BERKELEY LAB

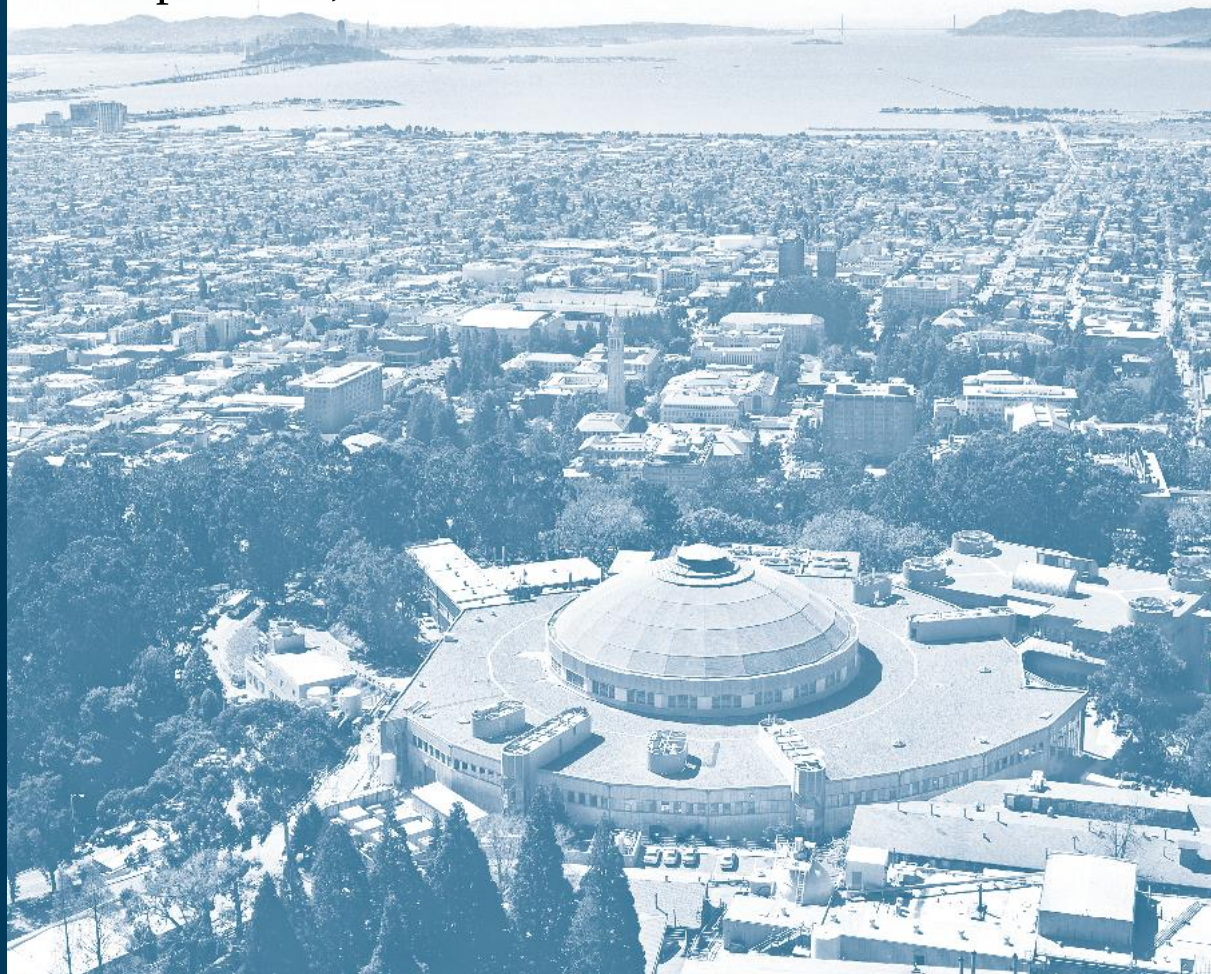
Lawrence Berkeley National Laboratory

Occupancy data analytics and prediction: a case study

Xin Liang, Tianzhen Hong,
& Geoffrey Qiping Shen

Lawrence Berkeley National Laboratory

Energy Technologies Area
September, 2016



Occupancy data analytics and prediction: a case study

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

Abstract

Occupants are a critical impact factor of building energy consumption. Numerous previous studies emphasized the role of occupants and investigated the interactions between occupants and buildings. However, a fundamental problem, how to learn occupancy patterns and predict occupancy schedule, has not been well addressed due to highly stochastic activities of occupants and insufficient data. This study proposes a data mining based approach for occupancy schedule learning and prediction in office buildings. The proposed approach first recognizes the patterns of occupant presence by cluster analysis, then learns the schedule rules by decision tree, and finally predicts the occupancy schedules based on the inducted rules. A case study was conducted in an office building in Philadelphia, U.S. Based on one-year observed data, the validation results indicate that the proposed approach significantly improves the accuracy of occupancy schedule prediction. The proposed approach only requires simple input data (i.e., the time series data of occupant number entering and exiting a building), which is available in most office buildings. Therefore, this approach is practical to facilitate occupancy schedule prediction, building energy simulation and facility operation.

Keywords: occupancy prediction; occupant presence; data mining; machine learning.

1 Department of Building and Real Estate, Hong Kong Polytechnic University, Hong Kong, China. Email: xin.c.liang@connect.polyu.hk

2* Building Technology and Urban Systems Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. Email: thong@lbl.gov, phone: 1(510) 486-7082

1 Introduction

Buildings are responsible for the majority of energy consumption and greenhouse gas (GHG) emissions around the world. In the United States (U.S.), buildings consume approximately 40% of the total primary energy [1]; while in Europe, the ratio is also about 40% [2]. In the last few decades, building energy consumption has continued to increase， especially in developing countries. In China, building energy consumption increased by more than 10% annually [3]. Large-scale commercial buildings have high energy use intensity, which can be up to 300 kWh/m^2 and 5-15 times of that in residential buildings [4]. Office buildings accounted for approximately 17% of the energy use in the U.S. commercial building sector [5]. Therefore, office buildings play an important role in total energy consumption around the world.

Occupant behavior is considered a critical impact factor of energy consumption in office buildings. Numerous previous studies emphasize the role that occupants play in influencing the energy consumption in buildings and the expected energy savings if occupant behavior was changed [6-8]. Masoso and Grobler [7] indicated that more energy is used during non-working hours (56%) than during working hours (44%), mainly due to occupants leaving lights and equipment on at the end of the day. More studies proved that different occupant behaviors can affect more than 40% of energy consumption in office buildings [9, 10]. Azar and Menassa [6] opined energy conservation events, which improve energy saving behaviors, can save 16% of electricity in the building.

Occupant behavior is likewise a critical impact factor of energy simulation and prediction for office buildings. Numerous simulation models and platforms have been developed and are widely used to predict building energy consumption during the design, operation and retrofit phases. However, the differences between real energy consumption and estimated value are typically more than 30% [11]. In some extreme cases, the difference can reach 100% [12].

The International Energy Agency's Energy in the Buildings and Communities Program (EBC) Annex 53: "Total Energy Use in Buildings: Analysis & Evaluation Methods" identified six driving factors of energy use in buildings: (1) climate, (2) building envelope, (3) building energy and services systems, (4) indoor design criteria, (5) building operation and maintenance, and (6) occupant behavior. While the first five factors have been well addressed, the uncertainty of occupant presence and variation of occupant behavior are considered main reasons of prediction deviations [12, 13].

Owing to the significant impacts on energy consumption and prediction in buildings, a number of studies focused on the occupant's energy use characteristics, which is defined as the presence of occupants in the building and their actions to (or do not to) influence the energy consumption [14]. D'Oca and Hong [15] observed and identified the patterns of window opening and closing behavior in an office building. Zhou, et al. [16] analyzed lighting behavior in large office buildings based on a stochastic model. Zhang, et al. [17] simulated occupant movement, light and equipment use behavior synthetically with agent-based models. Sun, et al. [18] investigated the impact of overtime working on energy consumption in an office building. Azar and Menassa [6] showed the education and learning effect of energy saving behavior, and proposed the impacts of energy conservation promotion on energy saving.

Before modelling occupant's energy use characteristics, there is a more essential research question: how to identify the pattern of occupant presence and predict the occupancy schedule? Without the answer to this question, the occupant's energy use characteristics cannot get down to the ground. However, due to the highly stochastic activities and insufficient data, it is difficult to observe and predict occupant presence. Previous studies did not pay enough attention to occupancy schedule and this question has not been well addressed. In general, three typical methods were applied to model occupant presence in

previous studies. First method is fix schedules. Occupants are categorized into several groups (e.g., early bird, timetable complier and flexible worker), then each group is assigned to a specific schedule [17]. Combining the schedules of each group proportionally can generate the schedule of the whole building. The second method assumes that occupant presence satisfies a certain probability distribution. The distribution can be Poisson distribution [16], binomial distribution [18], uniform distribution and triangle distribution [19]. The occupancy schedule can be obtained by a virtual occupant generation following the certain distribution. The third method is analyzing practical observation data. D'Oca and Hong [8] observed 16 private offices with single or dual occupancy and Wang, et al. [20] observed 35 offices with single occupancy.

Although these methods had advantages and improved occupancy schedule modeling, there are still some limitations: (1) the assumptions are not solid. Occupancy schedule is highly stochastic, it is inappropriate to simply define that occupants belong to a certain group or follow a certain distribution; (2) the previous research emphasized on summarizing rules of occupant presence, but less attention has been paid to predicting schedules in future. The results are not practical if they cannot guide future work; (3) the results of schedules lack validation with real data; (4) observed data mainly focused on a single or multiple offices, so the data are limited and results may be biased if applied to the whole building.

To bridge the aforementioned research gaps, this study proposes a data mining based approach to learning and predicting occupancy schedule for the whole building. Data mining can be defined as: "The analysis of large observation data sets to find unsuspected relationships and to summarize the data in novel ways so that owners can fully understand and make use of the data" [21]. Data mining methods have significant advantages in revealing underlying patterns of data, which has been widely used in various research and industry fields, such as marketing, biology, engineering and social science [22]. However, the

applications of data mining in occupancy schedule and building energy consumption is still underdeveloped. Some previous studies applied data mining methods to discover the pattern of occupant behavior [15, 23, 24], and others focused on interactions between occupants and energy consumption [8, 25, 26]. These studies demonstrated the strong power of data mining methods in recognizing pattern of occupant behavior and energy consumption areas, but the research area of occupancy schedule leaning and predicting still needs exploration.

The aim of this study is to present a new approach for occupancy schedule learning and predicting in office buildings by using data mining based methods. The process of this study includes recognizing the patterns of occupant presence, summarizing the rules of the recognized patterns and finally predicting the occupancy schedules. This study hypothesizes the identified patterns and rules by the proposed data mining approach are right. Namely, they can present the true characteristics of the occupancy data. This hypothesis is validated by comparing the accuracy of prediction between the proposed method and the traditional methods. If the accuracy of the prediction results is improved, it indicates the hypothesis is true.

This model only needs a few types of inputs, typically the time series data of occupant number entering and exiting a building. Another advantage of this model is that it allows for relatively simple operations, excluding probability distribution fitting and other complex mathematical processing. That means this method can be well adaptive to practical projects. The results of this study are critical to provide insight into the pattern of occupant presence, facilitate the energy simulation and prediction as well as improve energy saving operation and retrofit.

2 Methodology

2.1 Framework of occupancy schedule learning and prediction

Traditional methods of transforming data to knowledge normally used statistical tests, regression and curve fitting by a certain probability distribution. These methods are effective when data is small volume, accurate and standardized. However, when the volume of data is growing exponentially in recent years, these methods become slow and expensive. More seriously, when there is considerable missing data, the deviated data or the data format is disunion (e.g. the time steps are different, mix of numbers and words), these methods cannot be applied or cannot deduce satisfied results. Data mining is an emerging method which can process big data and unstructured data effectively and robustly. Machine learning, as a main method of data mining, is specifically good at identifying patterns and inducting rules. Since this study includes huge volume of data and aims to induct rules of occupancy schedules, data mining is selected as the research method.

Data mining, which is also named knowledge discovery in databases (KDD), is a relatively young and interdisciplinary field of computer science. It is the process of discovering new patterns from large data sets, involving methods at the intersection of pattern recognition, machine learning, artificial intelligence, cloud architecture, and data visualization [27]. Normally, the process of KDD involves six steps: (1) Data selection; (2) Data cleaning and preprocessing; (3) Data transformation; (4) Data mining; (5) Data interpretation and evaluation; and (6) Knowledge extraction [8].

This study proposes a data mining based approach to discover occupancy schedule patterns and extrapolate occupancy schedule from observed big data streams of a building. The framework of this proposed method includes six steps, illustrated in Figure 1.

Step 1: problem framing. The first step is to clarify problem definition, boundary, assumption and key metric of success. The research problem is defined as how to predict occupancy schedule from historical observed data. The scope of this study focuses on the schedule

prediction for weekdays in office buildings. The key metric of success is the similarity of prediction results to the observed data.

Step 2: data acquisition and preparation. The second step is to acquire, harmonize, rescale, clean and format data. Due to the failure of sensors and other interference factors, the raw data may contain missing data, error data and the unstructured data. Before data mining, the raw data should be pre-processed to get the valid data. In this study, the missing data is removed from the data set. Statistical methods (i.e., box plot and mean value) are used to investigate the characteristics of the data before data mining.

Step 3: methodology selection. Data mining involves various kinds of methods. Different methods target problems at different levels. According to the specific problem and data source, appropriate methods could be selected. In this study, machine learning method is adopted to discover patterns of occupant presence, while rule induction is used to summarize rules within the patterns. Software selection is essential to analyze data. Matlab 2015 and RapidMiner 6.5 are applied on a standard PC with Windows 7 to perform the data processing and data mining, respectively. RapidMiner is open source software with visualized interface and modularized operation for analytics and data mining. Due to its flexibility and accessibility, RapidMiner has been widely used in industry and academia.

Step 4: learning. This step is to discover the patterns of occupancy schedule and abstract the rules within the patterns. Clustering and decision tree are applied for pattern recognition and rule induction respectively. The details of processes and results of each step are illustrated in the learning phase in Figure 2.

Step 5: prediction. The observed data is split to a training set and a test set. The training set is used to train the model and identify the rules, shown in the predicting phase in Figure 2.

Based on the identified patterns and rules of occupant presence, the occupancy schedule can be predicted.

Step 6: validation. This step is to compare the prediction result to the test data set, shown in the validating phase in Figure 2. The more similar the two sets are, the better the method is. To quantitatively validate the proposed method, several metrics can be applied to measure similarity between prediction results and observed data, including mean, median, bias, RMSE (root mean squared error) and RTE (relative total error). The details of the metrics and validation will be introduced in Section 3.5.

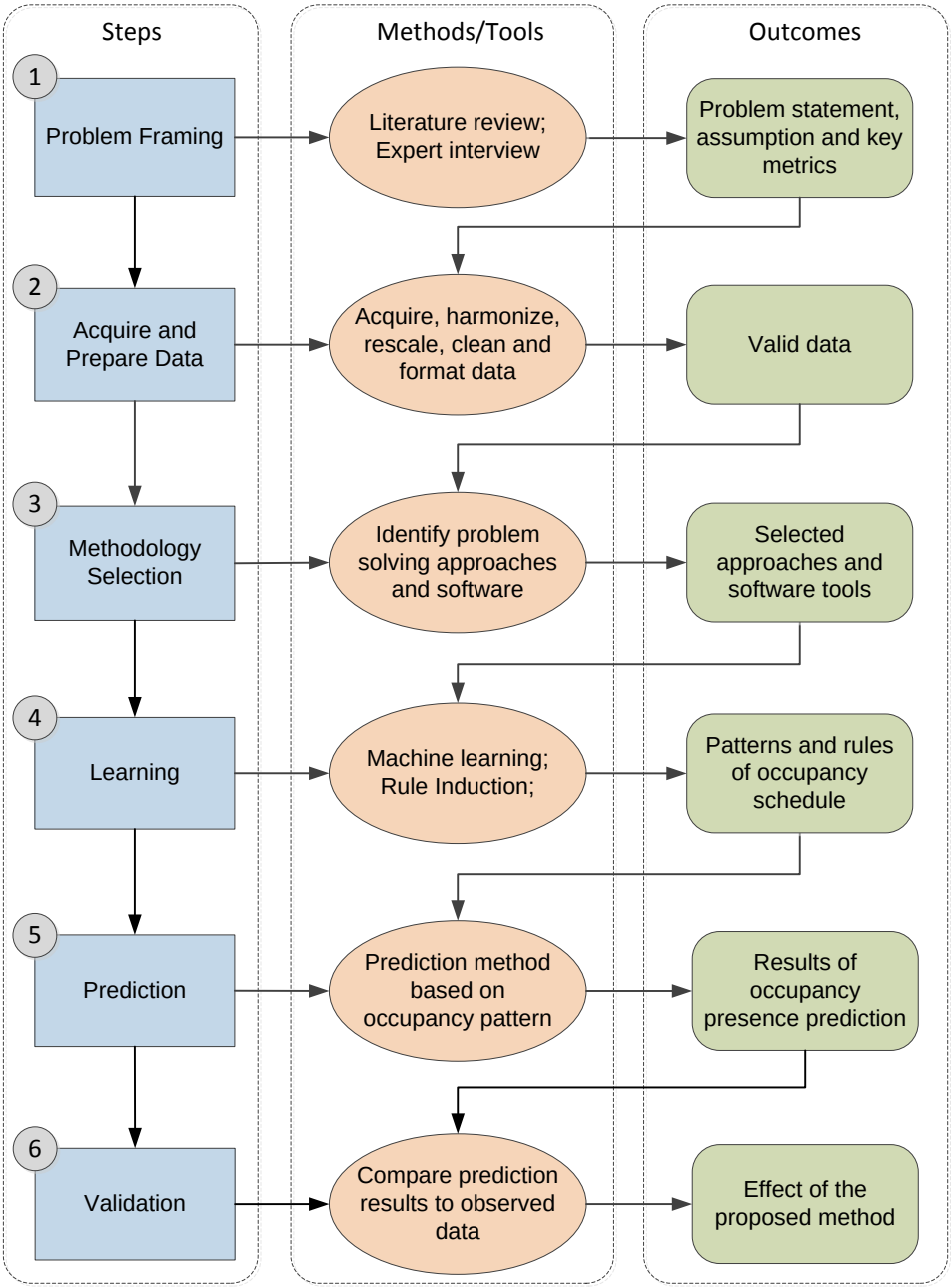


Figure 1 Framework of the proposed method for occupancy schedule learning and predicting

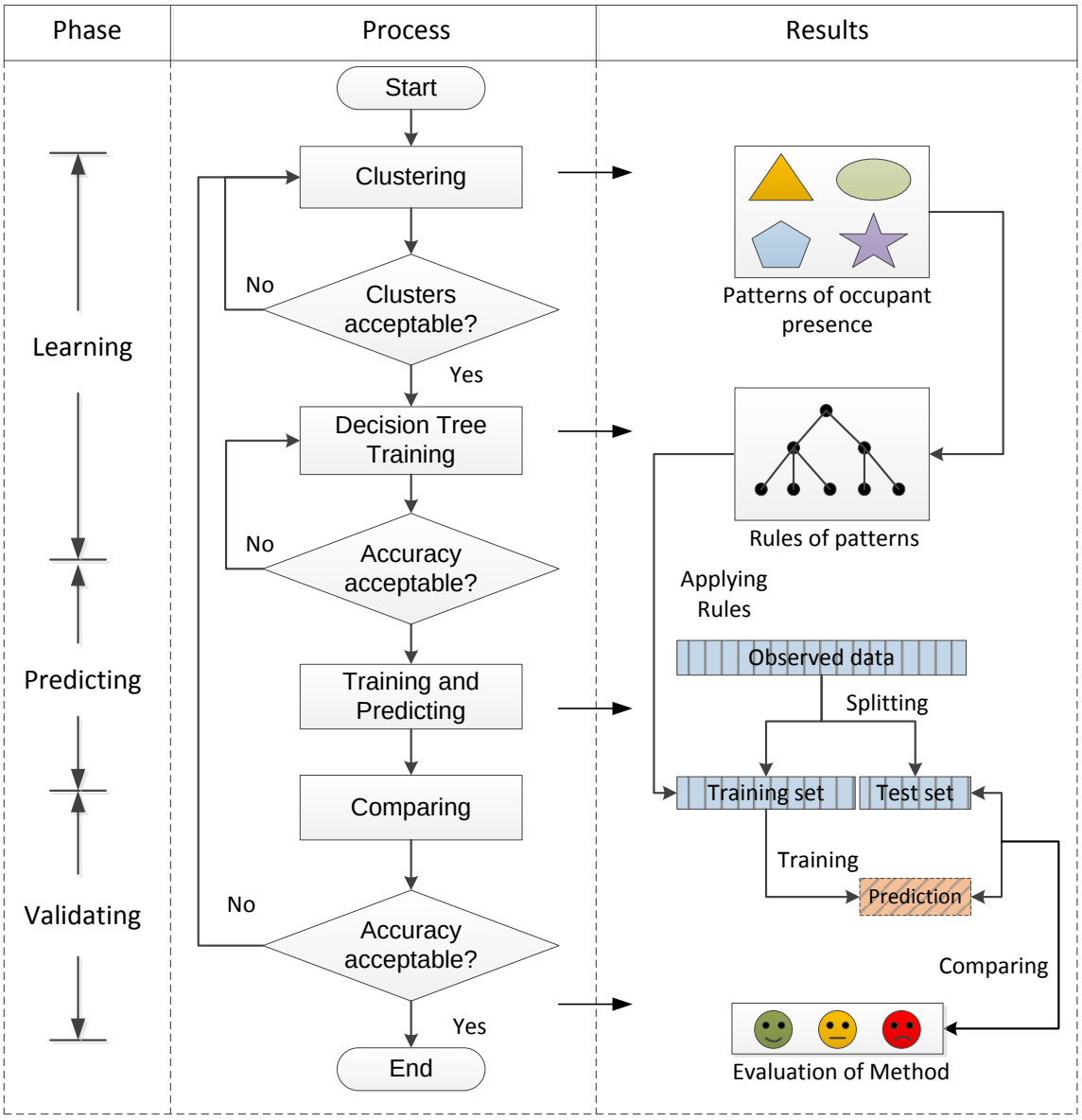


Figure 2 Processes of the proposed method and results

2.2 Machine learning

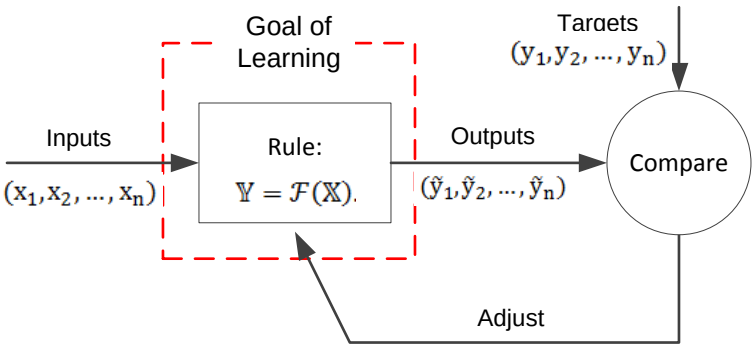
Machine learning is an important method of data mining [27], which allows computers to learn from and make predictions on data via observation, experience, analysis and self-training [27, 28]. It operates by building a model to make data-driven predictions or decisions, rather than following strictly static program instructions [29].

There are two types of machine learning, namely supervised learning and unsupervised learning [30]. The former one refers to the traditional learning methods with training data,

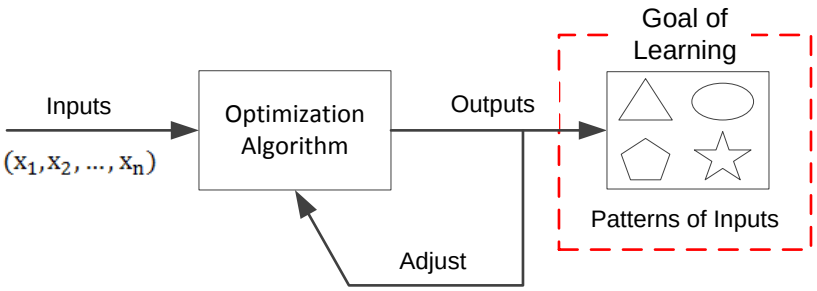
which is a known labeled data set of inputs and outputs. As a standard supervised learning problem, training samples $(\mathbb{X}, \mathbb{Y}) = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ are offered for an unknown function $\mathbb{Y} = \mathcal{F}(\mathbb{X})$. $\mathbb{X}$ denotes the "input" variables, also called input features, and $\mathbb{Y}$ denotes the "output" or target variables that trying to predict. The x_i values are typically vectors of the form $(x_{i1}, x_{i2}, \ldots, x_{in})$ which are the features of x_i, such as weight, color, shape and so on. The notation x_{ij} refers to the *j-th* feature of x_i. The goal of supervised learning is to learn a general rule $\mathcal{F}(\mathbb{X})$ that maps inputs $\mathbb{X}$ to outputs $\mathbb{Y}$, shown in Figure 3 (a). The typical algorithms of supervised learning include regression, Bayesian statistic, decision tree and etc.

The unsupervised learning refers to the methods without given labels to the learning algorithm, leaving it on its own to find structure in its input. In unsupervised learning, there is no "output" $\mathbb{Y}$ to train the function $\mathcal{F}(\mathbb{X})$. The goal of unsupervised learning is to discover hidden patterns in the input data $\mathbb{X}$ by its own features, shown in Figure 3 (b). In reality, numerous problems cannot obtain priori information of outputs. Therefore, unsupervised learning is widely used to solve this kind of problems recently.

This study uses both the supervised learning and the unsupervised learning in two steps. At the beginning, there is no label of occupancy schedule data, so the unsupervised learning method (i.e., clustering) is applied to identify patterns of occupant presence from the features of data. After that, the presence data have labels, which are the identified patterns. Then, the supervised learning method (i.e., decision tree) is applied to induct rules based on the labeled data.



(a) Supervised learning



(b) Unsupervised learning

Figure 3 Mechanism of machine learning

2.2.1 Cluster analysis

Cluster analysis is a typical unsupervised machine learning method, which aims to group data into a few cohesive clusters [31]. The criterion of clustering is the similarities among samples. The samples should have high similarities within the same cluster but low similarities in different clusters. The similarity is normally measured by distance. The shorter the distance between samples is, the more similar the samples are. There are various distance definitions, including the Euclidian distance, the Chebyshev distance, the Hamming distance, the dynamic time wrap distance and the correlation distance [32]. Appropriate distance type should be selected according to the specific problem. For example, The Euclidian distance is

commonly used for the direct geometrical distance. The correlation distance is good at triangle similarity. The dynamic time wrap is commonly used for the similarity of time-shift sequences. This study compares three kinds of distances, shown in Figure 10, and selects the Euclidian distance due to its best performance.

There are various clustering models, and for each of these models, different algorithms can be given [33]. Typical cluster models include connectivity based models (e.g., hierarchical clustering), centroid based models (e.g., k-means clustering), distribution based models (e.g., Gaussian distributions fitting) and density based models (e.g., Density-based spatial clustering of applications with noise) [34]. Among numerous clustering algorithms, the k-means clustering is the most commonly used, which is defined as follows.

1. Initialize cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}$
2. Repeat until convergence: {

   For every i, set

$$c_i = arg\ min_j \left\| \mathrm{x}_i - \mu_j \right\| \tag{1}$$

   For every j, set

$$\mu_j = \frac{\sum_{i=1}^{m} \alpha \cdot \mathrm{x}_i}{\sum_{i=1}^{m} \alpha}, \alpha = \begin{cases} 1 & if\ c_i = j \\ 0 & if\ c_i \neq j \end{cases} \tag{2}$$

   }

In the k-means algorithm, k (a parameter of the algorithm) is the preset number of clusters. The cluster centroids μ_j represent the positions of the centers of the clusters. Step 1 is to initialize cluster centroids, randomly or by a specific method. Step 2 is to find optimal cluster centroids and samples assigned to them. Two operations are implemented iteratively until convergence in this step. One operation is assigning each training sample x_i to the closest

cluster centroid μ_j, shown in Equation (1). The other one is moving each cluster centroid μ_j to the mean of the points assigned to it, shown in Equation (2).

The appropriate clustering algorithm for a particular problem needs to be chosen experimentally, since there is no defined "best" clustering algorithm [33]. The most appropriate algorithm for a certain problem can be selected by its performance. The performance of algorithms can be measured by the definition of clusters, namely the proportion of intra-cluster distance to inter-cluster distance. The Davies-Bouldin index (DBI) is used to evaluate different methods in this study. This index is defined in Equation (3).

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \tag{3}$$

where n is the number of clusters, c_i is the centroid of cluster i, σ_i is the average distance of all elements in cluster i to centroid c_i, and $d(c_i, c_j)$ is the distance between centroids c_i and c_j. The lower value of DBI means lower intra-cluster distances (higher intra-cluster similarity) and higher inter-cluster distances (lower inter-cluster similarity), therefore, the clustering algorithm with the smallest DBI is considered the best algorithm based on this criterion.

2.2.2 Decision tree learning

This study uses decision tree to induce the rules of occupant presence. Decision tree learning is a typical supervised machine learning algorithm in data mining [35]. It uses a tree-like structure to model the rules and their possible consequences. A main advantage of decision tree method is that it can represent the rules visually and explicitly. Figure 4 illustrates the structure of decision tree model, which includes three types of nodes (i.e., root node, leaf node and terminal node) and branches between nodes. The leaf nodes denote attributes of input, while branches denote the condition of these attributes. Each terminal node is a subset

of target variables $\mathbb{Y}$, which indicates two kinds of information: (1) classification of the target variables $\mathbb{Y}$, and (2) the probability of each subset. Based on the classification and probability, the rules of prediction can be inducted.

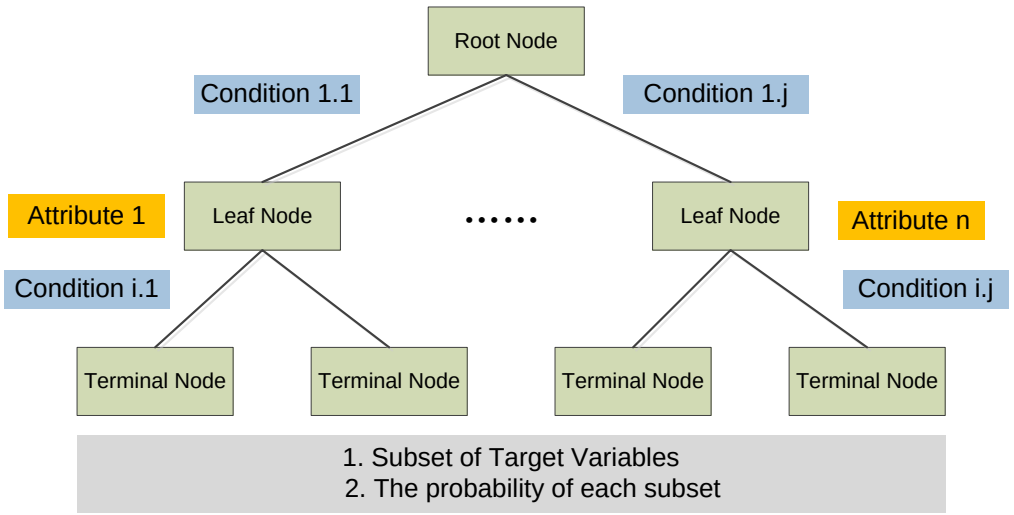


Figure 4 Graphical structure of decision tree model

Most algorithms for generating decision trees are variations of a core algorithm that employs a top down, greedy search through the entire space of possible decision trees. ID3 algorithm [36] and its successor C4.5 [37] are the most used methods. The key of these algorithms is the choice of the best attribute in each node. To measure the classification effect of a given attribute, a metric is defined, called *information gain*, which can be defined as follows [38].

$$Gain(S,A) = Entropy(S) - \sum_{\upsilon \in Values(A)} \frac{|S_\upsilon|}{|S|} Entropy(S_\upsilon) \tag{4}$$

$$\text{where } Entropy(S) = \sum -p_i \log_2 p_i \tag{5}$$

$Gain(S,A)$ represents the information gain of an attribute A related to a collection of samples S. $Values(A)$ is the set of all possible values of attribute A, and S_υ is the subset of S, which contains attribute A has value υ, namely $S_\upsilon = \{s \in S \,|\, A(s) = \upsilon\}$. p_i represents the proportion of S belonging to class i, and *Entropy* is a measure of the impurity in a collection

of training set. Given the definition of *Entropy*, the $Gain(S, A)$ in Equation (4) is the reduction in entropy caused by the knowledge of attribute A. Namely, $Gain(S, A)$ is the contribution of attribute A to the information of samples S. The highest value of information gain indicates the best attribute A in a specific node.

There are two steps of decision tree generation. First step is learning rules from training data based on the aforementioned C4.5 algorithm. Gain ratio method is employed to identify the best attribute in each node by minimizing the entropy. The confidence is set to 0.25 and the minimal gain is 0.1. The second step is predicting based on the rules learned from the first step, and validating results by testing data. If the accuracy is satisfied, the process is finished. Otherwise, the two steps are repeated to update decision tree until the result is satisfied. Cross-validation method [8] is used to evaluate the performance of decision tree in this study. The data set is divided into ten subsets. Seven subsets are used for training and the other three are used for testing. Then it repeats by exchanging subsets. The cross-validation can improve the accuracy and robustness of decision tree model.

2.3 Case study

A case study was conducted to demonstrate the proposed method. The office building of the case study is the Building 101 in the Navy Yard, Philadelphia, U.S., shown in Figure 5. The building is one of the nation's most highly instrumented commercial buildings. Building 101 in the Navy Yard is the temporary headquarters of the U.S. Department of Energy's Energy Efficient Building Hub (EEB Hub) [39]. 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 1.

Table 1 The profile of Building 101

Location	Philadelphia, U.S.
Size	6,410 m^2

Floor	3 floors
Constructed Year	1911
Building Usage	Office

Figure 5 Photo of Building 101

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 of Building 101, shown in Figure 6. The data format of raw sensor records is shown in Table 2. The set ($N_{i1,}$ $N_{i3,}$ $N_{i5,}$ N_{i7}) denotes the number of entering occupants, while the set ($N_{i2,}$ $N_{i4,}$ $N_{i6,}$ N_{i8}) denotes the number of exiting occupants at the *i-th* time step. Therefore, the number of total occupants in building at the *i-th* time step can be calculated by Equation (6).

$$N_{total} = \sum_{1}^{i}(N_{i1} - N_{i2} + N_{i3} - N_{i4} + N_{i5} - N_{i6} + N_{i7} - N_{i8}) \quad (6)$$

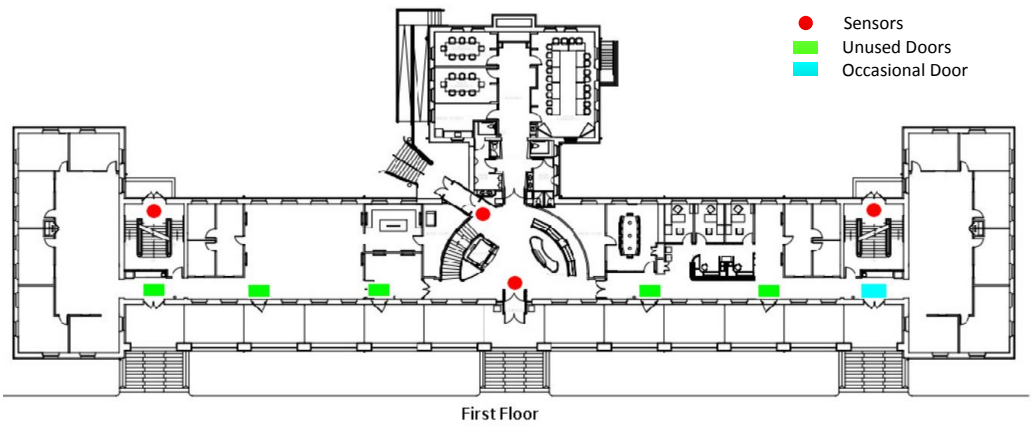


Figure 6 Sensor locations in Building 101

Table 2 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_{11}	N_{12}	N_{13}	N_{14}	N_{15}	N_{16}	N_{17}	N_{18}
1/1/2014 0:05	…..	…..	…..	…..	…..	…..	…..	…..
1/1/2014 0:10	…..	…..	…..	…..	…..	…..	…..	…..
…..	…..	…..	…..	…..	…..	…..	…..	…..
12/31/2014 23:50	…..	…..	…..	…..	…..	…..	…..	…..
12/31/2014 23:55	N_{i1}	N_{i2}	N_{i3}	N_{i4}	N_{i5}	N_{i6}	N_{i7}	N_{i8}

3 Results

3.1 General characteristics of occupant presence

This study uses the data from the year 2014 and the time step is five minutes. Due to the sensor failure and other reasons, there are some missing data, which is less than 1% of all samples. Based on the measured data of Building 101, general characteristics of the occupant presence were analyzed and compared among different conditions based on statistical method.

The daily 24-hour profile of occupant presence is the main target of this study. First, the hourly occupant presence of weekday and weekend is shown in Figure 7. The results show that the mean of occupant number is close to zero in the building during weekends and holidays and the variance is also low. It means there are normally few occupants in weekend and holiday. Therefore, when analyzing the occupancy schedule, this study excludes the data from weekend and holidays. In weekdays, the mean of occupant number is significantly changed over time. The variation range of occupant number is very large from 7 am to 4 pm in weekdays, which exceeds more than 30% of the mean. It indicates the main characteristics of occupant presence, dynamic, stochastic and highly variable. These characteristics lead to difficulty to understand and predict occupant presence based on traditional statistical methods.

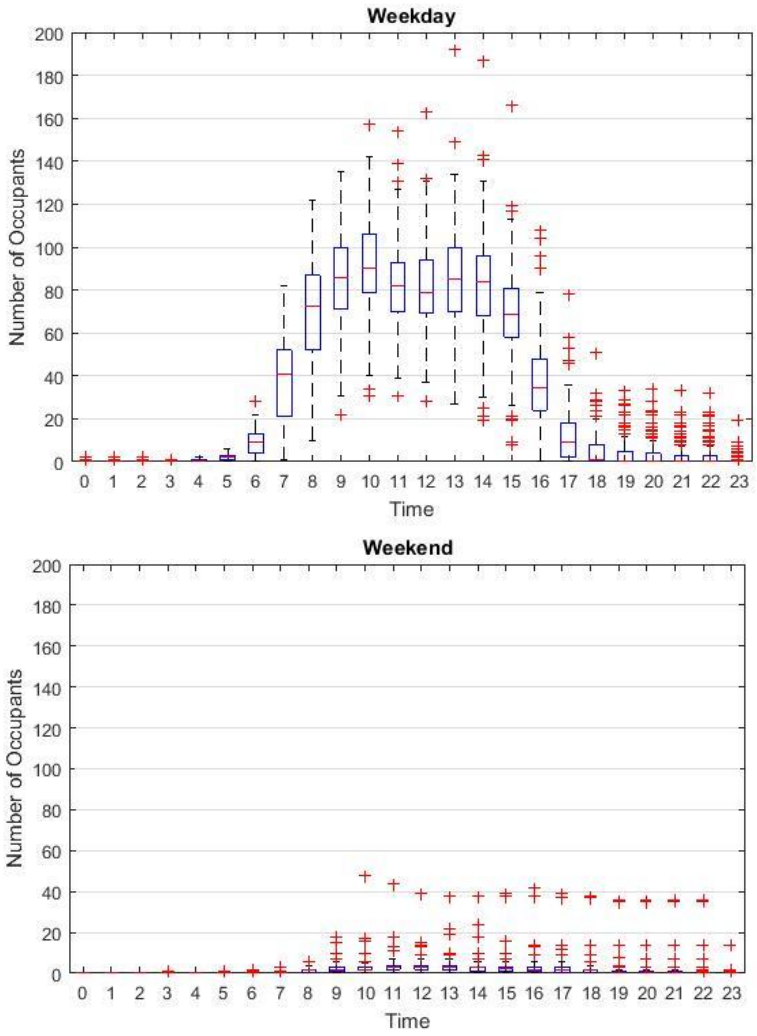


Figure 7 Hourly occupant presence during weekdays and weekends

Statistical results of hourly occupant presence from Monday to Friday are compared in Figure 8. It shows the features of each weekday are different. For example, the variance range at 11 am is much smaller on Tuesday and Thursday than Monday and Wednesday. The particular values (extremely high values) on Friday are significantly lower than that of the other four days. Although the occupancy features are different in each weekday, the averages of hourly occupant presence in each weekday are very similar except Friday. It indicates that traditional method, which only uses mean value to describe occupant presence (Figure 9), loses granularity of information.

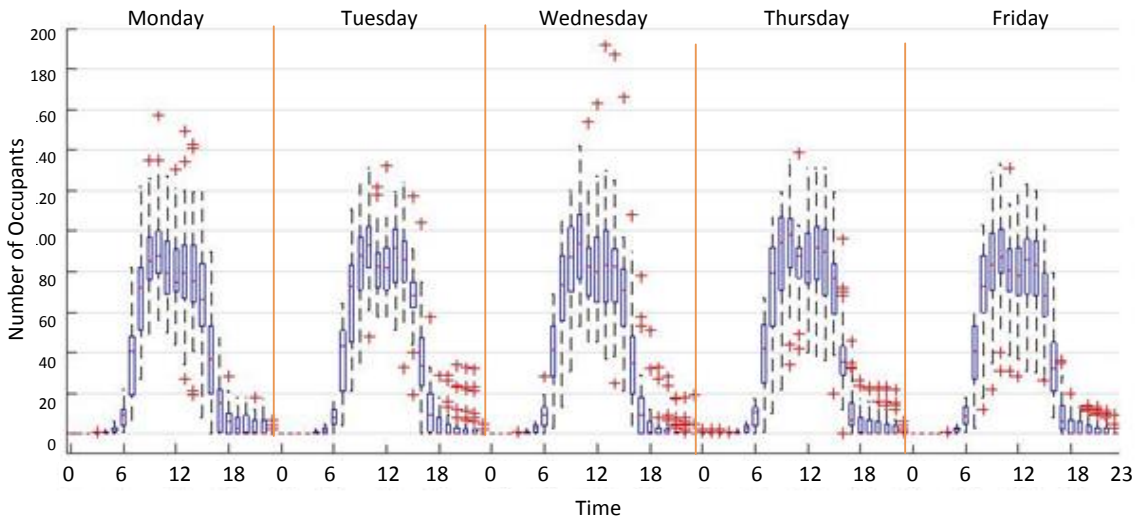


Figure 8 Hourly occupant presence from Monday to Friday

Figure 9 shows occupant presence in Building 101 has dual-peak feature (mainly due to occupants going out for lunch), which is similar to occupant schedules used in ASHRAE standard 90.1 [40]. It verifies the occupancy data in this case is not abnormal and has general adaption. But the peak in the afternoon is a bit lower than that in the morning (the peaks in morning and afternoon are the same in ASHRAE standard). In addition, the drop at noon is not as sharp as that in ASHRAE standard 90.1, and the slopes are likewise different. Therefore, ASHRAE standard schedule is not adaptable to variable buildings, it is necessary to adjust occupancy factor according to the data of a particular building.

The occupant presence curve can be divided into six periods:

- The night period (7 pm to 6 am): Few occupants are in the building, typically no occupant. The occupancy rate is normally less than 10% of the max value.
- The going-to-work period (7 am to 9 am): Occupants are arriving successively in this period. The occupancy rate is growing from 10% to 70%.
- The morning period (10 am to 12 pm): Occupants are working in the building and the occupancy rate stays around 80%.

- The noon-break period (12 pm to 1 pm): some occupants go out for lunch and the occupancy rate drops slightly to lower than 80%.
- The afternoon period (2 pm to 3 pm): Occupants are back to work in the building. The occupancy rate rises slightly higher than 80%, but is lower than that in the morning period.
- The going-home period (4 pm to 6 pm): Occupants are leaving office successively in this period. The occupancy rate is decreasing from 70% to 10%.

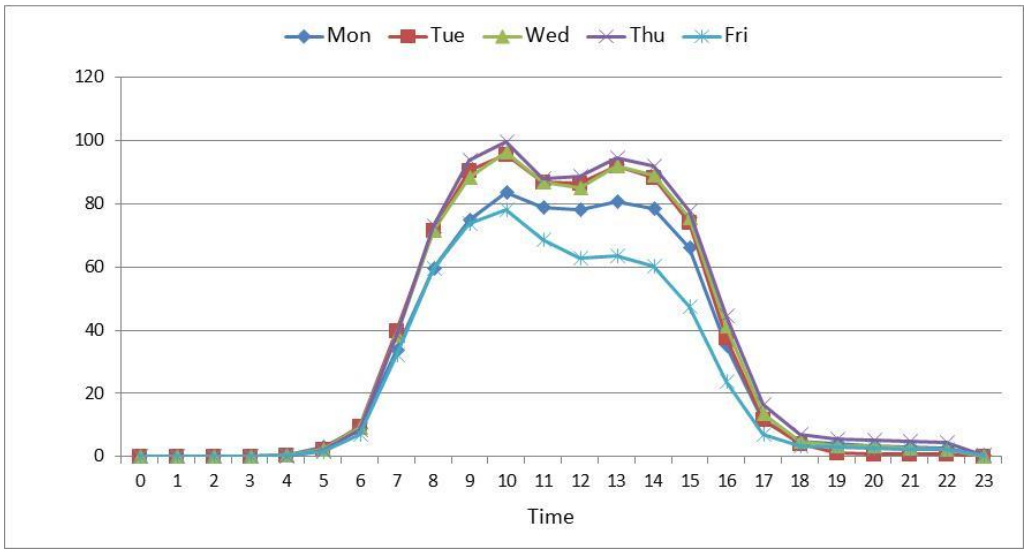


Figure 9 Mean of hourly occupants presence of weekdays

3.2 Patterns of occupant presence

This step is to discover the pattern of occupant presence during weekdays. The data mining software RapidMiner 6 is applied to disaggregate presence data to several clusters. In this study, BDI is used to find the optimal different k value in the k-means algorithm and distance metric. The k values are evaluated from 2 to 8 and the distance metrics are compared among Euclidean distance, correlation similarity and dynamic time wrap. The results indicate that k=4 with Euclidean distance metric is the optimal parameter in k-means algorithm for this data set, shown in Figure 10.

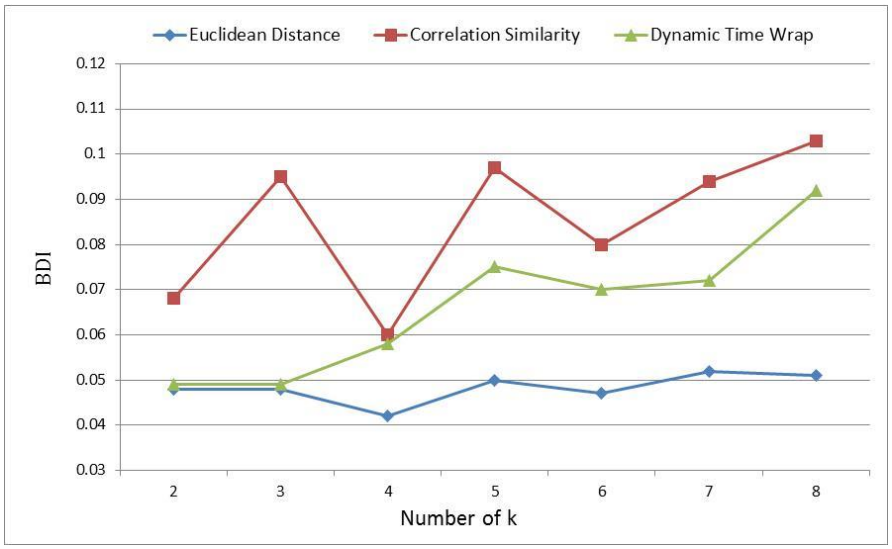


Figure 10 Performance of k and distance metrics evaluated by BDI

The four clusters of occupant presence data are shown in Figure 11. From the visualization of the clusters, four patterns of occupant presence are highlighted as following, and the characteristics of patterns are shown in Table 3:

- Pattern 1 represents the lowest occupancy rate and shortest working time. The occupants go to work latest and go home late in this pattern. The occupancy rate rises to 50% around early 10 am. In addition, there is no obvious noon-break drop of the curve in this pattern, since the occupant number decreases continuously since 11 am.
- Pattern 2 represents the highest occupancy rate and longest working time. The occupants go to work earliest and go home late in this pattern. The occupancy rate rises to 50% around early 8 am and decreases to 50% around 5 pm. The noon-break is around 12 pm.
- Pattern 3 represents the medium occupancy rate, medium working time, going-to-work later and going-home later. The occupancy rate rises to 50% around 9 am and decreases to 50% before 6 pm. The noon-break is around 2 pm.

- Pattern 4 is similar to Pattern 3, which likewise represents the medium occupancy rate and medium working time. But the main difference is that the going-to-work time and going-home time are about 1 hour earlier than that in Pattern 3. The occupancy rate rises to 50% around 8 am and decreases to 50% before 5 pm. The noon-break is around 1 pm.

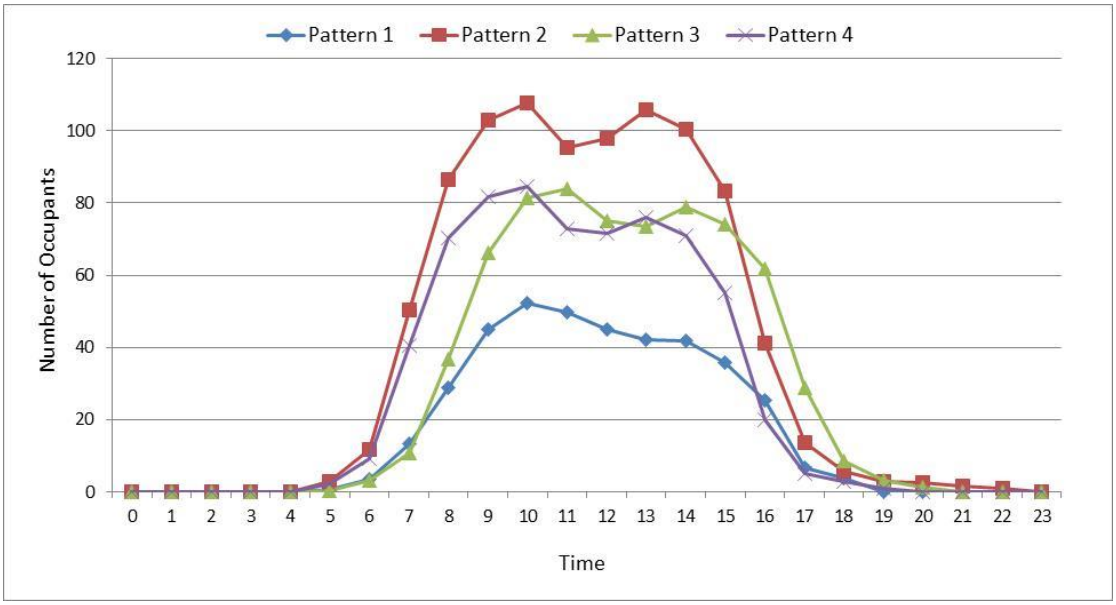


Figure 11 Patterns of occupant presence

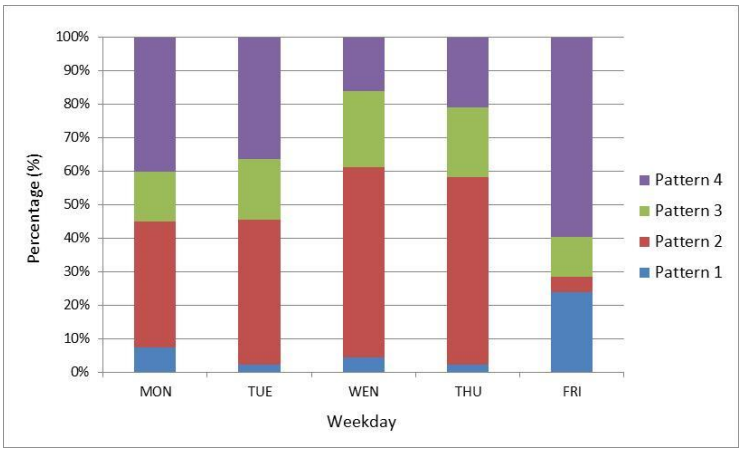
Table 3 Characteristics of occupant presence patterns

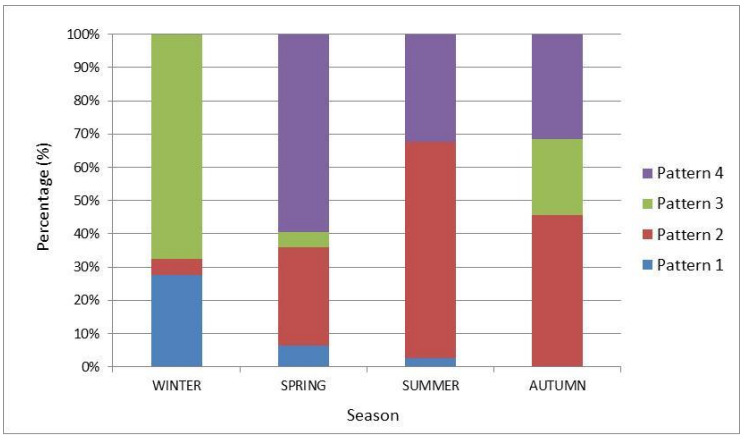
Pattern	Occupancy Rate	Working Time	Going to Work Time	Going home Time	Noon Break Time
Pattern 1	Lowest	Shortest	Latest	Earliest	NA
Pattern 2	Highest	Longest	Earliest	Later	12 pm
Pattern 3	Medium	Medium	Later	Latest	2 pm
Pattern 4	Medium	Medium	Earlier	Earlier	1 pm

3.3 Rules of patterns

Based on the recognized patterns of occupant presence, the rules of these patterns are induced in this step. According to data analysis, three influencing factors are used in the decision tree generation: the patterns are related to (1) seasons (temperatures); (2) weekdays; and (3)

daylight saving time (DST)[3]. Since the temperature information needs other data input but season information can be transformed from the existing data set (time step column), to simplify the proposed method, seasons are selected as an analysis factor. As shown in Figure 12, these three factors have strong relations with patterns. For example, most Pattern 1 happened on Friday and there is no Pattern 4 happened in winter. It means it is possible to induce the underlying rules of patterns from these factors.





3 Daylight saving time in USA starts on the second Sunday in March and ends on the first Sunday in November.

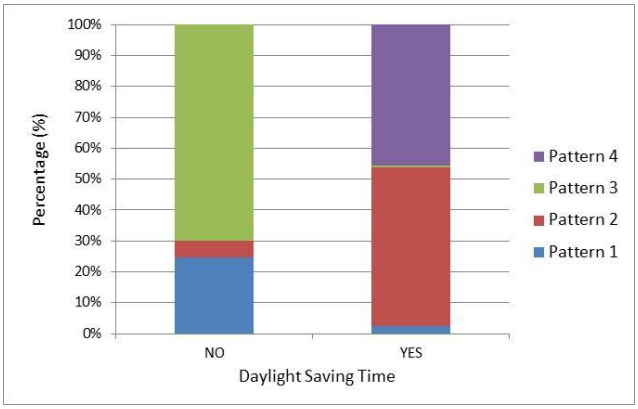


Figure 12 Relationship between occupancy patterns and weekdays, seasons and DST

Figure 13 shows the decision tree for classification of the patterns by the attributes. Any samples can be classified to different patterns top down along the path of the tree. The first decision level is season. If season is winter, the branch is to the terminal node. If not, the process will reach to the second decision level, namely weekday. After split by the weekday nodes, the final decisions can be generated. It needs to be noted that the DST is not included in the decision tree, which means DST cannot contribute enough information to reach the threshold of gain radio. Namely, DST is not a key attribute in the classification of patterns.

Not only the classification, but also the probability of the classification can be provided by the decision tree. In Figure 13, the lengths of different colors represent the probability of different patterns. For example, if the season is winter, the decision is Pattern 3. Behind this decision, there is more information of probability: the Pattern 3 is of the highest probability, Patterns 1 and 2 are of lower probabilities, and the probability of Pattern 4 is zero. Table 4 shows the rules of patterns in detail. 80% of all the training samples are correctly classified based on these rules. The result of the decision tree model shows relatively good performance to be further applied to prediction in the next step.

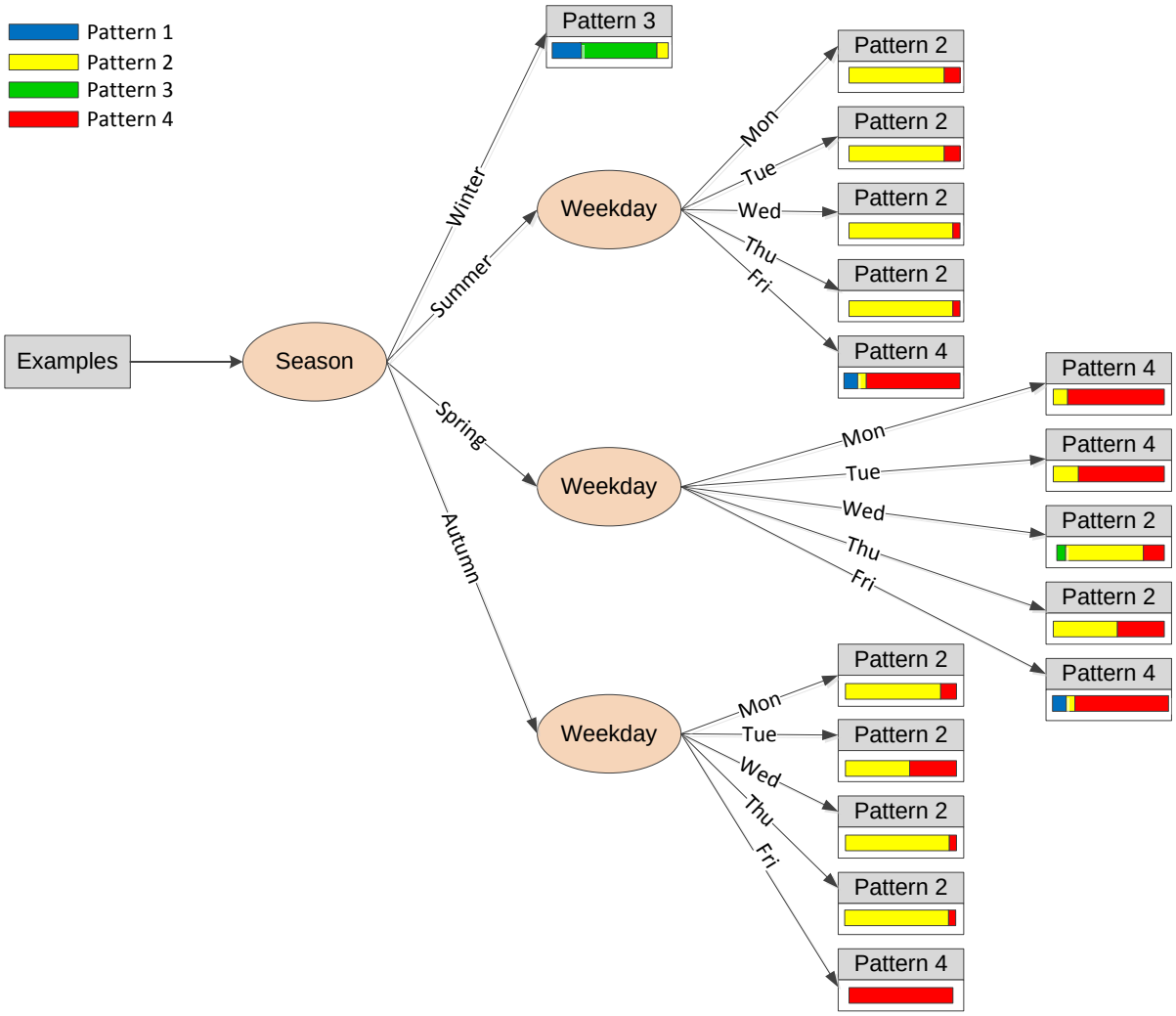


Figure 13 Decision tree for classification of occupant presence patterns

Table 4 Probabilities of patterns under different conditions

Season	Weekday	Probability of Pattern			
		Pattern 1	Pattern 2	Pattern 3	Pattern 4
Winter		28.9%	6.7%	64.4%	0.0%
Spring	Monday	0.0%	9.1%	0.0%	90.9%
	Tuesday	0.0%	25.0%	0.0%	75.0%
	Wednesday	0.0%	58.3%	8.3%	33.3%
	Thursday	0.0%	50.0%	0.0%	50.0%
	Friday	16.7%	8.3%	0.0%	75.0%
Summer	Monday	0.0%	66.7%	0.0%	33.3%
	Tuesday	0.0%	73.3%	0.0%	26.7%
	Wednesday	0.0%	86.7%	0.0%	13.3%
	Thursday	0.0%	86.7%	0.0%	13.3%

	Friday	14.3%	7.1%	0.0%	78.6%
Autumn	Monday	0.0%	50.0%	33.3%	16.7%
	Tuesday	0.0%	37.5%	25.0%	37.5%
	Wednesday	0.0%	62.5%	25.0%	12.5%
	Thursday	0.0%	71.4%	14.3%	14.3%
	Friday	0.0%	0.0%	16.7%	83.3%

3.4 Prediction of occupancy schedule

Based on the rules deduced by decision tree, the occupancy schedule can be predicted. Three prediction methods are compared in this study. The first is the mean-day method. The predictions depend only on the time of day. The method is presented by Equation (7), where t denotes the time of the day (e.g. 3 pm) and M_{day} denotes the mean value of all days. For example, the prediction for 3 pm is the average of all of the data for 3 pm in history. Therefore, there is no different profile for each day of the week, for different seasons or for other factors. This prediction method is simple and can be compared as a baseline.

$$Prdiction(t) = M_{day}(t) \tag{7}$$

The second method is mean-week method. The method is presented by Equation (8), where day denotes the day of samples and $M_{weekday}$ denotes the mean value of the assigned weekday. For example, the prediction of 3 pm on a Monday in spring is the average of all historical data for 3 pm on Monday.

$$Prdiction(weekday, t) = M_{weekday}(t) \tag{8}$$

The third method is the proposed method in this study, which is based on the probability of decision tree. The method is presented by Equation (9), where M_{pi} $(i = 1, 2, 3, 4)$ denotes the mean value of the Pattern i and P_{pi} denotes the probability of Pattern i. For example, the prediction of 3 pm on a Monday in spring is the expectation of all historical data for 3 pm based on probability of patterns.

$$Prdiction(day,t) = M_{p1}(t) \times P_{p1} + M_{p2}(t) \times P_{p2} + M_{p3}(t) \times P_{p3} + M_{p4}(t) \times P_{p4} \quad (9)$$

The visualized prediction of occupancy schedule based on the third method is shown in Figure 14. Since there are 16 terminal nodes in decision tree (Figure 13), there are 16 conditions of prediction.

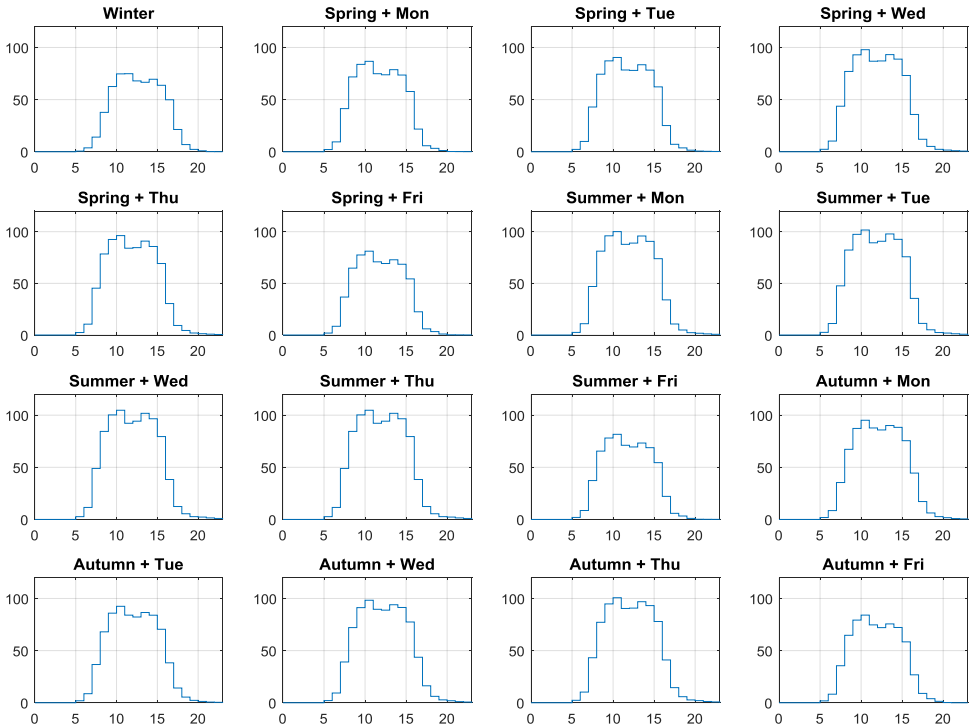


Figure 14 Prediction results based on the induced rules

3.5 Validation

Several statistical performance metrics are used to evaluate prediction. The definitions are described below.

The root mean squared error (RMSE) quantifies the typical size of the error in the predictions, in absolute units. The equation for RMSE is provided in Equation (10), where E_i is the observed data of occupants, $\hat{E}_i$ is the prediction results, and n is the total number of predictions.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(E_i - \hat{E}_i)^2}{n}} \tag{10}$$

The mean absolute error (MAE) is similar to RMSE, but places less emphasis on extreme values. The equation for MAE is provided in Equation (11).

$$MAE = \frac{\sum_{i=1}^{n}\left|E_i - \hat{E}_i\right|}{n} \tag{11}$$

The median error (medE) indicates whether the model has a systematic tendency to over- or under- predict. If the value of medE is 0, it means the prediction method does not have overall bias. The equation for medE is provided in Equation (12).

$$medE = median(E_i - \hat{E}_i) \tag{12}$$

The box plots of prediction errors ($E_i - \hat{E}_i$) with different methods are shown in Figure 15. In methods 1 and 2, the highest error ranges are around 8-9 am and 6-7 pm, which are the going-to-work and going-home periods respectively. It means the occupant number is in high uncertainty during these two periods. In method 3, the error ranges are significantly reduced, especially during going-to-work and going-home periods. It indicates the proposed method reduces the uncertainty and narrows the error range.

The mean errors of different methods are compared in Figure 16. In Methods 1 and 2, the range of mean errors is from -6 to 16, where the max and min values are at 8 am and 6 pm respectively. In Method 3, the range of mean errors is from -2 to 3, where the max and min values are at 2 pm and 8 am respectively. Figure 16 shows the proposed method improves the prediction accuracy significantly. It needs to be noted that the mean errors are very similar between Method 1 and Method 2, which means without pattern identification, the prediction accuracy cannot be improved significantly by only refining time scale. Namely, the patterns

and rules developed by data mining approach make the main contribution to the prediction accuracy.

The performances of each method based on the statistical metrics are shown in Table 5. According to RMSE and MAE metrics, the proposed method improves the prediction accuracy by around 30% compared to method 1 and 2. According to medE metric, methods 1 and 2 have positive systematic biases, but the proposed method has little bias. Therefore, all the metrics indicate the proposed method has better performance than traditional methods.

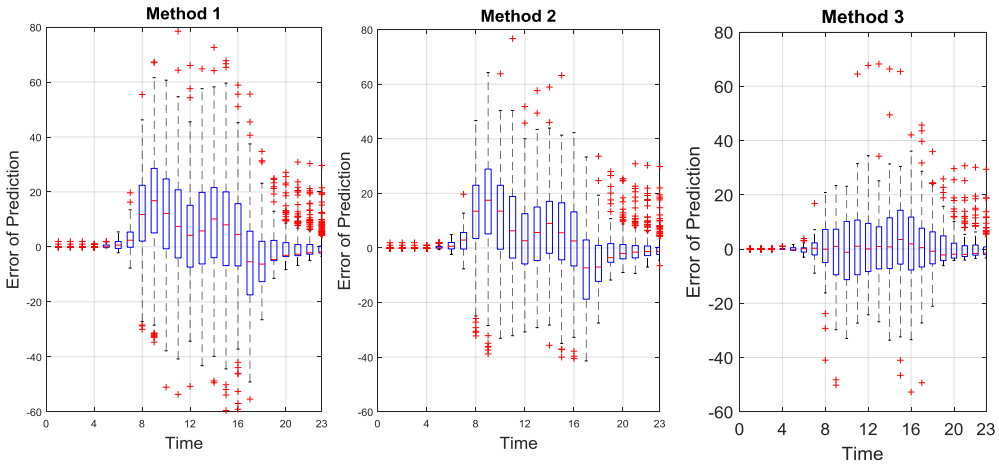


Figure 15 Errors of prediction of the three methods

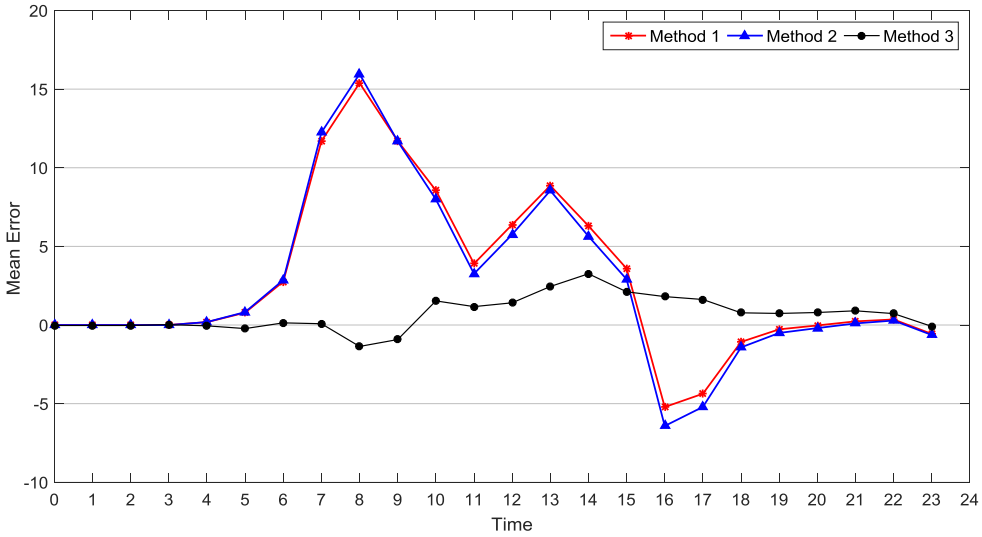


Figure 16 Mean errors of prediction of the three methods

Table 5 Performance of the three methods based on the statistical metrics

	Method 1	Method 2	Method 3
RMSE	68.4	60.4	48.5
MAE	8.5	7.6	5.8
medE	2.4	2.3	-0.07

4 Discussion

There are three main advantages of the proposed data mining based method. First, the underlying patterns and characteristics of data can be discovered by this method. The traditional data analysis methods can only show the statistical characteristics of the entire data set, as shown in Section 4.1. However, in reality, many data sets involve several subsets with various different characteristics. It is just like a bowl of mixed beans. It is difficult to describe the colors, shapes and other characteristics of the mixture. To understand the specific characteristics of the beans, the first step is to differentiate the types of beans by putting them in different bowls. That process is to discover the patterns of data. In this study, four patterns of occupant presence are recognized by cluster analysis. For each pattern, the characteristics can be identified clearly (Table 3).

Secondly, only simple input data is required for the proposed method. The only data input is the accessing records of the building (Table 2). The number of occupants can be calculated from the accessing records, and the attributes in this study (day of week and season) can be transformed from the time in raw data. Currently, this data is available in most commercial buildings for security reason. Data limitation is the main barrier in data mining, so the simple data requirement is a considerable benefit for adoption of the method.

Thirdly, this method can achieve more accurate prediction by relatively simple algorithms. The proposed method uses decision tree to induce rules of occupancy patterns. The decision tree method is straightforward with much lower complexity compared to other learning

methods (e.g., neural network algorithm). And method of weighted mean, which is likewise a most simple method, is used to predict occupancy schedule. The appropriate combination of the simple methods can obtain good performance and is helpful for applying the proposed method to real projects.

The results of this study can facilitate various applications. Energy simulation and prediction is a main direction. Numerous previous studies indicated that occupant presence can significantly impact energy consumption [6-8]. However, most current energy simulation programs use simplified and homogeneous occupant schedules often provided by standards (e.g., ASHRAE 90.1-2004 [40]). The prediction of occupancy schedule in this study can reduce uncertainty and improve accuracy of energy prediction. In addition, the results can facilitate energy efficiency retrofit. The number of occupants is an important factor to calibrate the energy saving after retrofit. The prediction results can help daily operation of buildings. For example, if the occupancy of a day is predicted to be Pattern 1, the start time of light, HVAC and other appliances can be delayed. Conversely, if the occupancy is predicted to be Pattern 2, the equipment should start earlier.

Besides the occupant presence, the proposed method can be employed in other applications. For example, to predict the patterns and prediction of energy use (e.g., electricity, gas and water), occupant behaviors (e.g., opening and closing windows, turning on and off lights) based on historical data. Furthermore, the method can be used in other domains, including the attendance of classes in university, purchasing habits of customers and travel behaviors on subway.

There are several limitations of this study. First is the reliability of the source data. Due to the sensor failure and other reasons, there are some missing data. And there is a small door used occasionally, shown in Figure 6, which cause the entering number and exiting number are not equal sometimes. Although the deviation is lower than 5%, it still impacts the accuracy of

results. In addition, due to data limitation, only one building is conducted as case study and the time span is one year.

5 Conclusions

Most commercial buildings have access control system, which is capable of providing data of accessing records in short time intervals. This data offers new opportunities to understand and predict occupant presence of buildings. However, few previous studies, paid attention to this area.

This study proposes a data mining based approach to learning and predicting the occupancy schedule of buildings. First, four typical patterns of occupant presence are discovered by cluster analysis. Then, the rules of the four patterns are induced by the decision tree method. Thirdly, based on the induced rules, the occupancy schedule is predicted based on the weighted mean of previous data by corresponding probabilities of patterns. Finally, the prediction results are validated by the observed data.

The proposed method in this study can be used in various applications both in building domain (i.e., occupant behavior) and other domains (i.e., marketing, transportation and energy) with available data records. Since the occupant presence is considered an essential factor in building operations as well as energy consumption in buildings, further research is recommended to improve building controls and energy prediction based on predicted occupancy schedule.

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.

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] A. Kashif, X. H. B. Le, J. Dugdale, and S. Ploix, "Agent based Framework to Simulate Inhabitants' Behaviour in Domestic Settings for Energy Management," in *ICAART (2)*, 2011, pp. 190-199.

[3] 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.

[4] THUBERC, "Annual Report on China Building Energy Efficiency," 2007.

[5] J. Laustsen, "Energy efficiency requirements in building codes, energy efficiency policies for new buildings," *International Energy Agency (IEA),* pp. 477-488, 2008.

[6] E. Azar and C. C. Menassa, "Agent-Based Modeling of Occupants and Their Impact on Energy Use in Commercial Buildings," *Journal of Computing in Civil Engineering,* vol. 26, pp. 506-518, 2012.

[7] O. T. Masoso and L. J. Grobler, "The dark side of occupants' behaviour on building energy use," *Energy and Buildings,* vol. 42, pp. 173-177, 2// 2010.

[8] S. D'Oca and T. Hong, "Occupancy schedules learning process through a data mining framework," *Energy and Buildings,* vol. 88, pp. 395-408, 2/1/ 2015.

[9] A. F. Emery and C. J. Kippenhan, "A long term study of residential home heating consumption and the effect of occupant behavior on homes in the Pacific Northwest constructed according to improved thermal standards," *Energy,* vol. 31, pp. 677-693, 4// 2006.

[10] H. Staats, E. van Leeuwen, and A. Wit, "A longitudinal study of informational interventions to save energy in an office building," *Journal of Applied Behavior Analysis,* vol. 33, pp. 101-104, Spring 2000.

[11] J. Yudelson, *Greening existing buildings*: McGraw-Hill New York, 2010.

[12] C. Turner and M. Frankel, "Energy performance of LEED for new construction buildings," U.S. Green Building Council2008.

[13] 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.

[14] P. Hoes, J. Hensen, M. Loomans, B. De Vries, and D. Bourgeois, "User behavior in whole building simulation," *Energy and Buildings,* vol. 41, pp. 295-302, 2009.

[15] S. D'Oca and T. Hong, "A data-mining approach to discover patterns of window opening and closing behavior in offices," *Building and Environment,* vol. 82, pp. 726-739, 12// 2014.

[16] 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.

[17] T. Zhang, P.-O. Siebers, and U. Aickelin, "Modelling electricity consumption in office buildings: An agent based approach," *Energy and Buildings,* vol. 43, pp. 2882-2892, Oct 2011.
[18] 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.
[19] T. Ryan and J. S. Vipperman, "Incorporation of scheduling and adaptive historical data in the Sensor-Utility-Network method for occupancy estimation," *Energy and Buildings,* vol. 61, pp. 88-92, 6// 2013.
[20] D. Wang, C. C. Federspiel, and F. Rubinstein, "Modeling occupancy in single person offices," *Energy and Buildings,* vol. 37, pp. 121-126, 2// 2005.
[21] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*: MIT Press, 2001.
[22] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed.: Morgan Kaufmann, 2011.
[23] W. F. van Raaij and T. M. M. Verhallen, "Patterns of residential energy behavior," *Journal of Economic Psychology,* vol. 4, pp. 85-106, 1983/10/01 1983.
[24] K. Van Den Wymelenberg, "Patterns of occupant interaction with window blinds: A literature review," *Energy and Buildings,* vol. 51, pp. 165-176, 8// 2012.
[25] Z. Yu, B. C. M. Fung, F. Haghighat, H. Yoshino, and E. Morofsky, "A systematic procedure to study the influence of occupant behavior on building energy consumption," *Energy and Buildings,* vol. 43, pp. 1409-1417, 6// 2011.
[26] Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling," *Energy and Buildings,* vol. 42, pp. 1637-1646, 10// 2010.
[27] T. H. Davenport and J. Kim, *Keeping Up with the Quants: Your Guide to Understanding and Using Analytics*: Harvard Business Review Press, 2013.
[28] R. Kohavi and F. Provost, "Glossary of terms," *Machine Learning,* vol. 30, pp. 271-274, 1998.
[29] C. M. Bishop, *Pattern recognition and machine learning*: springer, 2006.
[30] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," 1995.
[31] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "An Introduction to Classification and Clustering," *Cluster Analysis, 5th Edition,* pp. 1-13, 2011.
[32] J. Abello, P. M. Pardalos, and M. G. Resende, *Handbook of massive data sets* vol. 4: Springer, 2013.
[33] V. Estivill-Castro, "Why so many clustering algorithms: a position paper," *ACM SIGKDD explorations newsletter,* vol. 4, pp. 65-75, 2002.
[34] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Kdd*, 1996, pp. 226-231.
[35] O. Maimon and L. Rokach, "Data mining with decision trees: theory and applications," ed: USA: World Scientific Publishing, 2008.
[36] J. R. Quinlan, "Induction of decision trees," *Machine learning,* vol. 1, pp. 81-106, 1986.
[37] J. R. Quinlan, "C4. 5: Programming for machine learning," *Morgan Kauffmann,* 1993.
[38] T. M. Mitchell, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill,* vol. 45, 1997.
[39] EEBHUB. (Dec 17). *Energy Efficient Buildings Hub. http://www.buildsci.us/eeb-hub.html*. Available: http://www.buildsci.us/eeb-hub.html
[40] ASHRAE, *Energy Standard for Buildings except Low-RiseResidential Buildings*, 90.1-2004.