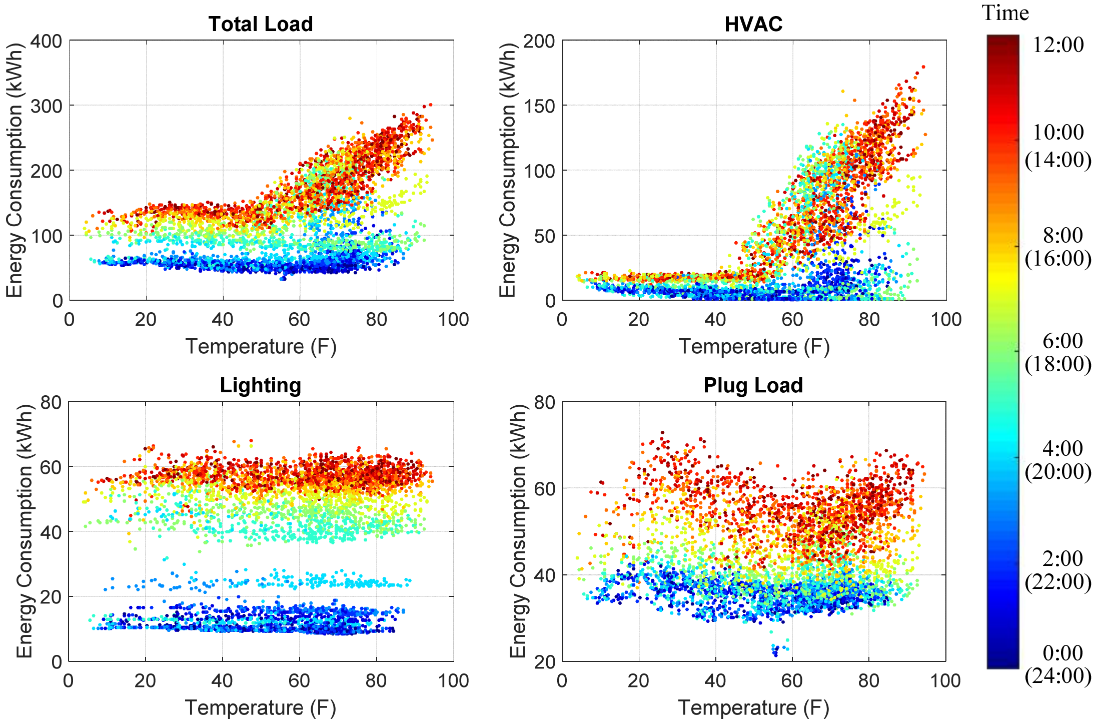

Figure 8 The correlation between the temperature and energy consumption

Besides the visualization of correlation by scatter plots, correlation analysis is adopted to calculate the correlation coefficients, shown in Table 4. In vertical comparison, the coefficient of occupant number (0.74) is 30% higher than that of temperature (0.44) in total energy use. This premium becomes greater in lighting and plug load systems, which are 60% and 81% respectively. The coefficients in HVAC system are approximately equal. Therefore, overall, the occupant number has much higher correlation with energy use than outdoor air temperature.

In horizontal comparison, the occupancy is more correlated to lighting and plug load systems, while the outdoor temperature is more correlated to the HVAC system. These results are consistent with common sense and previous studies [36, 37], because the lighting and plug load are controlled by occupants, but the HVAC system mainly depends on the outdoor air temperature.

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

Table 4 The correlation coefficients between occupancy/temperature and energy consumption

	Total Electric Load	HVAC	Lighting	Plug Load
Number of Occupants	0.74*	0.54*	0.73*	0.86*
Outdoor Air Temperature	0.44*	0.58*	0.13*	0.05*

* p-value <0.001

3.3 Accuracy of baseline models

The results in Section 3.2 have proved that the occupant number is highly correlated to energy consumption. The further question is whether the accuracy of baseline models can be improved by including the occupancy variable. To answer this question, Model 3, which uses the time, outdoor air temperature and occupancy variables, is implemented to compare with the previous methods. Since the volume of results is huge (a whole year in 1-hour intervals), it is difficult to show all the results. Therefore, one week of results is shown in Figure 9, with comparison of observed data and results of three models.

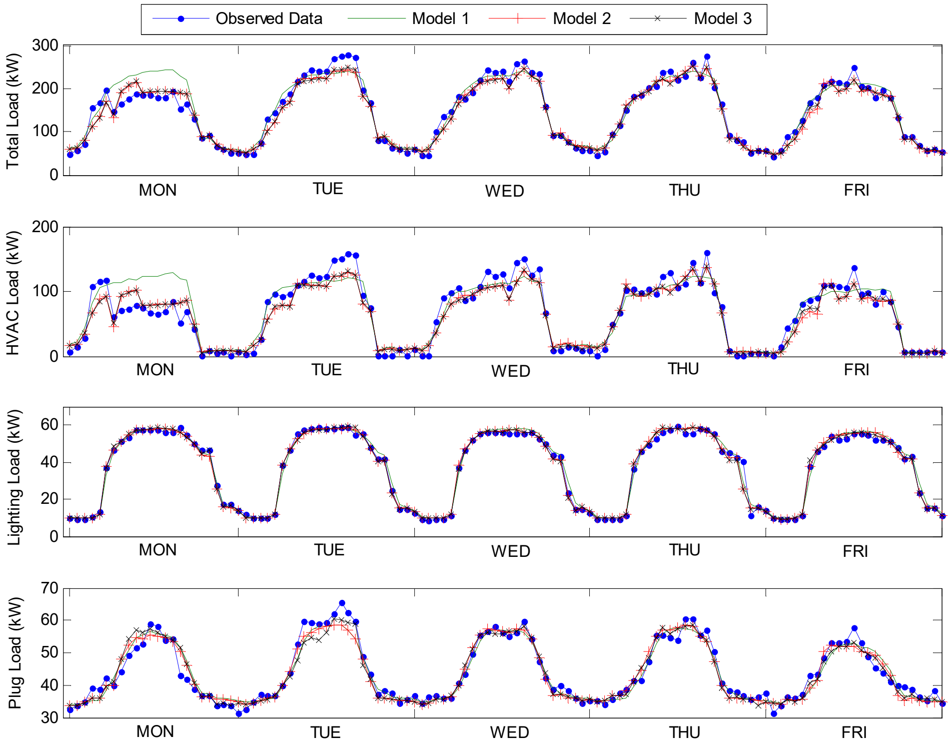

Figure 9 The observed load and predicted load by three models (from 11th-15th August 2014)

The accuracy of each model, measured by $\overline{CVRMSE}$, is shown in Figures 10-13. The lower value of $\overline{CVRMSE}$ indicates the higher accuracy.

- Figure 10 illustrates the accuracy of baseline models in total load prediction. The values of $\overline{CVRMSE}$ in Model 1 are around 0.25 during working time. The peak values are around 0.45, which are from 4 am to 6 am, and the valley values are around 0.1 which are at 8 pm. After including the outdoor temperature variable, the accuracy of Model 2 is improved significantly. The values of $\overline{CVRMSE}$ are mostly below 0.15, and higher $\overline{CVRMSE}$ values beyond 0.15 occur from 2 am to 6 am. The accuracy of

Model 3 is slightly improved from Model 2, and the shape of $\overline{CVRMSE}$ curve is very similar.

- Figure 11 illustrates the accuracy of baseline models in HVAC load prediction. It indicates that Model 1 is poor at HVAC load prediction. The $\overline{CVRMSE}$ values in Model 1 are mostly beyond 0.5, which means most prediction values deviate from observed value by more than 50%. The peak value is around 0.9 at 4 am. After including the temperature variable, the accuracy of Model 2 is improved significantly. The values of $\overline{CVRMSE}$ drop to below 0.2 during daytime (6 am to 6 pm), but the values of $\overline{CVRMSE}$ are still higher than 0.5 at night (from 7 pm to 3 am). By including the occupancy variable, the accuracy of Model 3 is not significantly improved in most time, except 7 pm to 12 am. The shape of $\overline{CVRMSE}$ curve is very similar. The big differences of accuracy in HVAC load prediction are probably caused by the operation schedule, which is related to neither occupancy nor outdoor temperature in this building. It will be discussed in detail in Section 4.
- Figure 12 illustrates the accuracy of baseline models in lighting load prediction. Model 1 performs well at lighting load prediction. The $\overline{CVRMSE}$ values in Model 1 are mostly below 0.1. But the $\overline{CVRMSE}$ values rise sharply at 6 am and 9 to 10 pm. After including the outdoor temperature variable, the accuracy of Model 2 is improved significantly at 6 am and 9 to 10 pm, which the $\overline{CVRMSE}$ values drop to around 0.15. By involving the occupancy variable, the accuracy of Model 3 is slightly improved in daytime (from 8 am to 6 pm), but not improved in other hours. The shape of $\overline{CVRMSE}$ curve is very similar.

- Figure 13 illustrates the accuracy of baseline models in plug load prediction. Model 1 performs well at lighting load prediction. The $\overline{CVRMSE}$ values in Model 1 are mostly below 0.1. The two peak values of $\overline{CVRMSE}$ occur at 8 am and 6 pm. After including the outdoor temperature variable, the accuracy of Model 2 is improved, which the $\overline{CVRMSE}$ values drop to below 0.09. By involving the occupancy variable, the accuracy of Model 3 is significantly improved, especially during working time (6 am to 7 pm). All the $\overline{CVRMSE}$ values of Model 3 drop to below 0.08. Different from the other two systems, the shape of $\overline{CVRMSE}$ curve of Model 3 for plug load is not similar to that of Model 2.

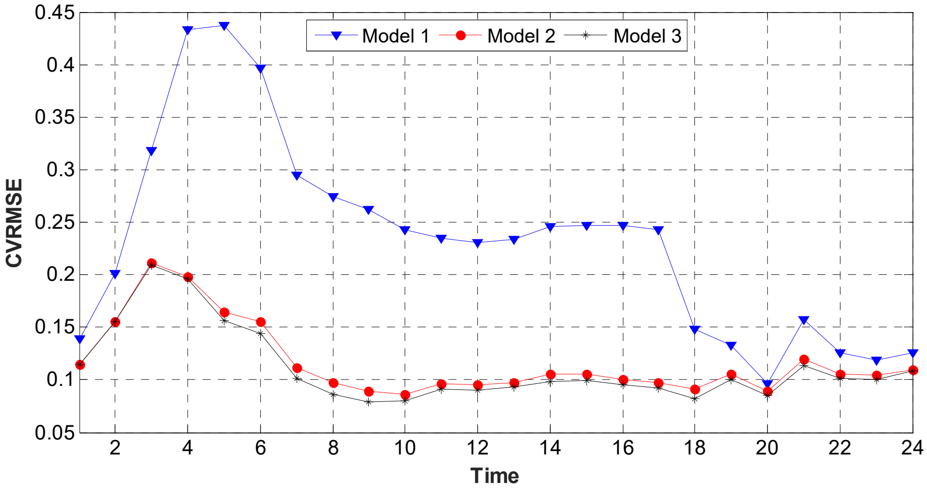

Figure 10 The accuracy of baseline models for total electric load prediction

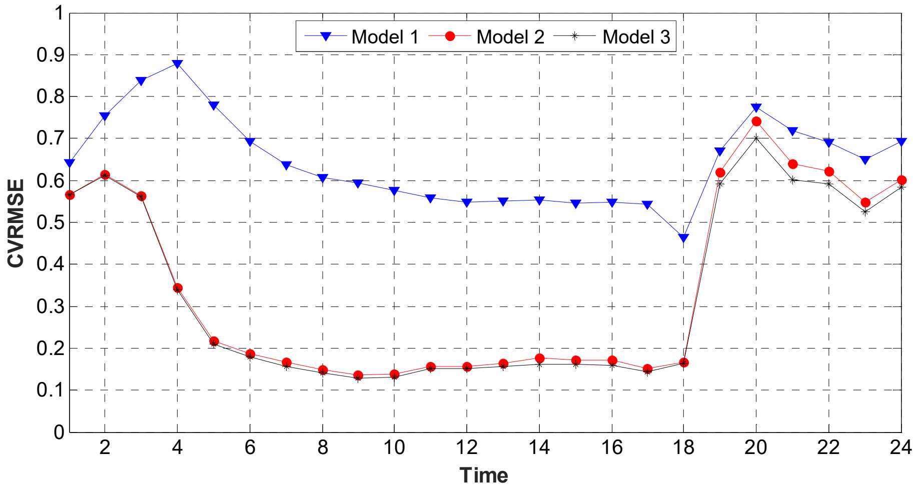

Figure 11 The accuracy of baseline models for HVAC load prediction

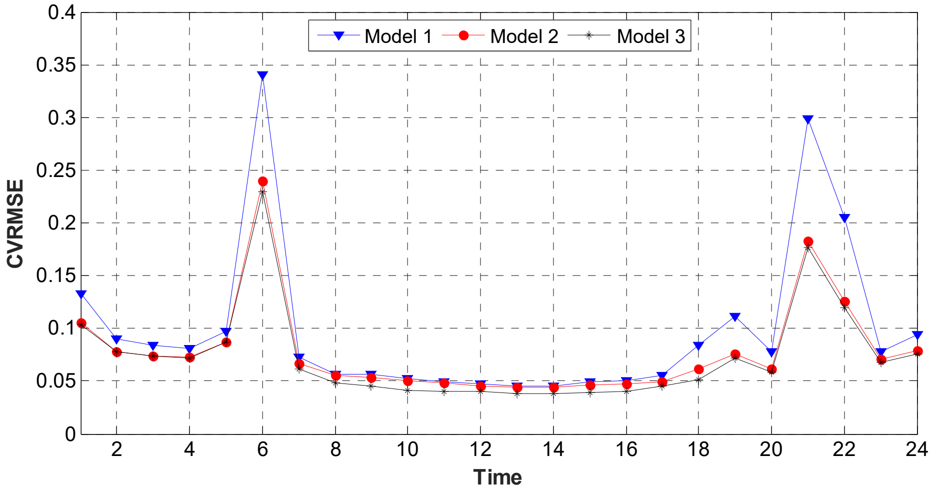

Figure 12 The accuracy of baseline models for lighting load prediction

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

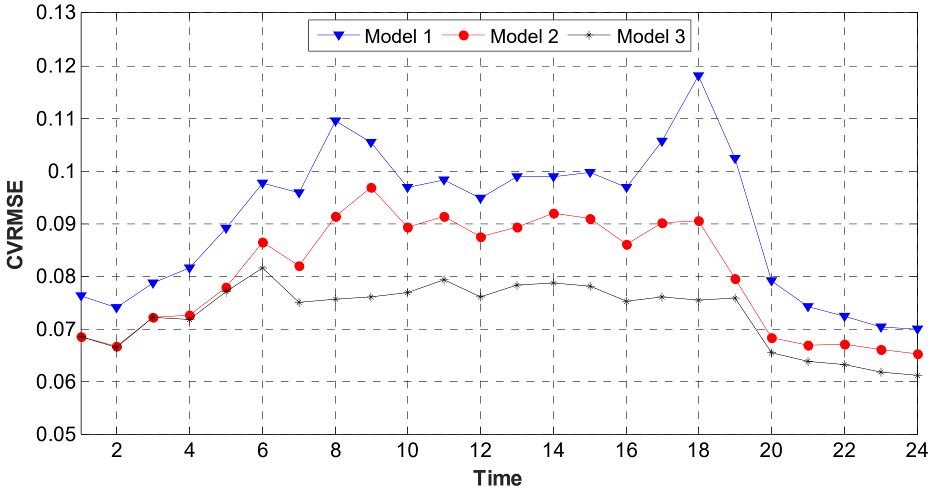

Figure 13 The accuracy of baseline models for plug load prediction

3.4 Contribution of the occupancy factor

The results in Section 3.3 confirmed the hypothesis that the occupancy data can improve baseline prediction. The next step is to quantify the contribution of the occupancy variable, and clarify whether this contribution is higher or lower than that of other factors. The result can help determine the dominant factors in baseline models. In this step, the contributions of occupancy and temperature are calculated and compared using the method introduced in Section 2.4.

Figure 14 illustrates the contribution of occupancy data on the accuracy of baseline prediction. The results show that occupancy data improves lighting and plug load prediction most significantly, especially during working time (8 am to 6 pm). But the improvement is not significant in HVAC load prediction, lower than 10%. Overall, the occupancy data improves the total energy prediction by around 10% during daytime (6 am to 6 pm), but less improvement at other times.

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

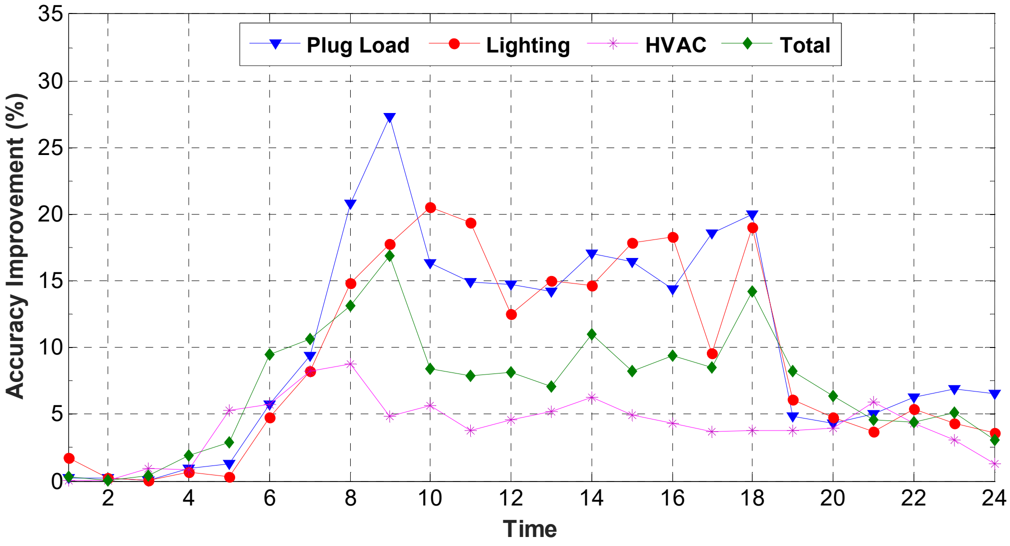

Figure 14 The contribution of occupancy data on the accuracy of baseline prediction

The statistical results of contributions of occupancy $Impact_{occ}$ and outdoor temperature $Impact_{temp}$ are shown in Table 5. According to the results, the outdoor temperature variable mainly contributes to HVAC and lighting load prediction, while the occupancy variable mainly contributes to lighting and plug load prediction. The mean contribution of occupancy variable on total energy prediction is 10%, which is much lower than the mean contribution of the outdoor temperature variable (63%). Occupancy has higher correlation with energy use but lower contribution on energy prediction, which seems inconsistent. The reasons of this problem will be discussed in Section 4.

Table 5 The statistical profile of the contributions by occupancy and temperature factors

	$Impact_{occ}$			$Impact_{temp}$		
	Max	Mean	Median	Max	Mean	Median
HVAC	9%	5%	5%	72%	46%	64%
Lighting	21%	12%	7%	38%	13%	10%
Plug Load	27%	15%	9%	20%	10%	9%
Total	18%	10%	8%	66%	44%	57%

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

3.5 Sensitivity analysis

There are three critical parameters influencing the prediction results in baseline models. First is the length of the training data period. The baseline models use previously observed data to train and fit models. The length of the training period will impact the training effect and further impact the accuracy of baseline prediction. The length should be neither too short nor too long [8]. If the training is too short, it cannot provide enough information to fit the model. If the training is too long, it may include useless or harmful information to the model. Since the building performance and occupant activities change over time, the data of the building in the distant past does not help predict the building performance in the future. For example, over a period of years, the base load of the building is likely changed. There will be a considerable bias if using data from years ago. The number of occupants and their energy use behaviors can be likewise changed during a long time, so the historical data can no longer reflect the current building performance.

The other two critical parameters are the piecewise number of occupancy and the outdoor temperature in regressions. As mentioned in Section 2.2, Model 3 is piecewise-continuous regressions of the occupancy and temperature variables. The piecewise number will impact model fitting. Fewer segments will sacrifice accuracy of the model, while too many segments can cause over-fitting and high computing cost. Therefore, how to define these segments appropriately is an important issue in the baseline model.

Sensitivity analysis is applied to evaluate the influence of these parameters on baseline models. Figure 15 shows the accuracy of baseline models during different training periods. The *CVRMSE* of Model 1 first increases with the training period and reaches the peak value at five months, then decreases. It can be explained that when the training period is five months, it uses training data from winter to predict the building performance in summer. As

Model 1 does not include the outdoor temperature variable, the prediction should be at lower accuracy. It verifies the aforementioned hypothesis, that longer training period may be harmful to accuracy. The *CVRMSE* of Models 2 and 3 fluctuate in short training periods, and reach convergence after three months. Their curves are almost coincident after three months, and Model 3 is slightly below Model 2. In short training periods (one to three months), Model 3 shows faster convergence and narrower range of fluctuation. It can bring not only technical but also economic benefits, since the time for data gathering can significantly impact the costs, investment return and payback period [8].

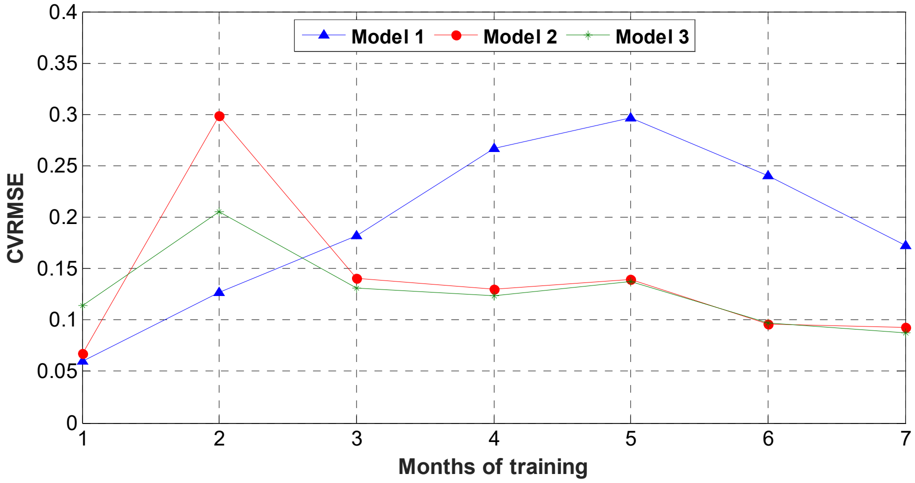

Figure 15 Accuracy of baseline prediction under different training periods

Figure 16 shows the accuracy of Model 3 under different piecewise number of occupancy and temperature data. For the temperature curve, The *CVRMSE* of Model 3 drops sharply from one segment to two segments, then decreases slowly with more than two segments, where the changes are lower than 2%. It means the piecewise number of temperature should be more than 2. For the occupancy curve, the *CVRMSE* of Model 3 stays stable over different piecewise numbers. Therefore, different segment definitions of the occupancy variable will not impact the accuracy of model significantly.

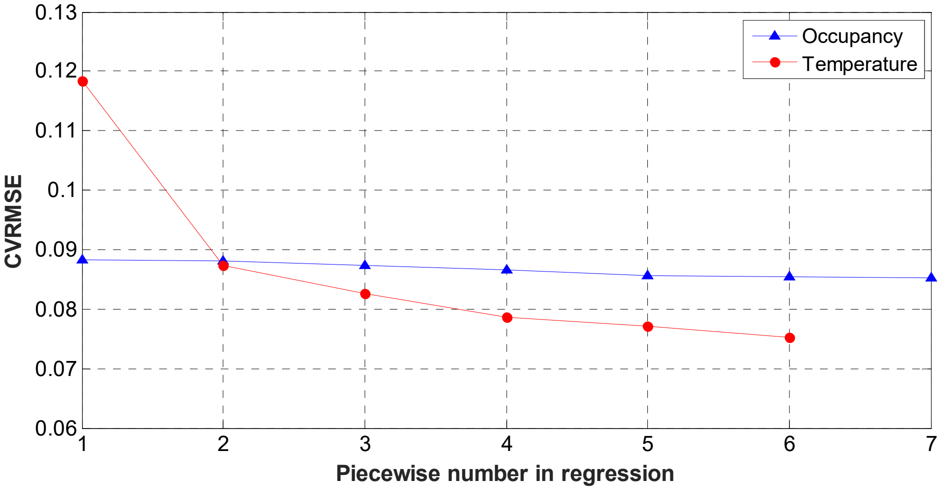

Figure 16 Accuracy of baseline prediction in Model 3 under different piecewise number of occupancy and temperature data

4 Discussion

According to the results in Section 3, a critical question needs to be answered: why occupancy is highly correlated with energy consumption but contributes less in baseline prediction (coefficient is 0.74 but contribution is 10%). It seems inconsistent and nonsensical, especially compared with the temperature variable (coefficient is 0.44 but contribution is 63%). There are two main reasons behind it. One is that the occupant number is highly correlated with time, so the time variable has provided most information of the occupancy. In the three baseline models, time of week is an important variable, which includes 120 one-hour time slots (24 hours multiply five weekdays). Although occupant number changes over time stochastically and is in high uncertainty, the occupant number in each time slot is in relatively low uncertainty. As shown in Figure 6, during 12 hours (7 pm to 7 am) of a day, the uncertainty of occupancy is close to zero. During the other 12 hours, the uncertainty of most occupancy data is under 20%. It means the time variable is highly correlated with occupancy and can provide most occupancy information. Therefore, the occupancy variable cannot provide much incremental information to the model.

The other reason is that the operation schedule mainly depends on time rather than occupancy. As shown in Figure 6, the operation time is much longer than occupied period. For example, the energy consumption significantly rises from 4 am, and the load has reached nearly 80% of mean peak load at 6 am, while there are fewer occupants in the building. Figure 6 can likewise verify this issue. Except plug load, the energy consumption can nearly reach peak value in the early morning and late afternoon. It means building systems are controlled by operation schedule rather than occupants. Therefore, the time variable is better to reflect energy consumption than the occupancy variable.

After clarifying the reasons of the last question, there is a further question: whether the occupancy variable can be removed from baseline model due to its less contribution. On the contrary, although the contribution of occupancy variable to accuracy is not as significant as temperature in this case, it can be an important indicator in M&V for energy efficiency retrofit. First, it can indicate the occupancy-related risk. According to the aforementioned first reason, the contribution of occupancy variable is lower when occupancy is more correlated to the time. In this case study, the low contribution of occupancy indicates the occupancy-related risk is low in this building, mainly because it is an office building and the occupancy is regular during one year. Conversely, the high contribution means the occupancy is highly uncertain and stochastic. If the occupancy-related uncertainty is very high (e.g., hotels), it needs to carefully consider the occupancy-related risk in retrofit decision making. Furthermore, it can clarify whether the changes of energy use are from retrofit or operation. Some buildings cannot achieve the energy saving target after retrofit, and common disputes are focused on whether it is caused by ineffective retrofit or inappropriate operation. If the occupancy does not significantly change but the contribution of occupancy is abnormally low, it indicates the operation schedule is inconsistent with the occupancy schedule and is more responsible for the excessive consumption.

There are three advantages of the proposed baseline model. First, this study defines the metrics to quantify the influence of occupancy. Numerous previous studies emphasized the impact of occupancy on M&V, however the quantified influence of occupancy is under-developed. Without this, it is difficult to improve baseline models as well as facilitate real projects of energy efficiency retrofit. Based on the proposed metrics, the contribution of different variables in the baseline models can be analyzed and compared. The proposed metrics can then be used to evaluate other factors in the baseline models.

Second, the proposed method zooms into the hourly performance and different systems of baseline models. Previous studies only provided the overall whole building results of baseline prediction, but the performance of model varies across hours and systems. For example, the load prediction for HVAC system is very accurate at daytime ($\overline{CVRMSE}$ is less than 0.2), but rises dramatically at night ($\overline{CVRMSE}$ is more than 0.6), shown in Figure 11. To improve baseline models, future research can pay more attention to these issues. Therefore, this method provides a “magnifying lens”, which can help diagnosis and trouble shooting.

Third, the proposed method requires simple input data and algorithm. Three types of data are needed in the model, namely the occupancy data (available in most commercial buildings for security reasons), energy consumption (most commercial buildings have electricity meters capable of providing short-interval data [8]) and the outdoor air temperature (available from local temperature sensor or weather stations). Data limitation is a main barrier in data mining, so the simple data requirement is a considerable benefit for modeling. In addition, this method only uses simple regression algorithm, which is easy to implement and fast in data processing.

The results of this study can be applied in energy efficiency retrofit projects. Before retrofit, it can offer suggestions of data collection, decision making and risk assessment. For example, if the projects are mainly for HVAC, occupancy factor can be ignored. However if the projects are mainly for plug load, it is necessary to collect occupancy data before retrofit since it influences the energy baseline significantly. For the risk assessment, the results of this study can also indicate the uncertainty of energy baseline model impacted by occupancy. If the uncertainty is relatively high, the investment strategy may be changed. After retrofit, the results of this study can improve the energy saving assessment by including the occupancy factor. It is critical for ESCOs, since the profits of ESCOs mainly depend on the calculated energy savings.

5 Conclusions

Baseline prediction is a key issue in M&V and energy efficiency retrofit of buildings. Occupancy, as a critical impact factor of energy consumption, has been emphasized in previous studies. However, few previous studies used the occupancy variable in baseline models or quantified the influence of occupancy variable on baseline prediction.

This study develops a new baseline model by including the occupancy data into the existing LBNL baseline model, and proposes metrics to quantify the accuracy of prediction and the impacts of variables. First, correlation between occupancy and energy consumption is visualized and analyzed by time series plot, scatter plot and statistical method. Then, the accuracies of the three baseline models are compared with the *CVRMSE* metric. Thirdly, based on the accuracy of models, the contributions of variables are quantified and compared. Finally, the sensitivity analysis is conducted to evaluate the influence of parameters in models.

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

The main findings are highlighted as follows:

1) *The correlation between occupancy and total building energy consumption is very high.* Occupancy is most correlated to plug load and lighting, with the correlation coefficients of 0.86 and 0.73 respectively. Outdoor air temperature has much lower correlation with energy consumption than the occupancy.

2) *The contribution of the occupancy variable is relatively low (lower than contribution of temperature).* It is mainly because the time variable can provide most information of occupancy and the operation schedule is inconsistent with the occupied time.

3) *The model including the occupancy variable shows faster convergence and narrower range of fluctuation in short training periods.* When training periods are getting longer, the results of the models with and without the occupancy variable are getting closer.

4) *The piecewise number of occupants in regression does not impact results significantly. But the piecewise number of the outdoor air temperature should be more than 2.*

There are several limitations of this study. First is the reliability of the source data. Due to the sensor failure and other reasons, there is some missing data. And there is a small door used occasionally, shown in Figure 5, which causes the entering number and exiting number to sometimes not be equal. Although the deviation is lower than 5%, it still impacts the accuracy of results. In addition, due to data availability, the case study only uses data from a single building and the time span is one year, so the results should be used with caution. Building 101 is a typical office building, the occupancy is regular over time. It cannot represent other building types with highly random occupancy (e.g., hotels). Third, there are various methods for energy prediction (e.g., change-point regression, ANN, SVR, etc.). The LBNL model is only an example method as function of energy prediction in this study to calculate the occupancy influence on energy prediction quantitatively. But based on the results of this study, occupancy data can be included in more methods to investigate the occupancy influence in further study.

Further research of occupancy in baseline prediction can focus on: (1) using larger data sets for potentially better results; (2) applying more methods and improving algorithm of the baseline model. It needs to consider the tradeoffs among result accuracy, algorithm complexity and length of training period; (3) comparing occupancy influences among different types of buildings and developing benchmarks of M&V for energy efficiency retrofit.

Acknowledgements

This research is funded by the National Natural Science Foundation of China (No. 71271184) and the Hong Kong Polytechnic University. It is also supported by the Assistant Secretary for Energy Efficiency and Renewable Energy of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 through the U.S.-China joint program of Clean Energy Research Center on Building Energy Efficiency. Authors appreciated Clinton Andrews of Rutgers University for providing the occupancy data of Building 101. This work is also part of the research activities of IEA EBC Annex 66, definition and simulation of occupant behavior in buildings.

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

References

[1] EIA. (2010, 09.03). *Annual Energy Review, DOE/EIA – 0384, 2010, Retrieved on 09.03.10 from. http://www.eia.doe.gov/aer/pdf/aer.pdf* Available: http://www.eia.doe.gov/aer/pdf/aer.pdf

[2] P. P. Xu, E. H. W. Chan, and Q. K. Qian, "Success factors of energy performance contracting (EPC) for sustainable building energy efficiency retrofit (BEER) of hotel buildings in China," *Energy Policy,* vol. 39, pp. 7389-7398, Nov 2011.

[3] J. Hong, G. Q. Shen, Y. Feng, W. S.-t. Lau, and C. Mao, "Greenhouse gas emissions during the construction phase of a building: a case study in China," *Journal of Cleaner Production,* vol. 103, pp. 249-259, 9/15/ 2015.

[4] C. C. Menassa and B. Baer, "A framework to assess the role of stakeholders in sustainable building retrofit decisions," *Sustainable Cities and Society,* vol. 10, pp. 207-221, 2014.

[5] UNEP. (2007). *Buildings Can Play Key Role in Combating Climate Change, SBCI-Sustainable Construction and Building Initiative, Oslo, 2007, Retrieved on 09.15.09 from. http://www.unep.org/Documents.Multilingual/Default.Print.asp*

[6] T. Hong, M. A. Piette, Y. Chen, S. H. Lee, S. C. Taylor-Lange, R. Zhang*, et al.*, "Commercial Building Energy Saver: An energy retrofit analysis toolkit," *Applied Energy,* vol. 159, pp. 298-309, Dec 1 2015.

[7] EPA, "Public law 109-58: 109th Congress: An act to ensure jobs forour future with secure, affordable, and reliable energy," 2005.

[8] T. Walter, P. N. Price, and M. D. Sohn, "Uncertainty estimation improves energy measurement and verification procedures," *Applied Energy,* vol. 130, pp. 230-236, 2014.

[9] P. Xu and E. H. W. Chan, "ANP model for sustainable Building Energy Efficiency Retrofit (BEER) using Energy Performance Contracting (EPC) for hotel buildings in China," *Habitat International,* vol. 37, pp. 104-112, 1// 2013.

[10] X. Xia and J. Zhang, "Mathematical description for the measurement and verification of energy efficiency improvement," *Applied Energy,* vol. 111, pp. 247-256, 11// 2013.

[11] J. L. Mathieu, D. S. Callaway, and S. Kiliccote, "Variability in automated responses of commercial buildings and industrial facilities to dynamic electricity prices," *Energy and Buildings,* vol. 43, pp. 3322-3330, 12// 2011.

[12] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, "Statistical analysis of baseline load models for non-residential buildings," *Energy and Buildings,* vol. 41, pp. 374-381, 2009.

[13] J. Granderson, P. N. Price, D. Jump, N. Addy, and M. D. Sohn, "Automated measurement and verification: Performance of public domain whole-building electric baseline models," *Applied Energy,* vol. 144, pp. 106-113, 4/15/ 2015.

[14] J. Granderson and P. N. Price, "Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models," *Energy,* vol. 66, pp. 981-990, 2014.

[15] D. E. Claridge, "A perspective on methods for analysis of measured energy data from commercial buildings," *Journal of solar energy engineering,* vol. 120, pp. 150-155, 1998.

[16] J. W. Taylor, L. M. De Menezes, and P. E. McSharry, "A comparison of univariate methods for forecasting electricity demand up to a day ahead," *International Journal of Forecasting,* vol. 22, pp. 1-16, 2006.

[17] S. Katipamula, T. A. Reddy, and D. E. Claridge, "Multivariate regression modeling," *Journal of Solar Energy Engineering,* vol. 120, pp. 177-184, 1998.

[18] J. K. Kissock, T. A. Reddy, and D. E. Claridge, "Ambient-temperature regression analysis for estimating retrofit savings in commercial buildings," *Journal of Solar Energy Engineering,* vol. 120, pp. 168-176, 1998.

[19] W.-S. Lee, "Benchmarking the energy efficiency of government buildings with data envelopment analysis," *Energy and Buildings,* vol. 40, pp. 891-895, 2008.

Liang X., Hong T., & Shen G.Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. Applied Energy, 179(2016), 247-260, DOI: 10.1016/j.apenergy.2016.06.141, October. (SCI, 5-Year Impact Factor: 6.222, Ranked 10/88 in Energy & Fuels, 6/135 in Chemical Engineering by JCR in 2015).

[20] W. Chung and Y. V. Hui, "A study of energy efficiency of private office buildings in Hong Kong," *Energy and Buildings,* vol. 41, pp. 696-701, 2009.

[21] W. Chung, Y. Hui, and Y. M. Lam, "Benchmarking the energy efficiency of commercial buildings," *Applied Energy,* vol. 83, pp. 1-14, 2006.

[22] A. Sabapathy, S. K. Ragavan, M. Vijendra, and A. G. Nataraja, "Energy efficiency benchmarks and the performance of LEED rated buildings for Information Technology facilities in Bangalore, India," *Energy and Buildings,* vol. 42, pp. 2206-2212, 2010.

[23] C. Martani, D. Lee, P. Robinson, R. Britter, and C. Ratti, "ENERNET: Studying the dynamic relationship between building occupancy and energy consumption," *Energy and Buildings,* vol. 47, pp. 584-591, 4// 2012.

[24] N. Li, G. Calis, and B. Becerik-Gerber, "Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations," *Automation in construction,* vol. 24, pp. 89-99, 2012.

[25] A. L. Pisello and F. Asdrubali, "Human-based energy retrofits in residential buildings: A cost-effective alternative to traditional physical strategies," *Applied Energy,* vol. 133, pp. 224-235, 11/15/ 2014.

[26] F. Oldewurtel, D. Sturzenegger, and M. Morari, "Importance of occupancy information for building climate control," *Applied Energy,* vol. 101, pp. 521-532, 1// 2013.

[27] X. Feng, D. Yan, and T. Hong, "Simulation of occupancy in buildings," *Energy and Buildings,* vol. 87, pp. 348-359, 1/1/ 2015.

[28] T. Hong, S. C. Taylor-Lange, S. D'Oca, D. Yan, and S. P. Corgnati, "Advances in research and applications of energy-related occupant behavior in buildings," *Energy and Buildings*.

[29] D. Yan, W. O'Brien, T. Hong, X. Feng, H. Burak Gunay, F. Tahmasebi*, et al.*, "Occupant behavior modeling for building performance simulation: Current state and future challenges," *Energy and Buildings,* vol. 107, pp. 264-278, 11/15/ 2015.

[30] E. Rey, "Office building retrofitting strategies: multicriteria approach of an architectural and technical issue," *Energy and Buildings,* vol. 36, pp. 367-372, Apr 2004.

[31] B. Gucyeter and H. M. Gunaydin, "Optimization of an envelope retrofit strategy for an existing office building," *Energy and Buildings,* vol. 55, pp. 647-659, Dec 2012.

[32] E. Miller and L. Buys, "Retrofitting commercial office buildings for sustainability: tenants' perspectives," *Journal of Property Investment & Finance,* vol. 26, pp. 552-561, 2008.

[33] N. Miller, J. Spivey, and A. Florance, "Does green pay off?," *Journal of Real Estate Portfolio Management,* vol. 14, pp. 385-400, 2008.

[34] J. A. Wiley, J. D. Benefield, and K. H. Johnson, "Green Design and the Market for Commercial Office Space," *Journal of Real Estate Finance and Economics,* vol. 41, pp. 228-243, Aug 2010.

[35] EEBHUB. (Dec 17). *Energy Efficient Buildings Hub.* *http://www.buildsci.us/eeb-hub.html*. Available: http://www.buildsci.us/eeb-hub.html

[36] X. Zhou, D. Yan, T. Hong, and X. Ren, "Data analysis and stochastic modeling of lighting energy use in large office buildings in China," *Energy and Buildings,* vol. 86, pp. 275-287, 1// 2015.

[37] K. Sun, D. Yan, T. Hong, and S. Guo, "Stochastic modeling of overtime occupancy and its application in building energy simulation and calibration," *Building and Environment,* vol. 79, pp. 1-12, 9// 2014.