Building and Environment 102 (2016) 179–192

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv

# Occupancy data analytics and prediction: A case study

Xin Liang [a, b], Tianzhen Hong [b, *], Geoffrey Qiping Shen [a]

[a] Department of Building and Real Estate, Hong Kong Polytechnic University, Hong Kong, China
[b] Building Technology and Urban Systems Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

ARTICLE INFO

*Article history:*
Received 6 January 2016
Received in revised form
12 March 2016
Accepted 25 March 2016
Available online 28 March 2016

*Keywords:*
Occupancy prediction
Occupant presence
Data mining
Machine learning

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.



## 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

* Corresponding author.
*E-mail addresses:* xin.c.liang@connect.polyu.hk (X. Liang), thong@lbl.gov (T. Hong).

http://dx.doi.org/10.1016/j.buildenv.2016.03.027
0360-1323/

[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 Fig. 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