Data Article

Datasets for occupancy profiles in apartment-style student housing for occupant behavior studies and application in building energy simulation

Leila Nikdel [a,*], Alan E.S. Schay [b], Daqing Hou [b], Susan E. Powers [a]

[a] Institute for a Sustainable Environment, 8 Clarkson Ave., Clarkson University, Potsdam NY USA
[b] Department Electrical and Computer Engineering, 8 Clarkson Ave., Clarkson University, Potsdam NY USA

ARTICLE INFO

ABSTRACT

Building energy simulation (BES) tools fail to capture diversity among occupants' consumption behaviors by using simple and generic occupancy and load profiles, causing uncertainties in simulation predictions. Thus, generating actual occupancy profiles can lead to more accurate and reliable BES predictions. In this article, occupancy profiles for apartment-style student housing are presented from high-resolution monitored occupancy data. A geo-fencing app was designed and installed on the cellphones of 41 volunteer students living in student housing buildings on Clarkson University's campus. Occupants' entering and exiting activities were recorded every minute from February 4 to May 10, 2018. Recorded events were sorted out for each individual by the date and time of day considering 1 for 'entered' events and 0 for 'exited' events to show the probability of presence at each time of day. Accounting for excluded days (234 days with errors and uncertainties), 1,096 daily occupancy observations were retained in the dataset. Two methods were used to analyze the dataset and derive weekday and weekend occupancy schedules. A simple averaging method and K-means clustering techniques were performed [1]. This article provides the

input datasets that were used for analysis as well as the outputs of both methods. Occupancy schedules are presented separately for each day of a week, weekdays, and weekend days. To show differences in students' occupancy patterns, occupancy schedules in 7 clusters for weekdays and 3 clusters for weekend days are provided. These datasets can be beneficial for modelers and researchers for either using provided occupancy schedules in BES tools or understanding occupant behaviors in student housing.

Specifications Table

Subject	Energy
Specific subject area	Building Energy Modeling and Simulation, Building Occupant Behaviors
Type of data	Tables
	Spreadsheet files
How data were acquired	Measurements by installing a geo-fencing app on cellphones
	Data analysis by SPSS and Excel
Data format	Cleaned measurement data, Analyzed
Parameters for data collection	To reduce privacy concerns, a random ID was assigned to each of the volunteer participants for anonymity. Students were assured that their specific location within the apartment complex was not known, just their movement into and out of the geo-fence boundary.
Description of data collection	A geo-fencing app was designed and installed on the cellphones of 41 volunteer students living in student housing buildings on Clarkson University's campus. Occupants' entering and exiting activities were recorded every minute from February 4 to May 10, 2018, with days in the semester breaks (February 21–25 and March 16–25) excluded.
Data source location	Clarkson University
	Potsdam, New York
	United States
	Latitude and longitude: 44.66866, -74.98493
Data accessibility	Repository name: Mendeley Data [12]
	http://dx.doi.org/10.17632/hx5mp695tv.1
Related research article	L. Nikdel, A. E. S. Schay, D. Hou, and S. E. Powers, "Data-driven Occupancy Profiles for Apartment-style Student Housing," Energy and Buildings, vol. 246, p. 111070, 2021, http://doi.org/10.1016/j.enbuild.2021.111070.

Value of the Data

- Identical residential buildings in shape, envelope, and location could have 300% difference in energy consumption due to occupant behaviors [2]. Generating data-driven occupancy profiles supports the need for a better understanding of occupant behaviors and consumption patterns in buildings as many human-building interaction activities require the presence of occupants [3]. Furthermore, BES tools fail to capture the diversity in consumption behavior by using simple and generic occupancy and load profiles [4], which cause uncertainties in simulation predictions [5]. Thus, generating actual occupancy profiles leads to more accurate and reliable BES predictions [6], which is very important in achieving targeted energy goals.
- Modelers and researchers can benefit from the generated occupancy profile data for application in BES tools for more accurate predictions as well as understanding and analyzing occupant behaviors in student housing. Accurate occupancy profiles can also be helpful and beneficial for demand and distribution management of electrical grids as identifying factors for demand and peak load [7].

- The data collection and analysis methods to derive occupancy profiles in this article can be used as a reference by similar studies that aim to obtain occupancy profiles or explore diversity in occupant behaviors. Various machine learning techniques can be applied to distinguish and characterize similar habit and behavior patterns among students. Occupancy schedules can be further correlated with occupants' consumption behaviors to identify personality profiles among students that can be used for behavior interventions.
- This data addresses the need to provide occupancy profiles for a population with specific household composition and occupation at a regional level which have not been explored before [3,4,8,9]. Errors and uncertainties in data collection and analysis methods explained in this article also can be helpful for future efforts and experiments.
- Datasets in this article illustrate differences of occupants' habits and behaviors among students living in apartment-style student housing. This further emphasizes on the necessity of occupant behavior studies and exploring diversity in occupant behaviors in buildings.

1. Data Description

This article has three spreadsheet files including tables for occupancy data before and after analysis. Each dataset is an MS Excel Workbook, with separate spreadsheet pages to organize data.

Dataset 1 includes cleaned data for occupancy schedules with the 1-minute resolution over 2 to 78 days for each of 36 participants in separate spreadsheets. A summary spreadsheet is provided in this dataset to show the number of total days, weekdays, and weekends for each participant.

Dataset 2 includes all days of data from all participants aggregated into 10-minute averages for all weekdays (843 days) and all weekends (253 days) in separate spreadsheets. For each day in this dataset, the date and the individual code for the participant are provided for potential further analysis.

Dataset 3 presents the results of both simple averaging and K-means clustering of the 10-minute occupancy data. For each method, occupancy patterns in apartment-style student housing are presented in two spreadsheets separately for weekdays and weekend days. Within each spreadsheet for simple averaging method, occupancy schedule for each day of week and for the average of all weekdays or weekend days are provided (Fig. 4 in [1]). These schedules provide the average occupancy probability (0–1) for the given 10-minute period across all students and all entries for the given day of the week (e.g., averaged over all Friday data).Within the spreadsheets for K-means clustering technique, clusters represent different occupancy patterns in weekdays (7 clusters) and weekend days (3 clusters) (Fig. 5 and Fig. 6 in [1]). Percent values in the column headings for clusters show dominant occupancy patterns vs. outliers in student housing.

2. Experimental Design, Materials and Methods

To derive occupancy profiles for students living in apartment-style student housing, an experiment was designed and conducted in four buildings of the Woodstock Village apartment complex on Clarkson University's campus (Potsdam NY, USA) from February 4 to May 10, 2018. To track occupants' entering and exiting activities into their apartments, a geo-fencing app was designed and installed on the cellphones of 41 volunteer residents. A specific geo-fence for Woodstock Village was defined around the whole village including all four buildings (Fig. 1 in [1]) to reduce privacy concerns and the Hawthorne Effect [4]. It was assured that students understood that their specific location within the Village was unknown. In addition, a random code was assigned to each individual for anonymity. Limitations with the data collection assumptions are elaborated in [1].

Occupants' entering and exiting activities were recorded every minute over the experiment period, with days in the semester breaks excluded (February 21–25 and March 16–25). All

Table 1

Examples of errors and assumptions to address data uncertainties.

Errors/ Assumptions	Examples from one participant				
	Event	Date	Time	Input for Schedule	Comments
Error Type 1	Exited: Woodstock	5/4/2018	15:26	'0' until the time for the next event (16:32)	First event was kept, the next two entries at the same time were ignored.
	Entered: Woodstock	5/4/2018	15:26	Ignored	
	Exited: Woodstock	5/4/2018	15:26	Ignored	
	Entered: Woodstock	5/4/2018	16:32	'1' until the time for the next event (17:47)	
Error Type 2	Exited: Woodstock	5/4/2018	17:47	Blank	Last event was kept, and for the events before that, it was left blank in the schedule since it was unknown what really happened at those times.
	Exited: Woodstock	5/4/2018	18:41	Blank	
	Exited: Woodstock	5/4/2018	23:00	'0' until the time for the next event (7:22 in the next day).	
Assumption 3	Entered: Woodstock	5/5/2018	7:22	'1' until the time for the next event (14:35).	Since there is no data for the following day, from the hour of the last event to midnight was left blank.
	Exited: Woodstock	5/5/2018	14:35	'0' until 15:00. Blank from 15:00 to midnight	
Assumption 4	Entered: Woodstock	5/7/2018	17:32	Ignored	The whole day was excluded – no data for the day before or after.
Assumption 5	Entered: Woodstock	5/9/2018	9:12	'1' until the time for the next event.	Blank from midnight to 9:12- no data for prior day.

recorded events for all participants in the server (6,664 events) were sorted out for each individual by the date and time of day using Python. Five participants were excluded from the experiment because of glitches in operating the geo-fencing app. To calculate probability of presence at each time of day, 1 was assigned for 'entered' events and 0 for 'exited' events. Therefore, 1,330 daily occupancy observations were generated for 36 individuals.

Data errors in recorded events were observed due to different possible reasons such as battery life of cellphones, wireless connections used to send the signals from cellphones to the server, or other unknown causes. Two types of errors in the dataset were identified (Table 1). Type 1 errors include both entering and exiting events recorded at the same time stamp. Type 2 errors include multiple sequential entering or exiting activity recorded with different time stamps. To clean the dataset, assumptions were made to deal with these errors:

1. A two-minute interval was considered as the minimum realistic time span between entering and exiting events. Thus, when two or more events were recorded in less than 2 minutes, only the first event was considered, and the rest were ignored.
2. Days with more than 3 errors of any type were excluded from dataset as the level of uncertainties deemed the occupancy observation for that day unreliable.
3. When the data for the following day after was missing, it was uncertain when the signal was lost. Thus, the last recorded event was considered for the hour of that event. The remaining hours for that day were left blank. (as shown in Table 1)
4. Days with only one recorded event were excluded from dataset when no data were recorded for the day before or after. (as shown in Table 1)
5. When data for the previous day was missing, times from midnight to the time of first event in that day were left blank. (as shown in Table 1)

The dataset was regenerated for all days and all individuals by incorporating all assumptions into the Python code. Ultimately, with excluding 234 daily observations with errors and uncertainties, 1,096 daily occupancy observations were retained in the dataset (Dataset 1). To prepare data for further analysis, the number of data points in a daily occupancy profile was reduced
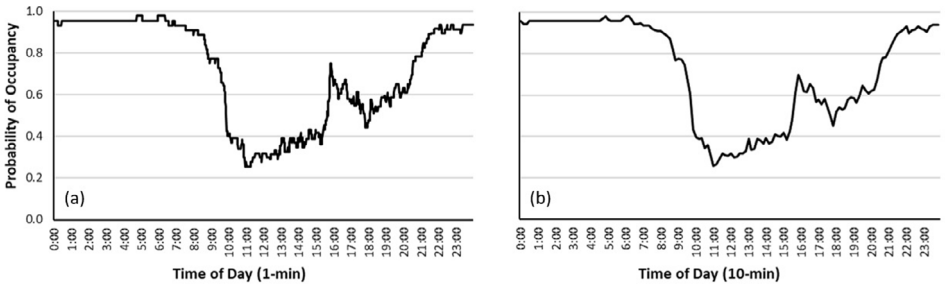
Fig. 1. Average weekday occupancy schedule for one individual with both (a) one-minute and (b) ten-minute data resolutions.
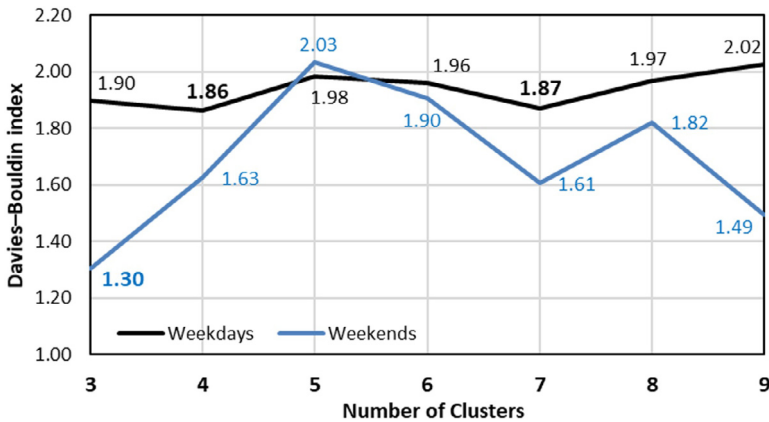
Fig. 2. Cluster distance performance analysis with the Davies–Bouldin Index.

to faciliate data analysis. Data for daily occupancy observations were aggregated from 1-minute into 10-minute intervals using Excel (Dataset 2). Fig. 1 illustrates that reducing the resolution did not impact the quality of data.

Dataset 2 was used as the input for further analysis to derive occupancy profiles for apartment-style student housing. Two methods were used to derive weekday and weekend occupancy schedules: a simple averaging method and a K-means clustering technique. Simple averaging was conducted in Excel and K-means clustering was performed with SPSS software. Mean values were calculated for each 10-minute interval in a day over all daily occupancy observations for a day of week (all Fridays for example). Number of daily observations for each day of a week ranged from 120 to 191 days. Mean values were also calculated separatelty for all weekdays (Mon. to Fri.-843 days) and all weekend days (Sat. & Sun.-253 days) (Dataset 3). For K-means clustering technique, the optimal number of clusters was determined using the Davies–Bouldin Index (DBI) [10] and the dendrogram produced by hierarchical clustering [11] of the dataset. The DBI was calculated considering 3 to 9 clusters for weekdays and weekend days separately (Fig. 2). For weekend days, three clusters was considered the optimal number of clusters as yeilded the lowest DBI (1.30). For weekdays, since both 4 and 7 clusters similarly yielded the lowest value for DBI (1.86 and 1.87 respectively), the dendrogram from hierarchical clustering of the dataset was used to determine the optimal number of clusters (Fig. 2 in [1]). Examining the dendrogram showed that the algorithm with 4 clusters produces distinct clusters, but also merges some occupancy patterns that have low similarities into one cluster. Hence, for weekdays, 7 clusters was considered as the optimal number of clusters for producing more realistic occupancy patterns (Dataset 3). The percentage of the number of observations allocated in a

cluster to the total number of observations indicates if that cluster represents a dominant or outlier occupancy pattern.

Ethics Statement

Informed consent (IRB# 17-28.2) was obtained for experimentation with human subjects.

CRediT Author Statement

Leila Nikdel: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization; **Alan E.S. Schay:** Software, Investigation, Data Curation; **Daqing Hou:** Conceptualization, Software, Writing - Review & Editing, Supervision, Funding acquisition; **Susan E. Powers:** Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Funding

References

[1] L. Nikdel, A.E.S. Schay, D. Hou, S.E. Powers, Data-driven occupancy profiles for apartment-style student housing, Energy Build. 246 (Sep. 2021) 111070, doi:10.1016/j.enbuild.2021.111070.

[2] A.S. Bahaj, P.A.B. James, Urban energy generation: the added value of photovoltaics in social housing, Renew. Sustain. Energy Rev. 11 (2007) 2121–2136, doi:10.1016/j.rser.2006.03.007.

[3] B.F. Balvedi, E. Ghisi, R. Lamberts, A review of occupant behaviour in residential buildings, Energy Build. 174 (2018) 495–505, doi:10.1016/j.enbuild.2018.06.049.

[4] D. Yan, T. Hong, International Energy Agency, EBC Annex 66: Definition and Simulation of Occupant Behavior in Buildings, May 2018 ISBN 978-0-9996964-7-7[Online]. Available: https://annex66.org/sites/default/files/2018FinalReport/Annex%2066%20Final%20Report%20-%20read.pdf.

[5] K. Sun, T. Hong, A framework for quantifying the impact of occupant behavior on energy savings of energy conservation measures, Energy Build. 146 (Jul. 2017) 383–396, doi:10.1016/j.enbuild.2017.04.065.

[6] A. Mahdavi, F. Tahmasebi, B. Gunay, W. O'Brien, S. D'Oca, Annex 66: Definition and Simulation of Occupant Behavior in Buildings_ Technical Report: Occupant Behavior Modeling Approaches and Evaluation, Nov. 2017 ISBN 978-0-9996964-1-5[Online]. Available: https://annex66.org/sites/default/files/2018FinalReport/Annex%2066%20Deliverable%20-%20Occupant%20behavior%20modeling%20approaches%20and%20evaluation.pdf. Accessed March 25, 2019.

[7] V. Aragon, S. Gauthier, P. Warren, P.A.B. James, B. Anderson, Developing English domestic occupancy profiles, Build. Res. Inf. 47 (4) (May 2019) 375–393, doi:10.1080/09613218.2017.1399719.

[8] M. De Simone, C. Carpino, Dafni Mora, Stephanie Gauthier, V. Aragon, Gülsu Ulukavak Harputlugil, IEA EBC Annex 66 – Subtask A Deliverable, Reference Procedures for Obtaining Occupancy Profiles in Residential Buildings, May 2018 [Online]. Available:https://annex66.org/sites/default/files/2018FinalReport/Subtask%20A%20Deliverable%20-%20Reference%20procedures%20for%20obtaining%20occupancy%20profiles%20in%20residential%20buildings.pdf.

[9] S. Carlucci, et al., Modeling occupant behavior in buildings, Build. Environ. 174 (2020) 106768, doi:10.1016/j.buildenv.2020.106768.

[10] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1 (2) (Apr. 1979) 224–227, doi:10.1109/TPAMI.1979.4766909.

[11] J. Sander, X. Qin, Z. Lu, N. Niu, A. Kovarsky, Automatic extraction of clusters from hierarchical clustering representations, in: Advances in Knowledge Discovery and Data Mining, 2003, pp. 75–87, doi:10.1007/3-540-36175-8_8. Berlin, Heidelberg.

[12] [Dataset] L. Nikdel, A.E.S. Schay, D. Hou, S.E. Powers, Datasets for occupancy profiles in student housing for occupant behavior studies and application in building energy simulation, Mendeley Data V1 (2020). doi: http://dx.doi.org/10.17632/hx5mp695tv.1.