Title

Smart thermostat data-driven U.S. residential occupancy schedules and development of a U.S. residential occupancy schedule simulator

Authors
Jung, Wooyoung
Wang, Zhe
Hong, Tianzhen
et al.

Publication Date
2023-09-01

DOI
10.1016/j.buildenv.2023.110628

SMART THERMOSTAT DATA-DRIVEN U.S. RESIDENTIAL OCCUPANCY SCHEDULES AND DEVELOPMENT OF A U.S. RESIDENTIAL OCCUPANCY SCHEDULE SIMULATOR

Wooyoung Jung[1,*], Zhe Wang[2, 3], Tianzhen Hong[4], and Farrokh Jazizadeh[5]

[1]The University of Arizona

[2]The Hong Kong University of Science and Technology

[3]HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute

[4]Lawrence Berkeley National Laboratory

[5]Virginia Tech

[*]Corresponding author

Abstract

Occupancy schedule is one of the key inputs in Building Energy Modeling (BEM) to reflect the interaction between buildings and occupants. Over the past decades, standardized occupancy schedules, developed mainly by engineering rule-of-thumb, have been widely used in BEM due to its simplicity and lack of real measured occupancy data. However, the BEM community has recognized their association with uncertainty and reliability in simulation results from BEM. This study introduces representative occupancy schedules in the U.S. residential buildings, derived from a large smart thermostat dataset and time-series K-means clustering, and an open-source tool to generate a stochastic residential occupancy schedule. Over 90,000 residential occupancy schedules were estimated from the ecobee Donate Your Data dataset. Then, the representative occupancy schedules were identified through clustering. This study further investigated the impacts of three parameters (day, house type, and state) on residential occupancy schedules. Then, a tool, the Residential Occupancy Schedule Simulator (ROSS), is developed using the representative occupancy schedules derived in this study. Details of this tool are presented in this paper. The derived representative occupancy schedules and the ROSS tool can help improve the energy modeling of residential buildings.

Keywords: Smart thermostat, Occupancy schedule, Residential building, Building energy modeling.

1 Introduction

A trend of adopting smart thermostats in residential buildings has been one of the movements of enhancing Human-Building Interaction (HBI), initiated by users. Consumers are attracted by this smart home device since they can have better control of Heating, Ventilation, and Air-Conditioning (HVAC) systems, improve their thermal comfort, and ultimately reduce their energy bills [1].

From the Building Energy Modeling (BEM) perspective, this trend provides great opportunities to analyze occupant behaviors or perceptions since some vendors store their smart thermostat data in their cloud server upon users' agreements. For example, vendors could keep track of the indoor cooling and heating temperature setpoints (hereinafter setpoints), understanding preferred air temperatures by the majority of the users. In other words, they could analyze the overall trends of setpoints per region or climate [2].

Even though occupancy could be considered the most intuitive feature of occupants' dynamics in buildings, compared to preferred environments or the location of occupants, its uniqueness in each space and building challenges the validity of BEM [3]. The American Society of Heating, Refrigerating, and Air-conditioning Engineers (ASHRAE) provides *standardized* occupancy schedules for several building types since 1989 [4], and they have been primarily applied in BEM [5]. However, several studies have demonstrated that actual occupancy schedules were distinct from such schedules, and such gaps contributed to unreliable building performance analyses [3, 6, 7].

In 2015, ecobee, one of the leading vendors of smart thermostats, launched a program called Donate Your Data (DYD) [8]. Through this DYD program, users can share their smart thermostat data for research and development. Then, researchers under the agreement gain access to their anonymized data. The continuous growth of user participation has contributed to the DYD dataset, containing over 104,000 ecobee smart thermostats until 2019. Each smart thermostat data – saved as a csv file – contains operational data such as setpoints, typical occupancy schedules set by users, motion, outdoor temperatures, heating or cooling mode, and others (more detailed information is in Section 3). In addition, the metadata contains floor area, house type, location (i.e., province/state and city), the maximum number of residents, etc. This DYD dataset provides an unprecedented opportunity for large-scale occupancy pattern analyses in residential buildings. Although several studies used the DYD dataset in recent years [2, 9-11], their focuses were unrelated to occupancy schedules. Some research efforts reported data-driven occupancy schedules [3, 6, 7], but they focused on either office/commercial buildings or university buildings and the sample size (i.e., the number of buildings) was relatively small (≤100). Recently, Mitra et al. [12] used the American Time Use Survey (ATUS) data to develop a typical residential occupancy schedule using over 190,000 interview data. They revealed a large gap between the standardized occupancy schedule and their results and identified the roles of age and the number of residents in shaping occupancy schedules. This study aimed to analyze the roles of location (i.e., state), house type, and day of the week, which will further advance the knowledge of data-driven occupancy schedules in residential buildings. In addition, gathering interview data is a cost-intensive and slow process. A data pool collected via smart thermostats is quickly expanded given its increasing market adoption, hence it is worth being analyzed from the BEM perspective.

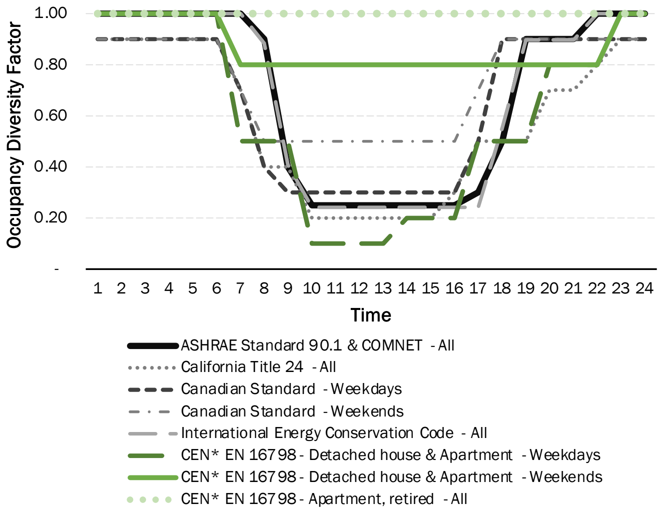
Therefore, this study contains two research efforts: (1) identifying data-driven representative occupancy schedules in residential buildings utilizing a large-scale smart thermostat dataset and (2) providing building energy modelers, engineers, or researchers with a tool to generate more realistic and representative occupancy schedules, for their residential building performance analyses. Regarding the first objective, we have explored whether any feature in the metadata (e.g., location, day, and house type) played a role in residential occupancy schedules. Hence, it could be the features that users should take into consideration. For example, the occupancy schedule for the multifamily building type, recommended by the ASHRAE standard 90.1 [5], has the same occupancy diversity factors (i.e., the probability of occupancy) throughout the week. We aimed to evaluate this generalization in our analyses. For the second objective, we have developed an open-source program, written in Python, which stochastically generates a residential occupancy schedule. Occupancy patterns in a residential unit are dynamic and complicate the analyses from BEM, hence we aimed to offer a tool that reflects reality. The stochastic schedule references representative occupancy patterns identified in this study, which can be generated by the tool. Therefore, building scientists, engineers, and modelers can make the most out of this study. It is important to note that this study focused on the data collected from the U.S. homes, which contains time-series data from 91,000 smart thermostats and was the majority at the time of data sharing from the vendor.

This paper is structured as follows. In Section 2, we introduced the occupancy schedules in standards and manuals. After that, we explained the occupancy sensing and modeling approaches emerged in recent years, paving the way for data-driven occupancy schedule analyses. This literature review shed light on the unique chance that the ecobee DYD dataset offers, especially in the context of developing residential occupancy schedules using data-driven approaches. Section 3 shows the approach and the results of identifying representative occupancy schedules. Section 4 presents the details of the Residential Occupancy Schedule Simulator (ROSS). In Section 5, the authors share their thoughts on a large smart thermostat dataset from the perspective of BEM. The last section shares the contributions of this study.

2 Literature review

2.1 Standardized occupancy schedules for residential buildings

As noted, the ASHRAE standard 90.1 has introduced reference occupancy schedules since 1989 [4] and other standards or manuals share typical occupancy schedules for several building types. The residential building type, often referred as the multifamily building type, has been commonly included, and Figure 1 and Figure 2 show the building-specific and space-specific occupancy schedules in multiple standards and manuals.



*European Committee for Standardization

Figure 1. The occupancy schedule for the residential building type suggested by standards and manuals.

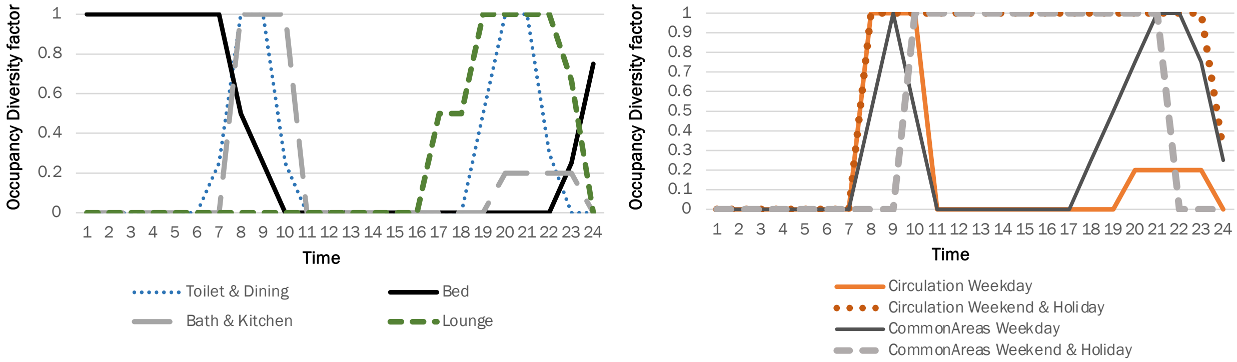


Figure 2. The space-level occupancy schedules for the multifamily building type suggested by the U.K. National Calculation Method (NCM).

The ASHRAE standard 90.1 and COMNET provide the same occupancy schedule and International Energy Conservation Code had a very similar schedule with marginal gaps – the diversity factor at 17:00 was the only difference. In contrast, the California's energy code proposes a distinct shape and different maximum and minimum values, compared to the others published in the U.S. Occupancy schedules by the Canadian standard have variations by day. The weekend schedule has the minimum value of 0.5 (higher than the minimum value of 0.3 in the weekday schedule), implying that the residential building type is more likely to be occupied on the weekend. Occupancy schedules in the European standard [13] take three factors into account – the day of the week, house type and occupation of the residents. Even though the house type does not affect the occupancy diversity factor, two house types are separately presented in the standard. On the other hand, the day of the week and retirement status plays a significant role. Lastly, the Simplified Building Energy Model (SBEM) in U.K. was the only resource authors found that introduces space-level occupancy schedules. The occupancy schedule for the bed space type follows a similar trend in other recourses – starting to have low occupancy diversity factors from 5:00-7:00 am and returning to the maximum value from 3:00-11:00 pm. However, it has zero value from 10:00 am to 10:00 pm, which is substantially different from the others.

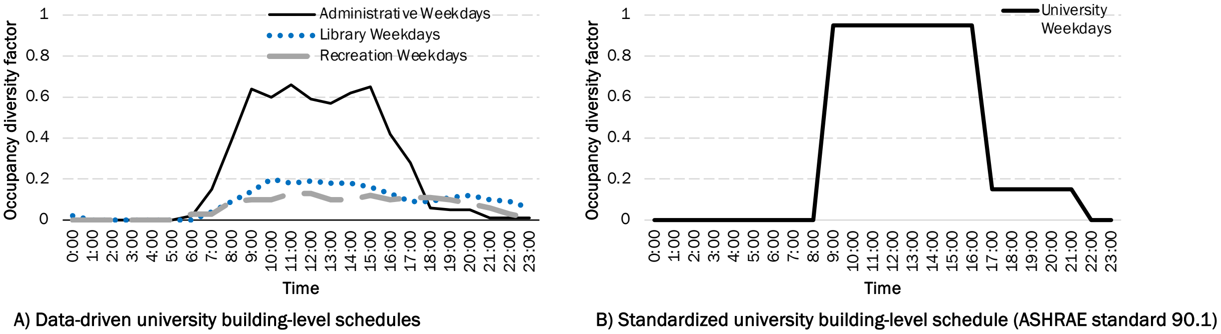
Understanding the gaps in these occupancy schedules is limited because none of the standards and manuals shares the raw data or the reasons for having such occupancy diversity factors. Deru et al. [4] shared the history of updating the occupancy schedules for multiple building types in the User's Manual for 90.1-2004 through a public review process. This means that these schedules were primarily decided by the engineering judgments and the agreements, rather than by data. Given the difficulty of obtaining large-scale occupancy data in each building type in the past, this process was reasonably acceptable.

2.2 Development of Occupancy Sensing and Modeling

In the last decade, one of the major interests in the field of HBI has been the development of accurate occupancy sensing techniques and occupancy schedules generation using data-driven approaches. The former allows building systems to instantaneously respond to occupancy dynamics and the latter opens up the potentials for predictive controls like pre-conditioning the buildings in the prediction of occupants' arrival times [14].

As synthesized in our review study [15], substantial advancements in occupancy sensing have been made in the last decade. A variety of sensor types and their compositions have been employed to investigate their potential for occupancy detection. Specifically, studies deployed the sensors are triggered by the variations of infrared radiation (i.e., motion) [16], ambient conditions [17], door states (i.e., open or closed) [17], WiFi uses [18], and electrical loads [19]. In the case of multiple sensing nodes, the average accuracy was almost 90% [15]. Examples were motion sensor networks [20] and combinations of motion and ambient conditions [17].

These technological advancements paved the way for developing data-driven occupancy schedules. Davis and Nutter [3] collected occupancy data from 11 university buildings for four to seven months through security cameras, doorway counting sensors, classroom scheduling data, and personal observations. Then, this study demonstrated that each building had distinct occupancy schedules. Similarly, Duarte et al. [6] collected the occupancy data in a large multi-tenant office building utilizing a total of 629 passive infrared (PIR) sensors over two years and presented occupancy schedules of several space types by day, showing that occupancy schedules could vary by day and space types.



A) Data-driven university building-level schedules

B) Standardized university building-level schedule (ASHRAE standard 90.1)

Figure 3. data-driven vs. standardized university building-level schedules from [3] and [21].

One important takeaway from these studies is that the data-driven occupancy schedules were quite different from the occupancy schedules used in standards and manuals. As in Figure 3, the data-driven occupancy schedules have lower peak values, and the shapes are distinct from the standardized schedules from ASHRAE standard 90.1. These gaps justify the necessity of reporting data-driven occupancy schedules. However, a common limitation has been the scale of measurements. Many investigated occupancy sensing potentials in several offices [17, 20] or multiple houses (≤100) [19, 22, 23]. Moreover, the data was collected for at most 8 months. In addition, most of the studies explored occupancy schedules in commercial or university buildings [3, 24-26], which might be because that such building types are more accessible by researchers. Also, the occupancy detection in residential buildings could raise privacy concerns. Although some studies have collected occupancy data in residential buildings [27, 28], the scale of the experiments was insufficient to demonstrate typical occupancy schedules, and the focus of the study was to evaluate the performance of a novel sensing system [27].

The ecobee DYD dataset, however, includes an unprecedented number of smart thermostats collected over several years. Also, the dataset is geographically well-distributed. Hence, the ecobee DYD dataset offers a great opportunity to analyze occupant behavior in residential buildings. With this dataset, Huchuk et al. [2] have analyzed the overall variations of setpoints depending on the season, region, and cost, and Huchuk et al. [9] developed several machine-learning models to infer occupancy and compared their performances. Also, Ueno and Meier [11] developed a method to create multiple representative heating and cooling setpoint schedules. However, none of these studies explored the potential of identifying representative occupancy schedules in residential buildings. Lastly, analyses of occupancy schedules via smart thermostats will contribute to energy simulations at various scales [29, 30].

3 Representative occupancy schedules derived from the DYD dataset

This section explains the first part of the research: identifying the representative occupancy schedules via a data-driven approach. The first subsection explains the methodology and the second subsection presents the results. It is worth mentioning that the occupancy schedules address the occupancy *status* not occupant *count* in this study.

3.1 Methodology

This study processed the ecobee DYD dataset to generate occupancy schedules from thermostat operational data. In so doing, this study came up with a couple of approaches that estimated occupancy in the ecobee DYD dataset. Then, a K-means clustering algorithm is applied to the schedules to identify clusters. More details are presented in the following subsections. As a note, this study utilized Python and its modules (e.g., Pandas, scikit-learn) for data processing.

3.1.1 ecobee DYD dataset description

The ecobee DYD dataset is composed of (1) the metadata file and (2) smart thermostat operation data files. Both were in the csv format. The metadata file contained the information organized in Table 1. In short, it had contextual information of each thermostat such as user, location (city, not mailing address), house, HVAC systems. The *Filename* field could be used to identify the thermostat operation data files. The ecobee DYD smart thermostat dataset is dominated by the U.S. residential buildings, followed by Canada and small portions of other countries.

Table 1. Metadata fields and their descriptions [31].

#	Fields	Description
1	Identifier	Unique value for each thermostat
2	Model	Generation of ecobee thermostat
3	User ID	Unique value given to a user account
4	Country	Country that the thermostat is in, inputted by the user
5	ProvinceState	Province or state that the thermostat is in, inputted by the user
6	City	City that this thermostat is in, inputted by the user
7	Floor Area	Floor area, inputted by the user
8	Style	Type of home, selected by the user (e.g., detached, townhouse, condo, etc.)
9	Number of floors	Number of floors, inputted by the user
10	Age of homes	Age of homes in years, inputted by the user
11	Number of occupants	Number of occupants, inputted by the user
12	installedCoolStages	Number of stages of the cooling system has (identified by wiring)
13	installedHeatStages	Number of stages of the heating system has (identified by wiring)
14	allowCompWithAux	User-configurable value to allow compressor and auxiliary stages to run simultaneously
15	Has Electric	Indicating whether an electric-based heat source is installed
16	Has a Heat Pump	Indicating whether a heat pump is installed

17	Auxiliary Heat Fuel Type	Fuel type for the auxiliary heat system
18	Number of Remote Sensors	Number of remote sensors connected to the thermostat
19	Filename	Name of the csv files associated with this thermostat
20	First Connected	First connection of the thermostat to the ecobee servers
21	eco+ enrolled	Current status in enrollment in eco+
22	eco+ slider level	Current eco+ slider level

Each thermostat operation data file is named after its unique identifier and contained the recorded data, described in Table 2, at a sampling rate of five minutes. When users have multiple remote sensors, each sensor reported temperature and motion data. The users joined the DYD program at different times, hence the amount of data collected in each thermostat varied. This study included all data points available in the dataset to maximize the representativeness in the results. It is worth mentioning that this study leveraged this five-minute interval for analyses at high temporal resolutions.

Table 2. Features in the smart thermostat data and their descriptions

Feature	Description
DateTime	Date and time at a 5-min interval
HVAC mode	The HVAC modes: heating, cooling, or automatic (i.e., heating or cooling is assigned automatically).
Schedule	Users are allowed to assign typical schedules that they are in the house (Home), are away from the house (Away), and sleep (Sleep) and different setpoints can be assigned and they get activated according to the schedule.
Event	A feature that users can employ to override the operational setup assigned in the schedule. For example, the heating setpoint is 72°F in the schedule, but a user can assign 74°F by this feature. The following are the automatic events activated by motion: *Smart home* is when the schedule has been set to Away and upon detection of motion in the space, the setpoints from the Home schedule are assigned. *Smart away* is when the Home schedule has been set and the motion sensors do not detect any motion for two hours.
Heating/Cooling temperature setpoint	Setpoints assigned by users. As noted, they could be from the schedule or event.
Humidity	Indoor humidity measured by the smart thermostat
Motion	Motion data sensed by the Passive InfraRed (PIR) sensors embedded in a smart thermostat and remote sensors. This feature is individually reported by each device.
Outdoor temperature/humidity	Outdoor temperature and humidity from the nearest weather station
Air temperature	Indoor temperature data sensed by a smart thermostat and remote sensors. This feature is individually reported by each device.
Fan	Runtime (seconds) for fan
AuxHeat	Runtime (seconds) for any heat source other than a heat pump
CompCool	Runtime (seconds) for any cooling
CompHeat	Runtime (seconds) for heat-pumps used in heating

It is worth noting that the DYD dataset this study utilized had data until August 2019, when the coronavirus disease 2019 was not dispersed in the US. Another note is Daylight Saving Time (DST) is taken into account automatically in the DateTime column. Hence, the authors did not take any additional efforts to process DST.

3.1.2 Occupancy schedules

Several thermostat features, listed in Table 2, could be associated with occupancy: Motion, Event, and Schedule. Specifically, the ecobee system possesses at least one motion sensor embedded in the thermostat and is possible to expand by connecting remote sensors (named SmartSensors by ecobee). Also, any user interactions with the thermostat such as assigning a new event reflects the presence of the user(s). Hence, this study processed such features as follows.

1. Motion sensors are triggered by movements in the indoor environment, so the *motion* feature was given the highest priority. This also means that the motion-induced events such as *Smart Home* or *Smart Away* (details in Table 2) were prioritized. Next, we leveraged user interaction with the thermostat. For example, every time a new event was assigned, it was considered occupied. Lastly, any setpoint adjustments also indicate the presence of residents in the house.
2. Then, the *schedule* feature was employed to address the challenging scenarios in that the motion data were insufficient to detect occupancy. For example, occupancy was not detected by the motion sensor when residents were sleeping. The solution was to rely on the *Sleep* schedule that indicated the typical time when the residents went to bed. Therefore, the first assumption made in this study was that residents always stayed at home once the *Sleep* schedule was activated. On occasions, users did not assign the *Sleep* schedule in their thermostat settings, and this study proceeded them without manipulating the raw data (a possible explanation of this setup by the users is stated in Section 5).
3. Another useful information from the schedule feature was the *Home* schedule. The PIR sensors are known for having false negative cases (reporting vacancy while the house is occupied), since they are incapable of recognizing stationary humans. Also, another possibility that PIR sensors could not capture motion was when the residents were out of the sensor range(s). In other words, the first three tiers might not fully present the true occupancy. Hence, the second assumption was formed: the resident(s) was always in the house during the *Home* schedule.

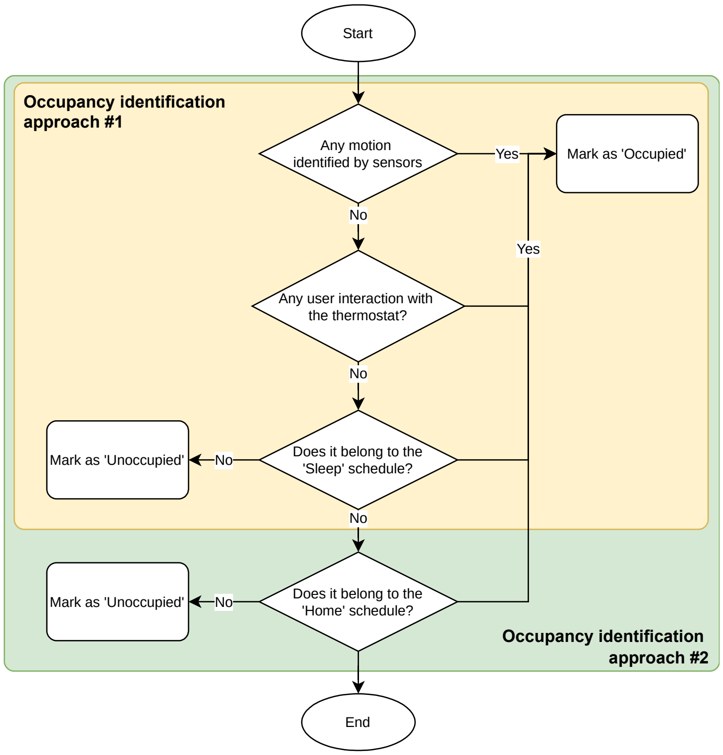
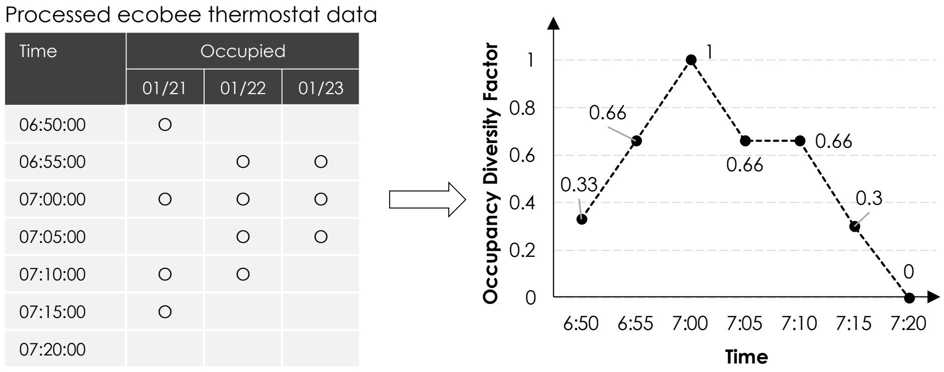

Start
Occupancy identification approach #1
Any motion identified by sensors
Yes
Mark as 'Occupied'
No
Yes
Any user interaction with the thermostat?
No
Mark as 'Unoccupied'
No
Does it belong to the 'Sleep' schedule?
No
Mark as 'Unoccupied'
No
Does it belong to the 'Home' schedule?
Occupancy identification approach #2
End

Figure 4. Flowchart of determining the occupancy state in each timestep.

Figure 4 presents the four steps that we proposed to identify the occupied timesteps from the raw dataset.

This study identified the occupancy status in each timeslot with these two assumptions. The first occupancy identification approach employed the first assumption and the second approach utilized two assumptions. Then, each smart thermostat data was processed with the two approaches. An occupancy diversity factor in each timestep was obtained by averaging the occupancy states from different days. As shown in Figure 5, as an example, when smart thermostat data has three days of data and the house was occupied in three days at 07:00 am, the occupancy diversity factor becomes one.


Processed ecobee thermostat data

Time	Occupied		
	01/21	01/22	01/23
06:50:00	O		
06:55:00		O	O
07:00:00	O	O	O
07:05:00		O	O
07:10:00	O	O	
07:15:00	O		
07:20:00			

Figure 5. Schematic of generating an occupancy schedule from a single ecobee smart thermostat data file

3.1.3 Data selection & grouping

A total of 104,693 ecobee thermostats data were shared with the authors. The majority of the data (91,747 ecobee thermostats) were collected in the U.S. (a total of 48 countries included), where this study focused on. To address the first objective, this study grouped the DYD dataset based on (1) day of the week, (2) house type, and (3) location (i.e., state). Even though the occupancy schedules in the standards or manuals often separate the weekday schedules from the weekend schedules (e.g., [21]), the data-driven schedules indicated that each day could hold a different schedule (e.g., [3]). Given that this study used the latter, each day of the week from Monday to Sunday was taken into consideration. Also, the DYD dataset had the following number of ecobee thermostats per each house type presented in parentheses: Apartment (3,060), condominium (3,252), detached (49,593), loft (350), multiplex (1,325), rowhouse (3,214), semi-detached (965), and townhouse (6,231). Once the users inputted their house type as 'other' (11,621), those thermostats were excluded from the analyses that gauged the impact of *house type* on occupancy schedules due to ambiguity.

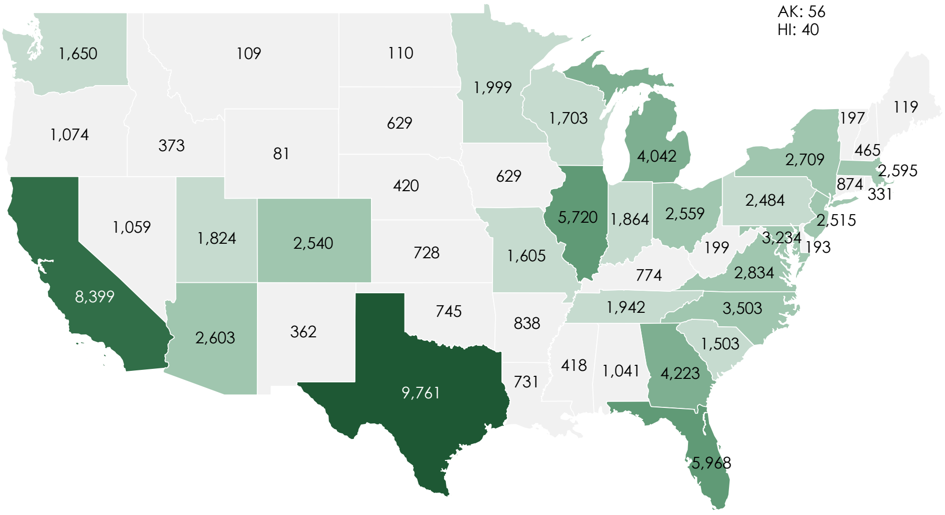


Figure 6. Number of ecobee thermostats for each state in the DYD dataset

Lastly, in order to assess whether the residents' location played a role in determining the occupancy patterns, the DYD dataset was grouped by state. As shown in Figure 6, four states (California, Florida, Illinois, and Texas) contained more than 5,000 thermostats so the DYD dataset in these four states was used to see the role of the location in deciding occupancy schedules. Considering the ASHRAE climate zones, these states were well distributed, covering the marine, hot-dry, hot-humid, mixed-dry, mixed-humid, and cold zones. In other words, this study also covers climate, as a hidden factor, under the umbrella of location.

After these classification processes, the daily occupancy schedule in each ecobee thermostat was generated following the steps in Section 0.

3.1.4 Clustering

After developing occupancy schedules, this study used the time-series K-means clustering method with the Euclidean distance [32] to identify the representative occupancy schedules. The K-means clustering is an unsupervised learning algorithm that aims to partition data points into K clusters based on their similarity. The similarity is calculated through a number of metrics and the Euclidean distance is the most common method for similarity measure in time-series clustering [32].A well-known limitation of the K-means clustering method is that users need to determine the number of clusters, K, as a hyper-parameter of the clustering algorithm. In our approach, the silhouette score utilizing the Euclidean distance [33] was used to decide the optimal number of clusters. The silhouette score quantifies how well data points fit their clusters by calculating the mean intra-cluster distance and the mean nearest-cluster distance: Equation (1), (2), and (3) show how the silhouette score is calculated.

$$s(i) = \frac{b(i) - a(i)}{\max\{\mathrm{a(i)}, \mathrm{b(i)}\}} \tag{1}$$

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j) \tag{2}$$

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j) \tag{3}$$

where data point i in the cluster C_I ($i \in C_I$), $|C_I|$ the number of points belonging to cluster i, $d(i,j)$ is the distance between datapoint i and j, C_J is the nearest cluster to datapoint i, and $s(i)$ is the silhouette score of data point i and the mean of all data points is the silhouette score.

Hence, to identify the optimum number of clusters within each group, we performed the K-means clustering by setting the number of clusters ranging from two to ten (a total of nine cases) and the silhouette score in each case was calculated. Then, the case with the maximum silhouette score was selected as the result in this study. Figure 7 shows this process in a flowchart.

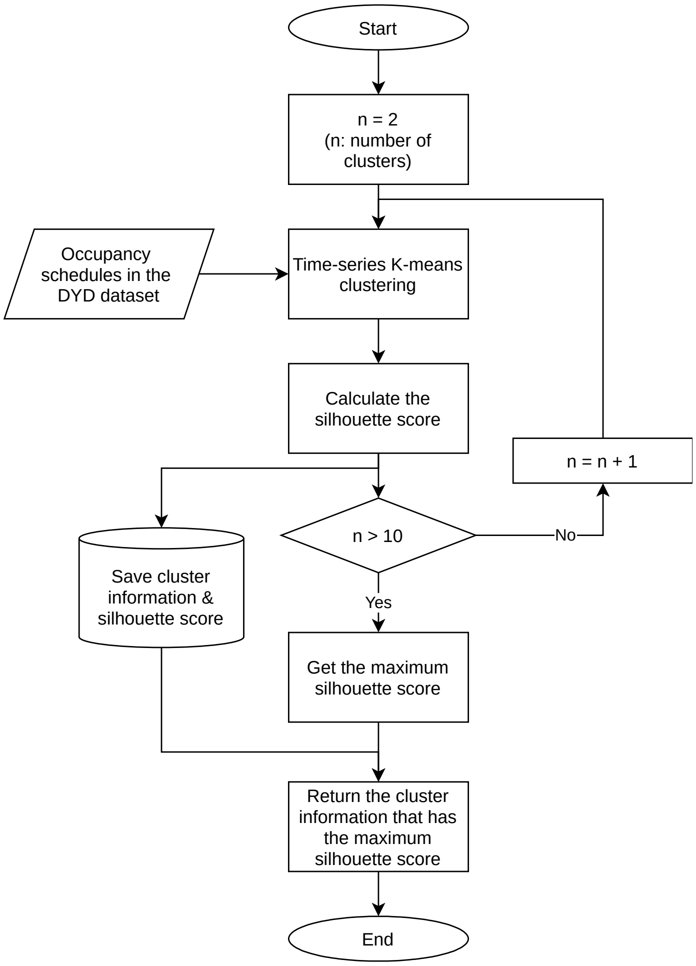

Start
n = 2
(n: number of clusters)
Occupancy schedules in the DYD dataset
Time-series K-means clustering
Calculate the silhouette score
n = n + 1
Save cluster information & silhouette score
n > 10
No
Yes
Get the maximum silhouette score
Return the cluster information that has the maximum silhouette score
End

Figure 7. Flowchart of the time-series K-means clustering method in this study

3.2 Results

Since this study attempted two occupancy identification approaches, the following subsection discusses the differences in results from both approaches. The second subsection shows how the clusters varied via different K values with the entire ecobee DYD dataset, demonstrating the variation of the clusters. The third subsection presents which parameter influences the occupancy schedules and the fourth subsection compares the representative occupancy schedules identified in this study against the occupancy schedules recommended by standards and manuals in the U.S. As noted, each user participated in the DYD program autonomously, so each thermostat possessed different numbers of data points included in the analysis. The average number of days included in each thermostat was 257 days per year.

3.2.1 Differences in two occupancy identification approaches

Before presenting the representative occupancy schedules found in this study, it is worth examining the difference caused by the two occupancy identification approaches. Figure 8 shows how two occupancy schedules were created from the same data as an example. The occupancy diversity factors from 10:00 pm to 6:00 am were similar in both cases as the ‘Sleep’ schedule was considered occupied. However, the occupancy schedules from 2:00 pm to 10:00 pm were different. The high occupancy diversity factor at 2:00 pm from the first approach indicates that motion was recognized frequently. However, it does not necessarily mean that the resident(s) stayed at home after 2:00 pm because of low occupancy diversity factors until 10:00 pm. It means that the resident(s) was sensed by one of the motion sensors and then disappeared. Another clue of what might have happened is having a high probability of motion at 8:00 am.

In other words, the resident(s) had recognizable motions at 8:00 am and 2:00 pm in front of the motion sensor regularly (e.g., commute). However, everything else is unclear. This ambiguity came from the absence of the ground truth data and limitations of PIR sensors such as the line-of-sight drawback or the incapability of recognizing stationary humans.

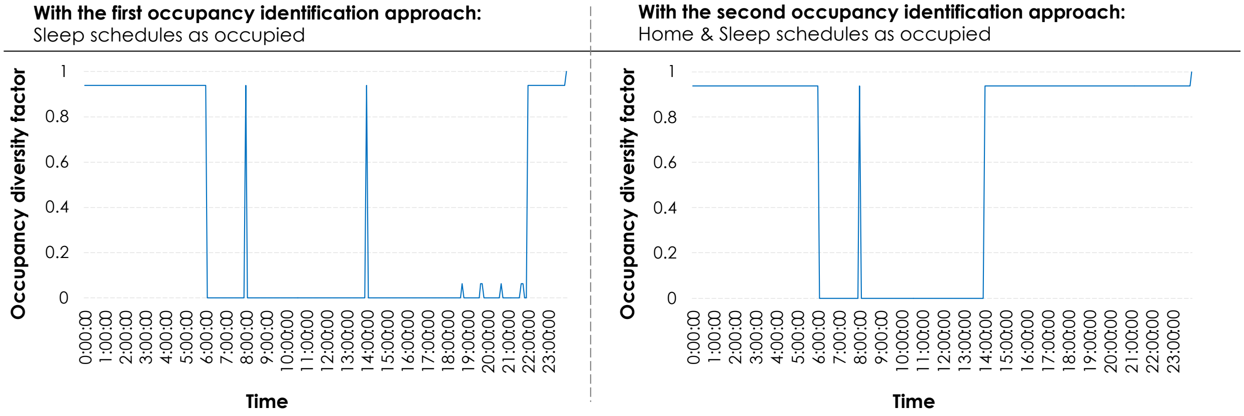
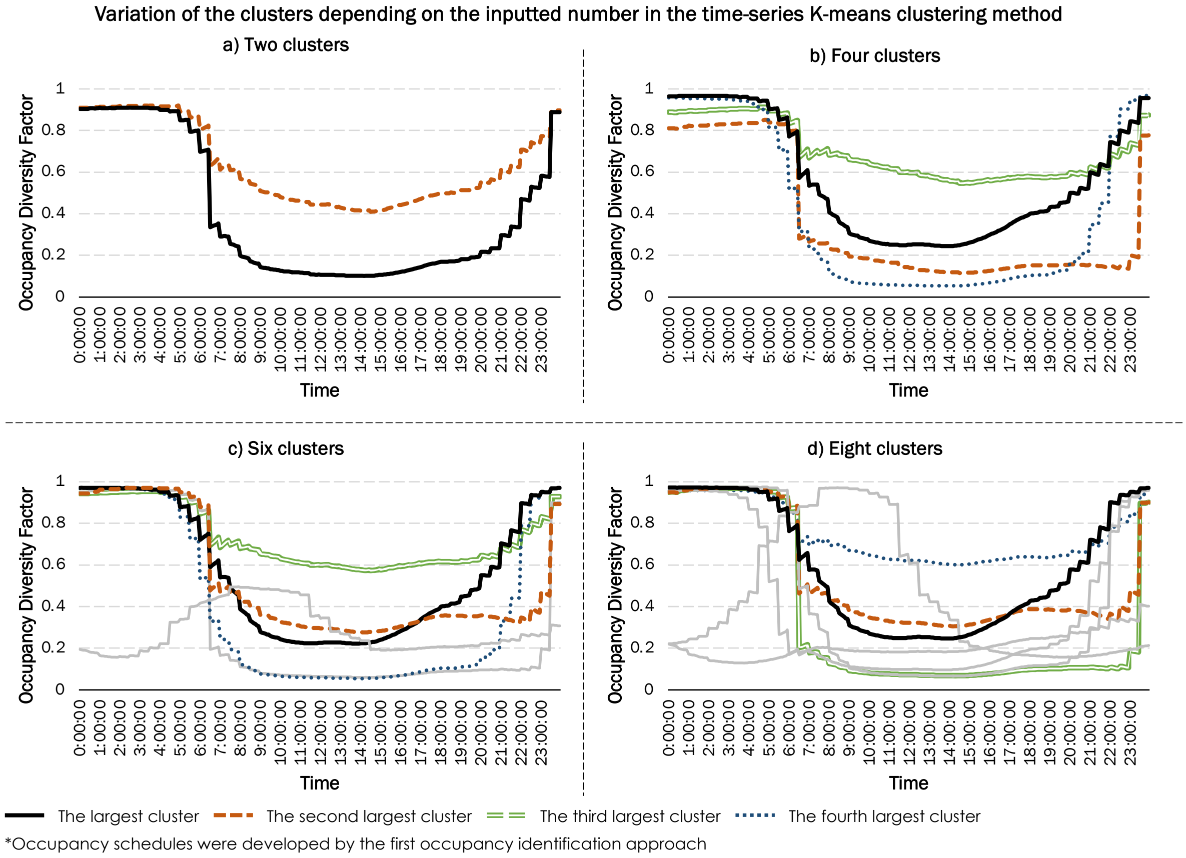


Figure 8. Occupancy schedules derived from an ecobee data file under two occupancy identification approaches.

This study has embodied such ambiguities in the analyses since revealing the limitations of the DYD dataset could stimulate discussions about how the smart thermostat dataset can be improved. The authors' thoughts are shared in the Discussion section.

3.2.2 Variations of representative occupancy schedules

This subsection presents the variations of clusters (i.e., representative occupancy schedules) when different numbers of clusters were fed into the K-means algorithm once all schedule data, including all days in the week, in the US were provided. These results explain the reasons for choosing the optimal number of clusters by the silhouette scores with more details.


Variation of the clusters depending on the inputted number in the time-series K-means clustering method

a) Two clusters

b) Four clusters

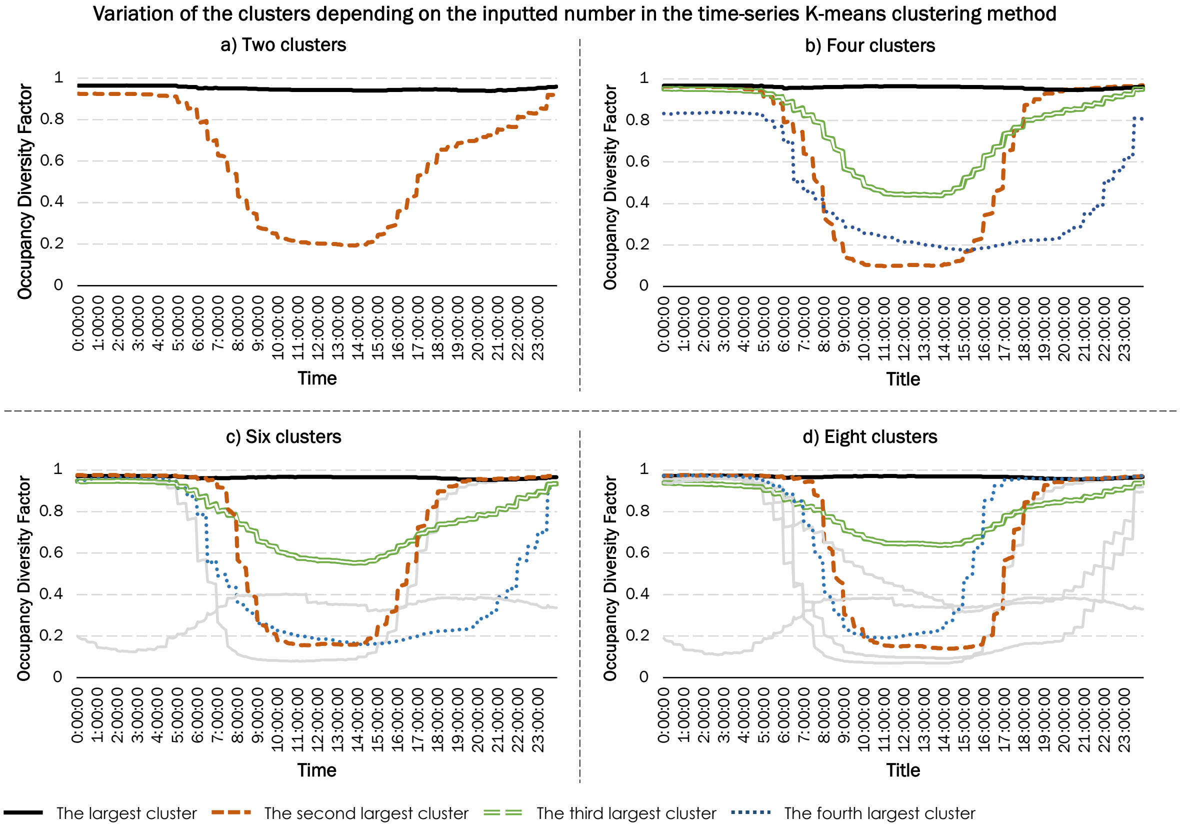
c) Six clusters

d) Eight clusters

*Occupancy schedules were developed by the first occupancy identification approach

Figure 9. Clusters identified through the time-series K-means clustering method using the first occupancy identification approach.

The largest cluster, with the black solid line, in the first approach, changed to a large extent when the number of clusters increased from two to four (from Figure 9 a) to b)). The occupancy diversity factors consisted of higher values from 06:00 am to 11:00 pm, looking like a mixture of two clusters in Figure 9 (a) and low occupancy diversity factors were clustered as the fourth cluster (the dotted blue line in Figure 9 (b)). Similarly, the second largest cluster among four had higher values when six or eight clusters were selected, meaning that large clusters were broken into smaller ones. However, four dominant clusters were similarly shaped although their order varied in some cases. There was an irregular pattern, having high values from 4:00 am to 2:00 pm in Figure 9 (c) and (d).



*Occupancy schedules were developed by the second occupancy identification approach

Figure 10. Clusters identified through the time-series K-means clustering method using the second occupancy identification approach.

The largest clusters from the second approach (considered home and sleep states as occupied) consistently had a horizontal line with high diversity factor values (close to 1) regardless of the number of clusters. The minimum value was 0.9389 at 08:45 pm in Figure 10 (a) and 0.9596 at 08:40 pm in Figure 10 (d). Also, the second and third largest clusters were similarly shaped when a different number of clusters was fed into the algorithm. The fourth largest cluster had a different shape once eight clusters were formed, but, considering its small population and the differences with other clusters, it does not impact the overall outcome.

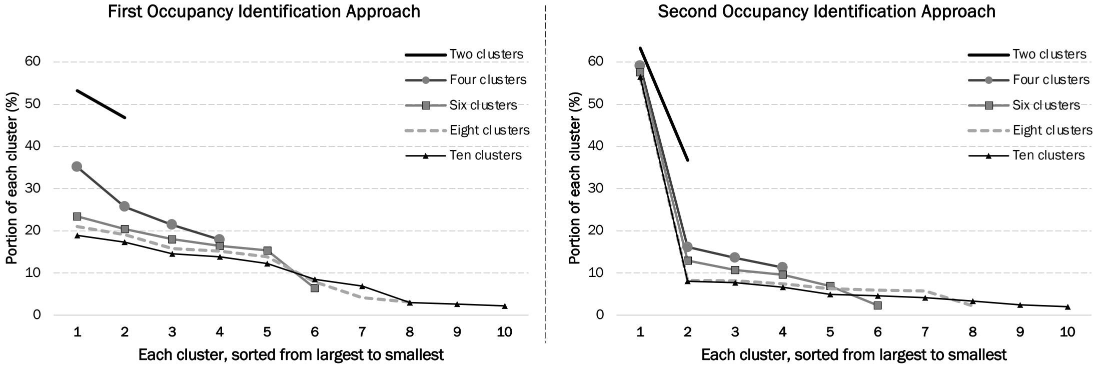


Figure 11. Population of each cluster depending on the number of clusters assigned in the K-means clustering method.

The aforesaid trends were also found once the population of each cluster is taken into consideration. In the first approach, there was no single cluster that dominantly held the schedules. As shown in Figure 9, the population of the largest cluster kept decreasing once higher initial K values were fed into the algorithm: from 53.2% to 18.9%. In contrast, the largest cluster for the second approach maintained the population >56.0% regardless of the K values inputted in the time-series K-means clustering method. In other words, a large portion of schedule data had similar patterns and could be clustered together.

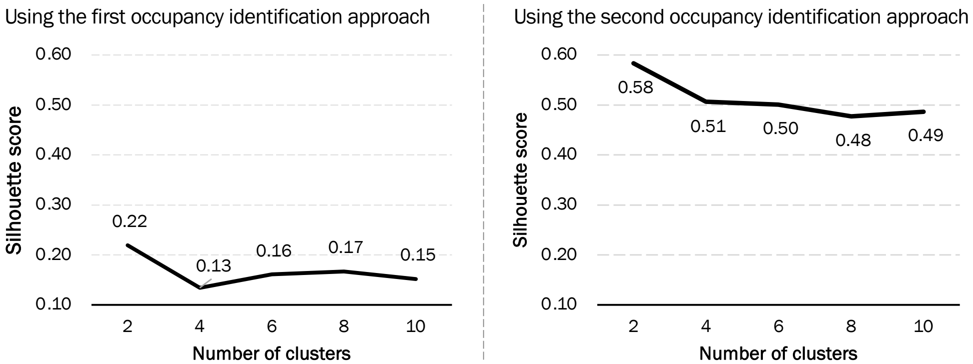


Figure 12. The silhouette scores of the clusters, obtained from two approaches.

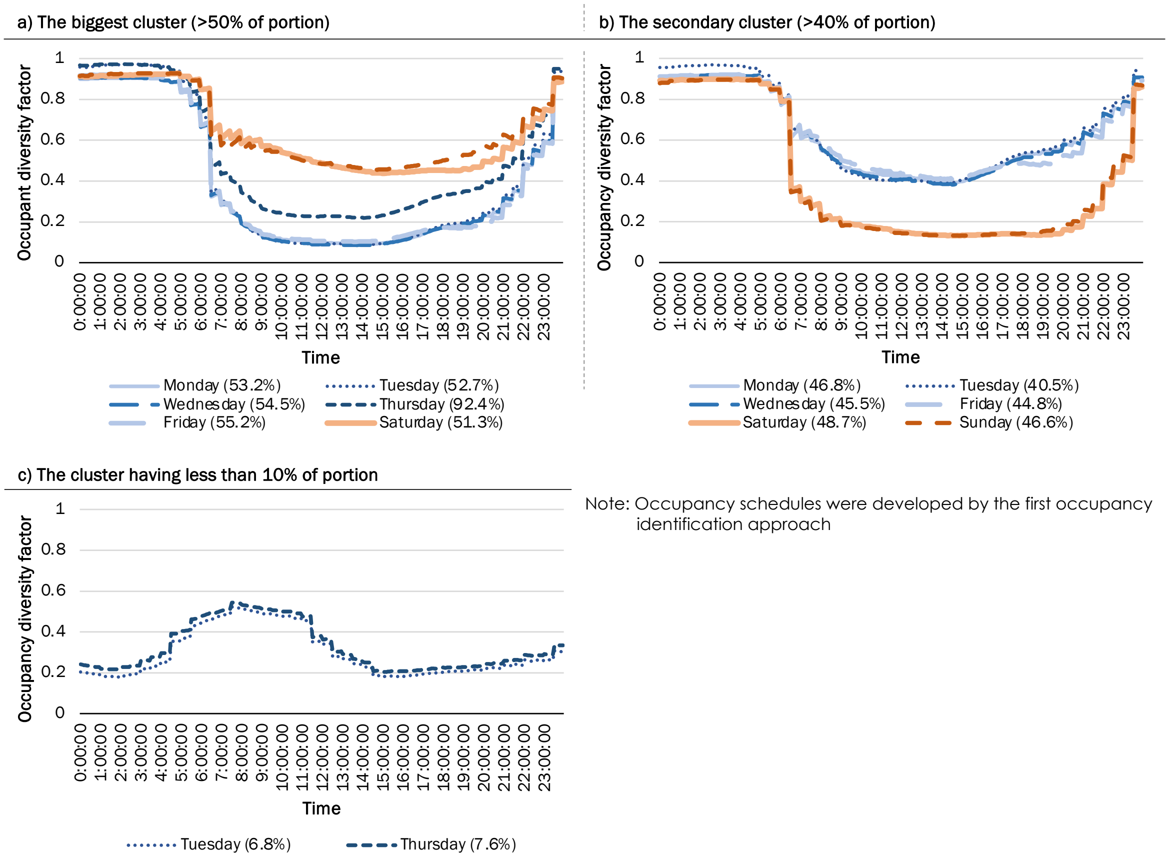
The silhouette scores demonstrate this trend (Figure 12). The silhouette scores from the first approach were much lower than the ones from the second approach because of diversity in clusters. In other words, the schedules from the first approach were formed fewer compact clusters and the largest cluster in the second approach contributed to the higher silhouette scores. In both approaches, the highest silhouette score was shown when two clusters were formed. Hence, the representative occupancy schedules were Figure 9 (a) and Figure 10 (a).

3.2.3 Analyses of impactful parameters

In this subsection, the optimal number of clusters was automatically selected as explained in Figure 7. Then, the impact of three parameters (day of the week, house type, and location) was examined.

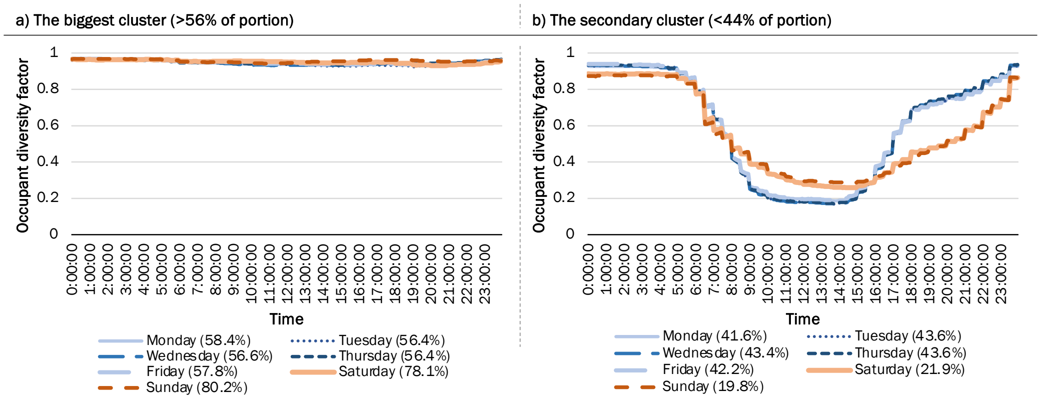
Day of the week: Figure 13 shows the representative occupancy schedules for each day. The occupancy schedules on Tuesdays were grouped into three clusters, but the rest were two clusters. The largest cluster found on each day was put in Figure 13 (a) to demonstrate how the dominant cluster varied by this parameter. Then, the rest are grouped based on their populations: Figure 13 (b) has the second largest clusters which were over 40%, and Figure 13 (c) has the clusters with populations less than 10%. In Figure 13 (a) and (b), occupancy schedules on Mondays, Tuesdays, Wednesdays, and Fridays had similar patterns. They overlapped with each other with marginal gaps. Occupancy schedules on Saturdays and Sundays were also similar.

The biggest cluster on Thursdays (i.e., the dark blue dashed line) had values lower than the weekend ones, but higher than the remaining weekday ones. The second cluster possessed a marginal population and showed an irregular pattern of having the peak value in the morning, which was also found on Tuesdays. This cluster was induced by the occupancy schedules that the sleep schedule was assigned in the morning by the users.



Note: Occupancy schedules were developed by the first occupancy identification approach

Figure 13. Representative occupancy schedules for each day using the first approach.



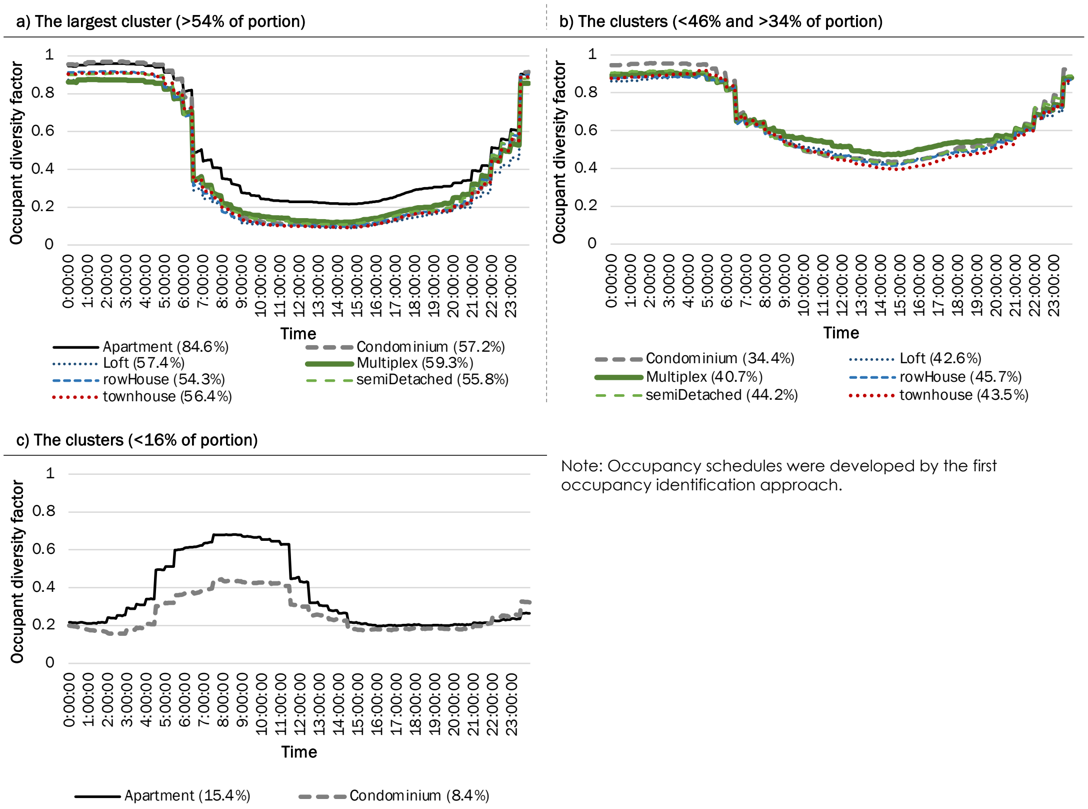
Note: Occupancy schedules were developed by the second occupancy identification approach

Figure 14. Representative occupancy schedules for each day using the second approach.

With the second approach (Figure 14), the dominant occupancy schedules had the same pattern: the house was mostly occupied throughout the day. The secondary clusters consisted of two patterns, distinguishing weekdays from weekends. During weekends, the houses were more likely to be occupied from 8:00 am to 4:00 pm but less occupied at other times.

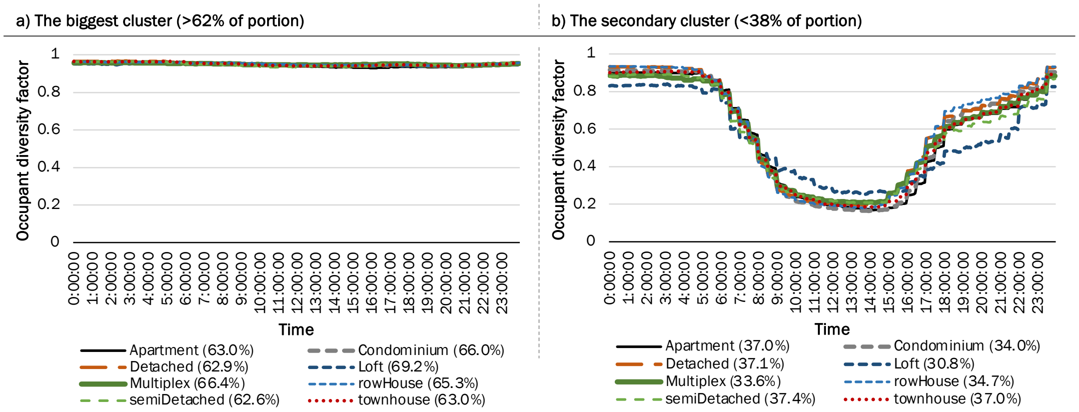
In short, except for the Thursday ones in the first occupancy detection approach (Figure 13 (a)), the patterns of occupancy schedules on weekdays were similar and the patterns of weekends were similar. These results indicated that the residential occupancy schedules are impacted by the *day* parameter.

House type: Figure 15 presents the representative occupancy schedules for each *house* type. The number of clusters was either two or three. The Condominium type was the only one with three clusters. When it comes to the largest cluster in each house type, the occupancy schedules almost overlapped with each other. The one from the Apartment type showed higher values, but the overall shape was similar. Given that its portion (84.6%) is significantly higher than other clusters, it could be the reason for having such results. When it comes to the clusters in Figure 15 (b), six clusters resemble each other closely. Two clusters in Figure 15 (c) were identified in the Apartment and Condominium types.



Note: Occupancy schedules were developed by the first occupancy identification approach.

Figure 15. Clusters identified in each house type using the first approach



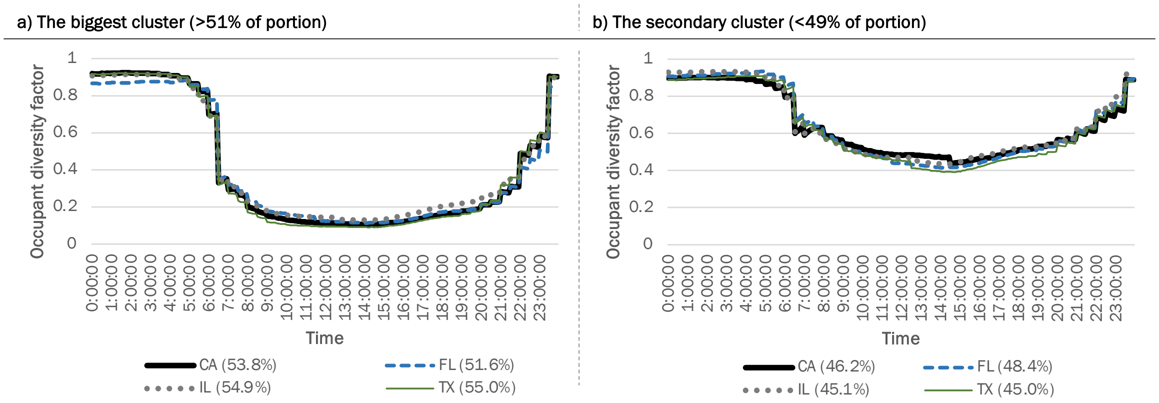
Note: Occupancy schedules were developed by the second occupancy identification approach

Figure 16. Clusters identified in each house type using the occupancy schedules made with the second approach.

In the second approach, the dominant schedules in different *house* types held a similar pattern: The house was almost always occupied throughout the day. The second cluster also showed a similar pattern that the houses started to be less occupied from 06:00 am, reached their minimum at around 02:30 pm, and regained their occupants until midnight. The loft type demonstrated a slightly distinct pattern: The occupancy diversity factors were lower in the morning, higher in the afternoon, and lower in the evening. However, the overall tendency was similar.

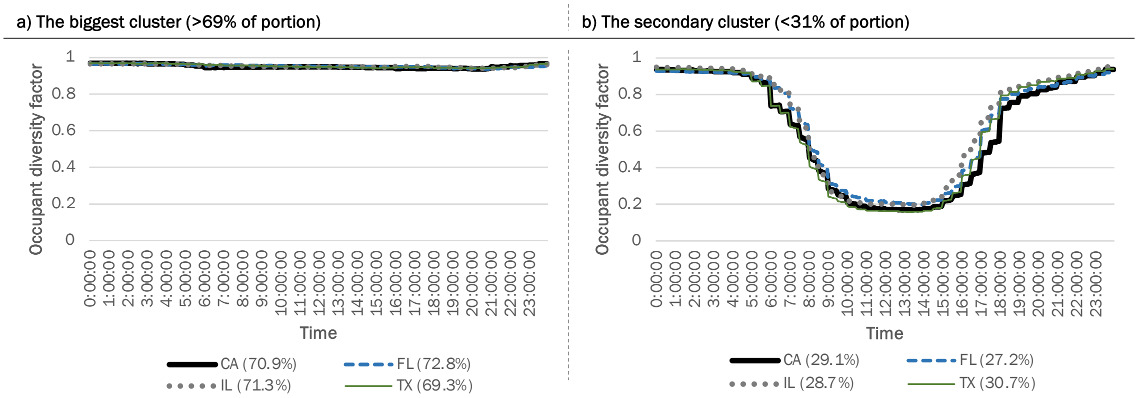
These results indicate that the *house* type was not an impactful parameter in residential occupancy schedules.

Location (state): The time zone was not taken into consideration in this analysis since the occupants would behave without consideration of time differences in their lives. In both approaches, two clusters were identified in each state and each cluster showed similar patterns regardless of location (Figure 17 and Figure 18). This result shows the insignificant role of location in residential occupancy schedules.



Note: Occupancy schedules were developed by the first occupancy identification approach

Figure 17. The clusters identified in four states using the occupancy schdules developed by the first approach.



Note: Occupancy schedules were developed by the second occupancy identification approach

Figure 18. the clusters identified in four states using the occupancy schedules developed by the second approach.

3.2.4 Comparison against U.S. standardized occupancy schedules for residential buildings

All available occupancy schedules in standards or manuals have a one-hour interval, hence the occupancy schedules clustered in this study were recalculated to have the same time interval. Then, the most dominant clusters, representing a weekday and weekend, were plotted in Figure 19.

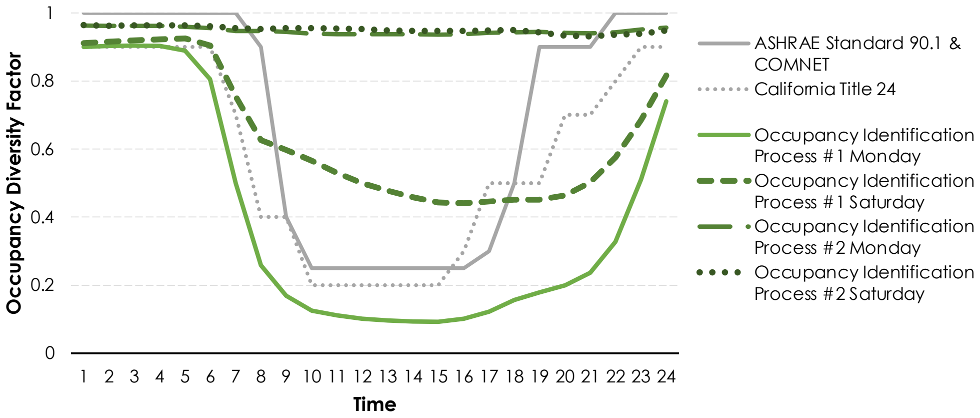


Figure 19. the data-driven occupancy schedules identified in this study from the two occupancy identification approaches, compared against standardized occupancy schedules.

The representative occupancy schedules identified in this research were very different from the occupancy schedules in standards and manuals. The data-driven schedules indicated that a single occupancy schedule may not represent the possible occupancy patterns in residential buildings. Specifically, multiple clusters were identified in different days of the week and their patterns were distinctively formed depending on the day. From the perspective of reliability of BEM, diverse occupancy schedules could be introduced, and the energy use intensity can be statistically analyzed.

4 Residential Occupancy Schedule Simulator (ROSS)

After the research efforts introduced in Section 3, the Residential Occupancy Schedule Simulator (ROSS) has been developed and made available open source at GitHub to the public:

https://github.com/humanbuildingsynergy/ResidentialOccupancyScheduleSimulator.

ROSS creates a stochastic residential occupancy schedule, enabling the BEM community an easier access to the data-driven stochastic residential occupancy schedules. The results in Section 3 demonstrated that the two occupancy identification approaches and day played a significant role in occupancy schedules, but house type and state did not. Hence, this tool asks users to choose the preferred occupancy identification approach, and the day of the week to customize the schedule upon their needs. Also, users can select three levels of randomness in ROSS: low, medium, and high. More details are given in the following subsections.

4.1 Methodology

4.1.1 Inhomogeneous Markov Chain and Inverse Function Method.

ROSS modified the method proposed by Page et al. [25] to stochastically generate a presence schedule (a series of zeros and ones). Specifically, ROSS excluded the consideration of having long periods of absence (e.g., vacations) in the model since such periods might not present the *representative* occupancy schedule. Other modifications are explained at the end of this subsection with the limitations of the original method.

The ROSS needs two inputs from users: *day of the week* and *randomness*. The former is used to identify the representative occupancy schedule on the chosen day. The latter is to determine the level of randomness imposed in stochastic schedules and uses the number of presence schedules stochastically generated through the inhomogeneous Markov chain using the clustered schedules identified in Section 3. The inhomogeneous Markov chain has three assumptions: (1) the state at the next time step solely relies on the current state, (2) the probability of transition is time-dependent, and (3) the probability of transition is from the present state to either the same state or its opposite state. Therefore, four probabilities of transition (T) exist at each time step: presence to presence (T_{pp}); absence to absence (T_{aa}); presence to absence (T_{pa}); absence to presence (T_{ap}).

Also, Equation (4) can be devised from the third assumption.

$$\begin{aligned} T_{pp}(t) + T_{pa}(t) &= 1 \\ T_{ap}(t) + T_{aa}(t) &= 1 \end{aligned} \tag{4}$$

Then, the probability of being present at the time step $t + 1$ is calculated as follows:

$$P(t+1) = P(t)T_{pp}(t) + \big(1 - P(t)\big)T_{ap}(t) \tag{5}$$

In other words, the probability of being present at the next time step is determined by either keeping the same presence state or by transitioning from absence to presence. Therefore, the probability of transition from presence to presence is calculated as:

$$T_{pp}(t) = \frac{P(t) - 1}{P(t)} T_{ap}(t) + \frac{P(t+1)}{P(t)} \tag{6}$$

In order to determine the values of $T_{ap}(t)$ and $T_{pp}(t)$, a parameter defined as the parameter of mobility, is introduced as the ratio between the probability of change of current state over that of no change (Equation (7)).

$$m(t) \coloneqq \frac{T_{ap}(t) + T_{pa}(t)}{T_{aa}(t) + T_{pp}(t)} \tag{7}$$

This parameter is considered to be a constant value, rather than a variable that keeps changing at each time step, in ROSS. Then, the transition probabilities of T_{ap} and T_{pp} can be calculated as shown in Equations (8) and (9). The remaining transition probabilities can be calculated from Equation (4).

$$T_{ap}(t) = \frac{m-1}{m+1} P(t) + P(t+1) \tag{8}$$

$$T_{pp}(t) = \frac{P(t) - 1}{P(t)} \left[\frac{m-1}{m+1} P(t) + P(t+1) \right] + \frac{P(t+1)}{P(t)} \tag{9}$$

Upon calculating all transition probabilities, an inverse function method is used to decide the occupancy state at each time step. The inverse function method works as follows: (1) Under the first assumption, the current state determines which transitional probabilities are utilized (either the pair of $T_{ap}(t)$ and $T_{aa}(t)$ or another pair of $T_{pp}(t)$ and $T_{pa}(t)$, (2) a random value is generated from a uniform distribution, where the stochasticity is implemented, (3) this random value is compared against the two transitional probabilities after building a cumulative distribution function, then (4) the occupancy state, either presence or absence, gets determined.

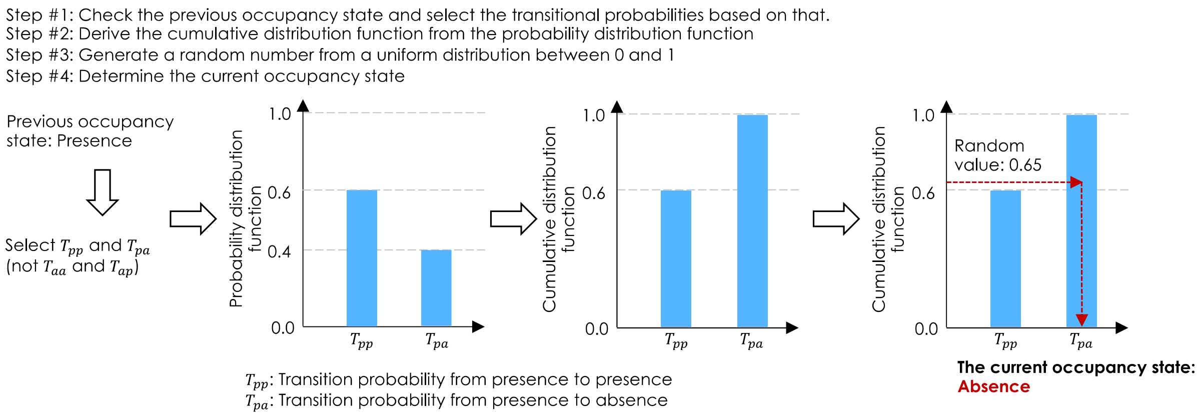

Step #1: Check the previous occupancy state and select the transitional probabilities based on that.
Step #2: Derive the cumulative distribution function from the probability distribution function
Step #3: Generate a random number from a uniform distribution between 0 and 1
Step #4: Determine the current occupancy state
Previous occupancy state: Presence
Select T_{pp} and T_{pa} (not T_{aa} and T_{ap})
Probability distribution function
Cumulative distribution function
Cumulative distribution function
1.0
0.6
0.4
0.0
T_{pp}
T_{pa}
Random value: 0.65
T_{pp}: Transition probability from presence to presence
T_{pa}: Transition probability from presence to absence
The current occupancy state:
Absence

Figure 20. Schematic representation of how presence schedules are created.

In Figure 20, the previous state is presence, hence $T_{pp}(t)$ and $T_{pa}(t)$ are taken into consideration. A random value was created from a uniform distribution determines the current state. In the figure, the probability of having the same state at the next step (T_{pp}) is 0.60. Since a random value is 0.65, above 0.60, the occupancy state becomes absence at the next time step. In the end, the simulator generates a single presence schedule (i.e., a time series of presence or absence). A stochastic occupancy schedule is created by running the simulator several times, decided by the users (details in Section 4.2) and averaging the presence schedules. In other words, ROSS averages multiple time series of zeroes and ones to get a new time series representing the occupancy diversity factors in each time step as shown in Figure 21.

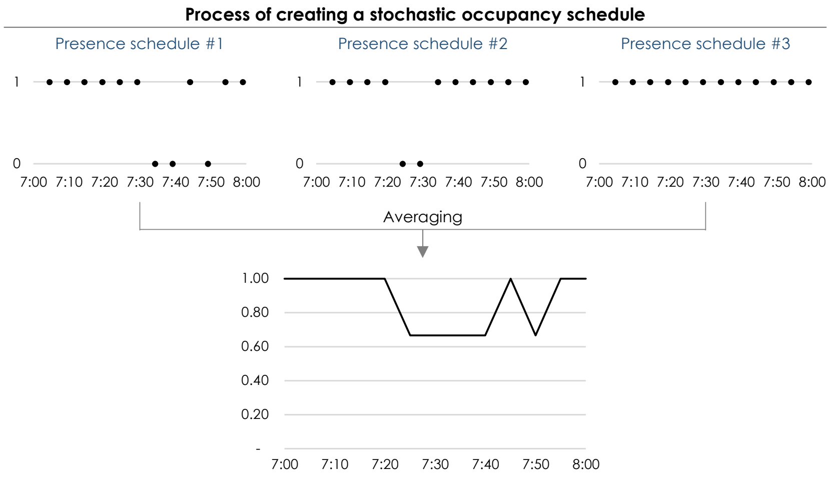


Figure 21. Process of creating a stochastic occupancy schedule through stochastic presence schedules

4.1.2 Modifications in ROSS

The original approach in [25] had several issues that need to be addressed by ROSS. As explained in [25], for some cases, the transitional probabilities could become larger than one. This frequently happens when there is a large gap between $P(t)$ and $P(t+1)$. In this study, such cases were tackled by updating the parameter of mobility until the transitional probabilities become less than one. In addition, when the parameter of mobility is set to near 0.0, the frequency of having the absence state increases significantly. Hence, ROSS uses 1.0 as the default value.

The second limitation was having the consistent occupancy state at the beginning of the stochastic presence schedule. Specifically, the initial occupancy state had to be presence or absence to determine the occupancy state at the next timestep. This would not represent the stochasticity of the occupancy schedule the ROSS creates. To tackle this limitation, using the fact that 24 hours of the day repeat, the ROSS creates 25 hours of occupancy states. Specifically, the simulator starts from 00:00 am and ends at 00:55 am not 11:55 pm and uses the probabilities of the first one hour for the last one hour. Then, the first one hour of the occupancy states was replaced with the last one hour of the occupancy states (Figure 22), and then, the first 24 hours of data were extracted as the results. This modification maintains the use of inhomogeneous Markov chain but addresses the limitations of having the same initial state.

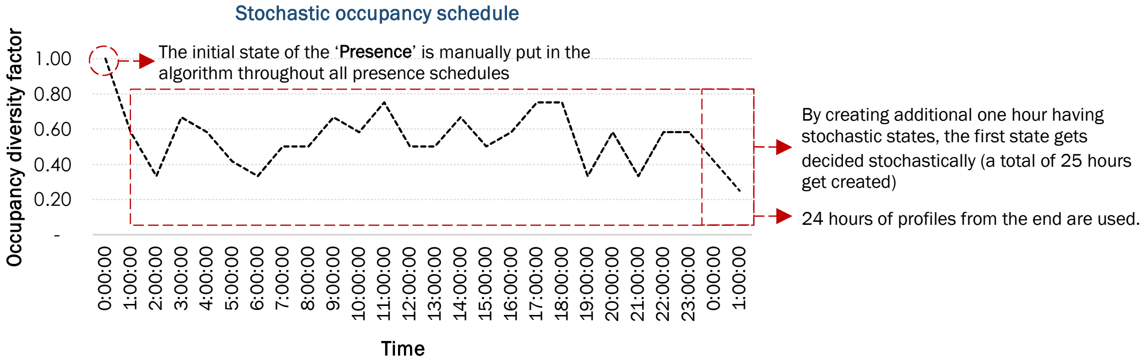
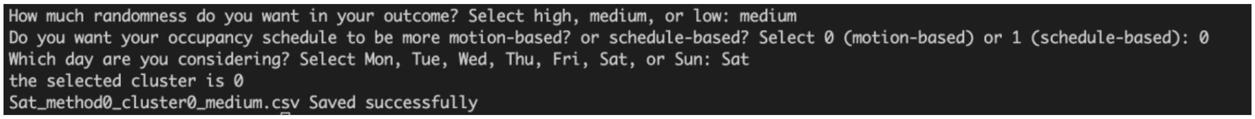


Figure 22. Schematic illustration: addressing the problem of having the same initial state in the inhomogeneous Markov chain.

4.2 *Execution and Example Results*

ROSS mainly works in the command line interface (Figure 23) and generates a csv file that has a stochastic occupancy schedule at a five-minute interval. The first three lines in Figure 23 are to obtain inputs from the user. The randomness is decided by the number of presence schedules calculated in the occupancy schedule.



```
How much randomness do you want in your outcome? Select high, medium, or low: medium
Do you want your occupancy schedule to be more motion-based? or schedule-based? Select 0 (motion-based) or 1 (schedule-based): 0
Which day are you considering? Select Mon, Tue, Wed, Thu, Fri, Sat, or Sun: Sat
the selected cluster is 0
Sat_method0_cluster0_medium.csv Saved successfully
```

Figure 23. The command line interface of ROSS

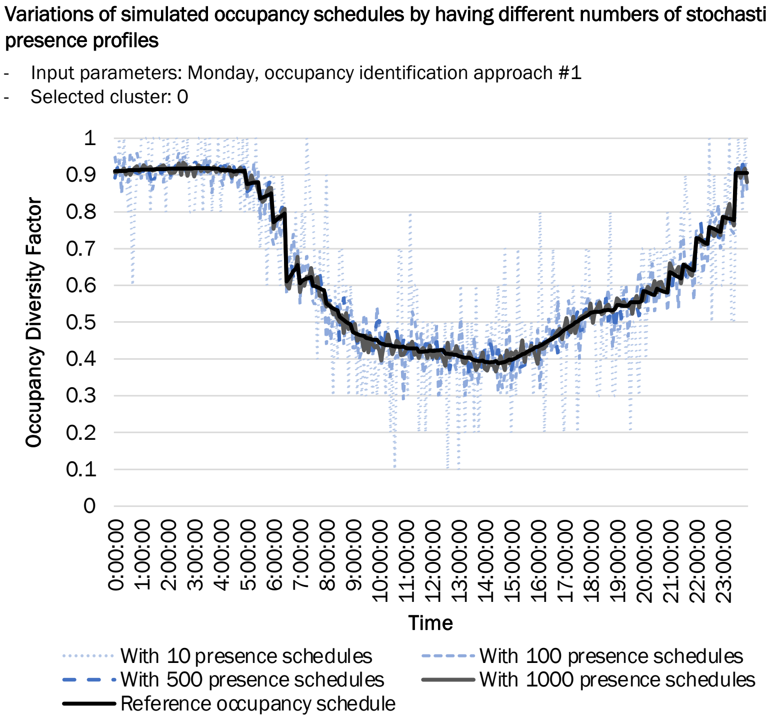
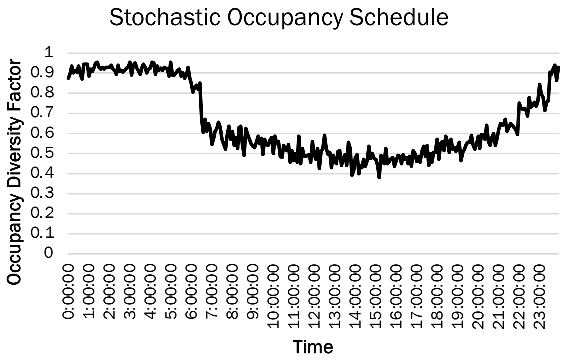


Figure 24. Variations of occupancy schedules, generated by ROSS, depending on the number of presence schedules in calculations.

ROSS averages all stochastic presence schedules by referencing one of the representative occupancy schedules, hence the number of presence schedules influences the final schedule considerably. As shown in Figure 24, when the number of presence schedules increased, the final schedule became much similar to the referenced schedule. Hence, users get to select three levels of randomness in ROSS for their applications: low, medium, and high. Each option creates different numbers of presence schedules (low: 1000, medium: 100, and high: 10). These values were heuristically determined by the authors and are devised for the sake of users' convenience and their specific needs. Users can adjust the numbers in the code upon their needs, when necessary. The second question refers to the occupancy identification approach. As the first approach relied heavily on motion data, this approach is named motion-based and the second approach is based on user-inputted schedules, so it was referred to as schedule-based. Lastly, the user gets to select the day of the week.

Once the simulator has all the input parameters and gets to run, it selects one of the representative occupancy schedules (the fourth line in Figure 23) for reference. Then, a stochastic occupancy schedule is saved in the *result* folder with the name indicated in the fifth line. Figure 25 presents example results of ROSS that users would obtain.



- Input parameters: medium randomness, occupancy identification approach #1, and Sunday

Figure 25. An example results from ROSS

5 Discussion and Limitations

Smart thermostats play an increasingly important role to realize occupant-centric and grid-interactive buildings and this research is one of the pioneer studies that examined the potentials of a large-scale smart thermostat dataset from the perspective of BEM or occupant behavior modeling. Hence, it is worth sharing some of the insights of using such a dataset.

The most challenging part of this study was the absence of the ground truth. Although the authors came up with assumptions to process and reason about the available data, we might have missed some situations or scenarios. These situations could be short visits to the houses, occupants being inactive in houses, occupants being out of the range of motion sensors, pets activating the motion sensors, and so forth. It was possible to come up with other assumptions (e.g., once a motion gets detected during the home schedule, it is assumed that residents stayed in the house during the whole 'Home' schedule), but ambiguity remained. This study could not validate some of the occupancy schedules identified in this study such as the ones with the peaks in the morning (e.g., Figure 13 (c)). It is possible that the occupants with such thermostat settings might utilize the residential building for a different purpose. Another schedule, almost always occupied throughout the day (e.g., Figure 16 (a)) could be used by retired occupants, similar to the schedule proposed by CEN EN 16798 [13], but it could be commuters who assigned the schedules in the thermostats, which may not correctly represent their occupancy. The second approach reduces the potential of representing the actual occupancy dynamics because the *Away* schedule can only be impacted by the sensor readings. As a consequence, the results demonstrated that half of the sample population stayed at home, which may not be entirely true. As such, the second approach can only be used to households who behave very regularly along with the predefined schedules in the thermostats. Again, the absence of ground truth data complicates the interpretation of the results.

However, one of the key contributions of this study is to reveal the roles of three parameters (day of the week, house type, and location) in occupancy schedules. The insignificant roles of house type and location indicate that such house-related features may not be relevant to how occupants occupy their houses. The metadata of the ecobee dataset contained valuable information about each user such as number of residents, floor area, age of home, and source of cooler and heater. However, when it comes to analyzing occupant behavior, insufficient occupant-related data was collected by the vendor. Specifically, possible influential factors could be occupation, education level, age of residents, etc. However, such metadata was missing. When it comes to the day of the week, the results from this study were slightly different from the previous studies that collected data from an extensive number of occupancy sensors (e.g., [6]), which could adopt the dynamics of occupancy with high resolutions. Given that the customers had varied numbers of occupancy sensors and they might not be configured most effectively, this research might not reveal the occupancy schedules with the highest resolution.

Another limitation is the consumers of the smart thermostat may not be a purely random sampling of the whole population, or in other words, the sample set might be biased. Specifically, a smart thermostat is often adopted by early adopters or house owners. Hence, occupant behavior analyses, even in this study, may not fully represent all potential patterns. This limitation would be slowly resolved when smart thermostats are widely adopted in residential buildings. Also, it is worth noting that the results of this study (i.e., representative occupancy schedules in the simulator) should be updated consistently to absorb more demographic features despite its unprecedented scale.

This study used the operational data before COVID-19, so the occupancy schedules during such a pandemic crisis were not included. This crisis heavily influenced the occupant behaviors; therefore, it could be a future direction of this research. As mentioned, this study used the thermostat data collected in the U.S. and it is possible that the residents in other countries occupy their homes differently.

Lastly, the ecobee DYD dataset did not contain energy use data. For example, the results of this study could have been verified using energy use data. The occupied houses tend to consume more energy, hence some of the assumptions could be improved. Also, one of the goals of occupant behavior analyses is to estimate/predict energy use in buildings. Among the factors impacting energy use behaviors in buildings, occupant behavior is the area where the least knowledge has been developed. Hence, enhancing the energy data collection process expands the possible analyses of energy behavior in buildings through a large dataset. However, it is clear that the ecobee DYD dataset opened a gate to see data-driven representative occupancy schedules at an unprecedented scale. The most valuable attribute of this dataset is that it is totally free. The researchers do not need to install or maintain additional sensors to collect the data. With some adjustments, building modelers, analysts, and researchers will definitely benefit from this data-sharing campaign. Lastly, from a sustainability perspective, this digital transformation of building systems, as exampled by this study with smart thermostats, will lead to a new paradigm shift of trading energy between prosumers – expanding the use of renewable energy resources – or developing new regional policies for occupant health at diverse scales [34, 35].

When it comes to ROSS, a series of updates will take place after internal discussions, and interactions with the users in GitHub. First of all, the current version randomly chooses one of the clusters available in the day of the week (e.g., Monday has two schedules). This allows additional stochastic characteristics in the tool, but the users might end up referencing the schedule that might not reflect their needs. Also, the authors are considering sharing all the presence schedules generated in the middle of the algorithm as they contain the stochastic nature. Then, the users can diversify their analyses utilizing the stochastic occupancy schedules. Another consideration is offering a graphical user interface as some prospective users are not familiar with the command line operation. This update will invite more users to ROSS. Also, the ecobee DYD dataset is expanding consistently, hence this research can be replicated for the sake of updating ROSS. Since ROSS is written in Python and outputs csv files, they can readily be integrated with building energy simulation programs like EnergyPlus. For example, the Python Energy Management System in EnergyPlus allows manipulation of input values through Python programming and the Schedule:Compact objects, often used for occupancy schedules, can be updated with few lines of codes.

6 Conclusion

Occupant behavior substantially impacts the energy performance in buildings and the presence of occupants is a fundamental aspect of human building interaction that influences building operations. However, occupancy schedules in residential buildings were inadequately investigated due to privacy issues and the lack of data. This study leveraged the large-scale smart thermostat data from the ecobee DYD dataset to identify the typical occupancy schedules in residential buildings through a data-driven approach. Further, this research developed an open-source program that stochastically generates a residential occupancy schedule. This study shed light on the potentials/limitations of analyzing a large dataset gathered by smart thermostats. Having ambiguity in the collected dataset was the starting point and remained a challenge. Moreover, the potential directions of how the dataset could be improved are discussed. As a future research direction, energy use intensity of diverse types of residential buildings can be analyzed with the occupancy schedules from ROSS and the validity of the occupancy schedules from ROSS can be assessed.

Acknowledgment

This material is based upon work supported by the National Science Foundation under grant #1663513. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Wang's work was supported by the Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083). Hong's contribution to this work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Office, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Reference

1. Malekpour Koupaei, D., Song, T., Cetin, K.S., and Im, J., *An assessment of opinions and perceptions of smart thermostats using aspect-based sentiment analysis of online reviews.* Building and Environment, 2020. **170**: p. 106603.
2. Huchuk, B., O'Brien, W., and Sanner, S., *A longitudinal study of thermostat behaviors based on climate, seasonal, and energy price considerations using connected thermostat data.* Building and Environment, 2018. **139**: p. 199-210.
3. Davis, J.A. and Nutter, D.W., *Occupancy diversity factors for common university building types.* Energy and Buildings, 2010. **42**(9): p. 1543-1551.
4. Deru, M., Field, K., Studer, D., Benne, K., Griffith, B., Torcellini, P., Liu, B., Halverson, M., Winiarski, D., Rosenberg, M., Yazdanian, M., Huang, J., and Crawley, D., *U.S. Department of Energy Commercial Reference Building Models of the National Building Stock.* 2011, National Renewable Energy Laboratory.
5. American Society of Heating, R.a.A.-c.E., *Standard 90.1 User's manual*, in *Building Envelope.* 2016: Atlanta, GA.
6. Duarte, C., Van Den Wymelenberg, K., and Rieger, C., *Revealing occupancy patterns in an office building through the use of occupancy sensor data.* Energy and Buildings, 2013. **67**: p. 587-595.
7. Duarte, C., Van Den Wymelenberg, K., and Rieger, C., *Revealing occupancy patterns in office buildings through the use of annual occupancy sensor data.* 2013, Idaho National Laboratory (INL).
8. ecobee. *Donate your data.* 2020; Available from: *https://www.ecobee.com/donate-your-data/*.
9. Huchuk, B., Sanner, S., and O'Brien, W., *Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data.* Building and Environment, 2019. **160**: p. 106177.
10. Huchuk, B., O'Brien, W., and Sanner, S., *Exploring smart thermostat users' schedule override behaviors and the energy consequences.* Science and Technology for the Built Environment, 2020: p. 1-24.
11. Ueno, T. and Meier, A., *A method to generate heating and cooling schedules based on data from connected thermostats.* Energy and Buildings, 2020. **228**: p. 110423.
12. Mitra, D., Steinmetz, N., Chu, Y., and Cetin, K.S., *Typical occupancy profiles and behaviors in residential buildings in the United States.* Energy and Buildings, 2020. **210**: p. 109713.
13. Standard, E., *CEN Standard 16798.* 2019.
14. Scott, J., Brush, A.J.B., Krumm, J., Meyers, B., Hazas, M., Hodges, S., and Villar, N., *PreHeat: controlling home heating using occupancy prediction*, in *Proceedings of the 13th international conference on Ubiquitous computing.* 2011, Association for Computing Machinery: Beijing, China. p. 281–290.
15. Jung, W. and Jazizadeh, F., *Human-in-the-loop HVAC operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions.* Applied Energy, 2019. **239**: p. 1471-1508.
16. Hailemariam, E., Goldstein, R., Attar, R., and Khan, A., *Real-time occupancy detection using decision trees with multiple sensor types*, in *Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design.* 2011, Society for Computer Simulation International: Boston, Massachusetts. p. 141–148.

17. Yang, Z., Li, N., Becerik-Gerber, B., and Orosz, M., *A systematic approach to occupancy modeling in ambient sensor-rich buildings.* SIMULATION, 2013. **90**(8): p. 960-977.

18. Li, D., Balaji, B., Jiang, Y., and Singh, K., *A wi-fi based occupancy sensing approach to smart energy in commercial office buildings*, in *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*. 2012, Association for Computing Machinery: Toronto, Ontario, Canada. p. 197–198.

19. Kleiminger, W., Mattern, F., and Santini, S., *Predicting household occupancy for smart heating control: A comparative performance analysis of state-of-the-art approaches.* Energy and Buildings, 2014. **85**: p. 493-505.

20. Dodier, R.H., Henze, G.P., Tiller, D.K., and Guo, X., *Building occupancy detection through sensor belief networks.* Energy and Buildings, 2006. **38**(9): p. 1033-1043.

21. American Society of Heating, R.a.A.-c.E., *ANSI/ASHRAE/IES Standard 90.1-2019 -- Energy Standard for Buildings Except Low-Rise Residential Buildings*. 2019.

22. Kleiminger, W., Beckel, C., and Santini, S., *Household occupancy monitoring using electricity meters*, in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2015, Association for Computing Machinery: Osaka, Japan. p. 975–986.

23. Kleiminger, W., Beckel, C., Staake, T., and Santini, S., *Occupancy Detection from Electricity Consumption Data*, in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*. 2013, Association for Computing Machinery: Roma, Italy. p. 1–8.

24. Wang, D., Federspiel, C., and Rubinstein, F., *Modeling occupancy in single person offices.* Energy and Buildings, 2005. **37**: p. 121-126.

25. Page, J., Robinson, D., Morel, N., and Scartezzini, J.L., *A generalised stochastic model for the simulation of occupant presence.* Energy and Buildings, 2008. **40**(2): p. 83-98.

26. Yang, Z. and Becerik-Gerber, B., *Modeling personalized occupancy profiles for representing long term patterns by using ambient context.* Building and Environment, 2014. **78**: p. 23-35.

27. Soltanaghaei, E. and Whitehouse, K., *WalkSense: Classifying Home Occupancy States Using Walkway Sensing*, in *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*. 2016, Association for Computing Machinery: Palo Alto, CA, USA. p. 167–176.

28. Yang, Z., Ghahramani, A., and Becerik-Gerber, B., *Building occupancy diversity and HVAC (heating, ventilation, and air conditioning) system energy efficiency.* Energy, 2016. **109**: p. 641-649.

29. Hosseinihaghighi, S., Panchabikesan, K., Dabirian, S., Webster, J., Ouf, M., and Eicker, U., *Discovering, processing and consolidating housing stock and smart thermostat data in support of energy end-use mapping and housing retrofit program planning.* Sustainable Cities and Society, 2022. **78**: p. 103640.

30. Dabirian, S., Panchabikesan, K., and Eicker, U., *Occupant-centric urban building energy modeling: Approaches, inputs, and data sources - A review.* Energy and Buildings, 2022. **257**: p. 111809.

31. ecobee, *Donate Your Data Researcher Handbook*, ecobee, Editor. 2019. p. 1-9.

32. Aghabozorgi, S., Seyed Shirkhorshidi, A., and Ying Wah, T., *Time-series clustering – A decade review.* Information Systems, 2015. **53**: p. 16-38.

33. Rousseeuw, P.J., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 1987. **20**: p. 53-65.

34. Koley, S., *Challenges in Sustainable Development of Smart Cities in India.* Sustainability, 2020. **13**(4): p. 155-160.

35. Bheemarasetti, S. and Patruni, R.P., *19 - DER, energy management, and transactive energy networks for smart cities*, in *Solving Urban Infrastructure Problems Using Smart City Technologies*, J.R. Vacca, Editor. 2021, Elsevier. p. 411-432.