Science and Technology for the Built Environment

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uhvc21

Stochastic residential occupancy schedules based on the American Time-Use Survey

Diba Malekpour Koupaei, Kristen S. Cetin & Ulrike Passe

Science and Technology for the Built Environment, (2022) **28**, 776–790

ISSN: 2374-4731 print / 2374-474X online
DOI: 10.1080/23744731.2022.2087536

Stochastic residential occupancy schedules based on the American Time-Use Survey

DIBA MALEKPOUR KOUPAEI[1], KRISTEN S. CETIN[2]*, AND ULRIKE PASSE[3]

[1]*Department of Civil, Construction and Environmental Engineering, Iowa State University, Ames, IA, USA*
[2]*Department of Civil and Environmental Engineering, Michigan State University, East Lansing, MI, USA*
[3]*Department of Architecture, Iowa State University, Ames, IA, USA*

The energy performance of residential buildings is closely correlated with occupants' behavior and their schedules. Moreover, the energy consumption of heating, ventilation, and air conditioning (HVAC) systems, which are by far the biggest contributor to energy consumption in residential buildings, is controlled by the presence or absence of occupants in a building. Thus, accurate occupancy presence and activity profiles are important to determine actual energy demands and corresponding control schedules for residential buildings. Conventionally, building energy simulation tools typically use a single generic and static occupancy profile to represent a building's occupancy schedule, regardless of day type or household size. However, literature in the field suggests that there is significant potential for improvement to allow for more flexibility and accuracy in the calculation of occupancy. The objective of this study is thus to develop a stochastic building occupancy model and propose it as a realistic replacement for the conventional generic static schedules. This three-state stochastic occupancy model is based on the 2019 American Time-Use Survey (ATUS) and considers the differences among weekdays, Saturdays, and Sundays. In this model, survey respondents are clustered based on the number of residents in their household and a first-order inhomogeneous Markov chain technique is used to generate occupancy presence schedules. The models' results are then validated against the original ATUS data in terms of state probabilities, state durations, and number of state changes throughout the course of a day.

Introduction

In the U.S., residential buildings accounted for more than 22% of total energy consumption, 38% of total electricity use, and 19% of total greenhouse gas (GHG) emissions in 2020 (U.S. Energy Information Administration (EIA), 2021). Moreover, energy consumption from residential buildings is projected to continue to maintain similar levels moving forward, despite technological advancements. This includes a projected average increase of 0.1% per year for the period of 2018–2050 under a business as usual scenario (U.S. Energy Information Administration (EIA), 2019). Thus, the building energy sector, in general, and residential buildings in particular, represents a significant opportunity for accelerating the energy transition and ensuring a low-carbon future (Zhang et al. 2018). The prediction of buildings' energy use, both current and future, plays a crucial role in the realization of this opportunity. Factors that influence a building's energy performance and that are commonly utilized for making such predictions include (1) climate conditions, (2) building envelope, (3) building energy systems, (4) indoor design criteria, (5) building operation and maintenance, and (6) occupant behavior (Yoshino, Hong, and Nord 2017). Of all of these factors, occupant behavior is commonly cited as a major contributor to uncertainty in buildings' energy use predictions and simulations (Hong et al. 2016; Yan et al. 2015). This uncertainty often results in a 'gap' between actual and predicted energy consumption of buildings (Hoes et al. 2009; Hong et al. 2016; Yan et al. 2015).

To address this challenge and the resulting performance gap, in recent years numerous efforts have been made to improve occupancy models in terms of predicting occupants' energy-related behaviors and presence schedules (Hong et al. 2016; Yan et al. 2015). Yan et al. (2015) provides an overview of these efforts, and categorizes them into the following four key areas of improvement: (1) occupant monitoring and data collection, (2) model development, (3) model evaluation, and (4) model implementation into building simulation tools (Yan et al. 2015).

Previous efforts in this field include the development of various stochastic and deterministic models to simulate occupancy and occupant-building interaction at the building or urban scale. Stochastic-based models use statistical

Received June 28, 2021; accepted June 3, 2022
Diba Malekpour Koupaei, MS, Student Member ASHRAE, is a PhD Candidate. **Kristen S Cetin, PhD,** Member ASHRAE, is an Assistant Professor. **Ulrike Passe, Dipl.-Ing,** is a Professor.
*Corresponding author e-mail: cetinkri@msu.edu
Supplemental data for this article is available online at https://doi.org/10.1080/23744731.2022.2087536.

probabilities derived from real data to estimate the change in occupancy level or the occurrence of a specific occupant-building interactions (Happle, Fonseca, and Schlueter 2018). Researchers use different techniques to sense, collect, and manage relevant data about building occupants with different levels of granularity (Yan et al. 2015). Yang, Santamouris, and Lee (2016) suggested the following five overarching categories of occupancy data collection: (1) questionnaires, surveys, or interviews, (2) radio frequency (RF) occupancy sensors, (3) infrared, ultrasound, or video cameras, (4) Carbon dioxide (CO_2) sensors, and (5) global positioning system (GPS), cellular data, wireless local area network (WLAN), and Bluetooth. In IEA-EBC Annex 66 (Yan et al. 2017), a similar classification system was used to categorize occupant sensing technologies, including (1) threshold and mechanical, (2) image-based, (3) motion sensing, (4) radio-based environmental, (5) mixed sensing, (6) human-in-the-loop, and (7) consumption sensing. (Hong et al. 2017) used a different approach for categorizing occupant behavior sensing methodologies, according to which such efforts are either categorized as (1) physical sensing, or (2) nonphysical sensing methods. In this categorization system, survey questionnaires and self-reported data are subjective measurements made by nonphysical sensing methods, while objective measurements, such as smart meters, building data, and indoor and outdoor environmental data, are made by physical sensing methods. Among these data collection methods, time use surveys (TUS) have been used to investigate residential building occupancy and occupant behavior patterns. TUSs are nationwide surveys that document the activities of individuals throughout the day and provide a high-resolution source of data for the daily routine of a representative sample of the population. They have been used in many fields and are projected to have a significant potential for use in building research (Grandjean, Adnot, and Binet 2012).

In the context of buildings' energy modeling, the primary purpose of bottom-up TUS-based models is to simulate individual households' occupancy patterns and model the occupants' energy-related activities to estimate the household's energy demand. The resulting occupancy profile and/or the energy consumption characteristics could be later extrapolated and used as a realistic input to obtain the energy demand of a simulated community or city (Diao et al. 2017; Swan and Ugursal 2009). As such, bottom-up probabilistic models, such as Markov chains, are more suitable for modeling the stochastic nature of human activities (Nijhuis, Gibescu, and Cobben 2016). The Markov chain technique is used as the core of developed models to mimic human behaviors, where it simulates a sequence of events based on the present state and the probability of transition between states (Brémaud 2013). The Markov chain models can generally be divided based on their transition probability matrices. Accordingly, the homogeneous Markov chain models use a fixed time-independent transition probability compared with the non-homogeneous Markov chain, where the transition probability change with time (Brémaud 2013).

The accuracy of TUS-based Markov chain models depends on the size of the population sample used to calculate the transition probability matrices, Change-Matrix-Time, which is defined as per how many minutes a different transition matrix is used, and the simulation time step (Adamopoulou, Tryferidis, and Tzovaras 2016; Flett and Kelly 2016; Osman and Ouf 2021; Zhou et al. 2017). Markov chain models' order is thus another property of a Markov chain that is determined based on the memory status, where a first-order Markov chain is a memory-less model as the Markov property of the subsequent state is influenced only by the current state (Osman and Ouf 2021). Many of the proposed occupancy behavioral models are first order Markov chains (e.g. Richardson, Thomson, and Infield 2008). Alternatively, in the second-order Markov chain, the subsequent state's probability is estimated by two precedent states (Torriti 2014). More recent studies in the field (e.g. Flett and Kelly 2016; Ramírez-Mendiola, Grünewald, and Eyre 2019) have explored the benefits of such models compared to the more common first-order models. While the use of higher-order models was previously shown to be slightly more effective in predicting occupancy with accuracy, the added benefits of such techniques when compared to a first-order model are not significant and the added complexity is also avoided (Flett and Kelly 2016).

Accordingly, this study focuses on the model development and implementation areas of improvement in the aforementioned occupant behavior modeling framework and attempts to bridge the gap between predicted and actual energy consumption of buildings by proposing an accurate yet practical stochastic modeling approach based on the 2019 American Time Use Survey (ATUS) (U.S. Bureau of Labor Statistics (BLS) 2007). The proposed first-order inhomogeneous Markov chain model generates stochastic domestic occupancy presence data with the same characteristics as the ATUS data when given the following two input parameters: (1) type of day (weekday, Saturday, or Sunday) and (2) the number of household members. In addition to the availability of the required inputs, the main advantages of the proposed model compared to those that have been previously proposed (For a complete review of these efforts please refer to Hong et al. (2016) and Yan et al. (2015)) are related to the compatibility of the resulting outputs' formatting for implementation in conventional energy simulation tools. In the following sections of this manuscript, first, the methodology for developing the introduced model is discussed in detail. Then, the resulting model is validated and analyzed.

Methodology

As noted in the introduction, the aim of this study is to propose a model that generates stochastic occupancy data with the same statistical characteristics as the time-use survey on which it is based, notably in terms of state probabilities, state durations, and number of state changes during the course of a day. The developed model distinguishes three important states of occupancy (corresponding to 'absent' (A), 'present and active' (P), 'present and inactive' (S)) and is based on the 2019 ATUS data. ATUS is a yearly survey,

Table 1. A sample diary from the ATUS tagged with presence states.

TRCODE	Activity	TEWHERE	Presence state
030112	Picking up/dropping off household children	3	A
180301	Travel related to caring for & helping household children	12	A
020201	Food and drink preparation	1	P
110101	Eating and drinking	1	P
020203	Kitchen and food clean-up	1	P
010201	Waiting associated w/eating & drinking	−1	P
120303	Television and movies (not religious)	1	P
010101	Sleeping	1	S

supported by the U.S. Bureau of Labor Statistics and conducted by the U.S. Census Bureau, that measures the amount of time people spend doing various activities such as working, watching television, and sleeping (U.S. Census Bureau 2012). Information collected by the ATUS includes the start and stop times of each activity (in minutes), where each activity occurred, and whether the activity was done for a person's profession. Additional information on each respondent, including age, gender, occupation, and region of residence, is also available (U.S. Bureau of Labor Statistics (BLS), 2019). It should be noted that while researchers have explored the limitations of TUSs for such applications before, TUS datasets remain the sole source for occupancy and activity data with a sufficient breadth of respondents to be representative of the overall population and also smaller sub-populations and are thus used as the database in this study (Flett and Kelly 2016; Torriti 2014).

The model proposed in this study uses a first-order time-inhomogeneous Markov chain technique and is based on the 2019 ATUS. The ATUS data consists of 24-hour long diaries recoded at irregular time intervals depending on specific activity durations starting at 4 am. The proposed model, however, adopts a 10-min resolution and is configured to run from midnight to midnight. Overall, the development of the proposed occupancy presence model can be divided into three subsequent steps: (1) data cleaning and processing, (2) data clustering, and (3) model development procedure. In the following sections, these steps are described in detail.

Step 1: Data cleaning and processing procedure

The goal of this step is to prepare and process the ATUS data for use in the proposed occupancy presence model. For the specific purposes of this study, only the Respondent and Activity files from the ATUS database have been used as input. In these files, the "TUCASEID" (a 14-digit identifier) identifies each unique household, and every entry includes activity-level information related to the specific activity code, location, duration, as well as start and stop times (U.S. Bureau of Labor Statistics (BLS), 2019). In the following sections, different sub-steps in this data cleaning procedure are discussed.

Step 1-1: Defining the presence states of the respondents for all activities

First, a presence state of either present and active (P), present and asleep (S), or away (A) is allocated to each diary entry according to the respondents' presence and activity level during that specific activity. This information is derived from the "TEWHERE" variable in the ATUS Activity File. In the U.S. Bureau of Labor Statistics (BLS) (2019), the TEWHERE variable is described as "where were you during the activity?" and a value equal to 1 for this variable stands for presence at the "respondent's home or yard" (U.S. Bureau of Labor Statistics (BLS), 2019). It should be noted that this variable is not collected for activities with activity codes of 0101xx (sleeping or sleeplessness), 0102xx (washing, dressing and grooming oneself), 0104xx (personal/private activities), 500105 (respondent does not provide information), or 500106 (gap/cannot remember) (U.S. Bureau of Labor Statistics (BLS), 2019). In such cases, a value of −1 is assigned to the TEWHERE variable in the ATUS Activity File which indicates that the activity was "out of universe" for the "where" question (U.S. Bureau of Labor Statistics (BLS), 2019). Since most of the listed unregistered activities are most probable to be happening in one's private living space, in this study it is assumed that a value of −1 for TEWHERE is considered to be present at home. Then, activity levels are determined and a presence state of "Sleep" or "Present" is allocated to each of these activities. Table 1 shows how a sample diary is interpreted by this protocol.

Step 1-2: Creating 24-hour diaries for all respondents

While each respondent's entire diary input in the 2019 ATUS database is meant to capture at least a full 24-hour period (starting at 4:00 a.m.), not all are 24 hours long. In other words, the duration of the last activity recorded determines the entire length of a respondent's diary. However, for the specific purposes of this study, it was necessary to remove the extra parts of the diaries, resulting in one 24-hour long diary per respondent. It should be noted that for each activity recorded, two ATUS variables recorded in the ATUS Activity File are essential for this step: (1) "TUSTARTTIM" which stands for "activity start time," and (2) "TUSTOPTIME" which is defined as "activity stop time" (U.S. Bureau of Labor Statistics (BLS), 2019). According to the U.S. Bureau of Labor Statistics (BLS) (2019), both of these variables can take any valid time value between 00:00:00 and 24:00:00 (U.S. Bureau of Labor Statistics (BLS), 2019).

To achieve one 24-hour long diary per respondent, in each respondent's diary, input the TUSTOPTIM for the last

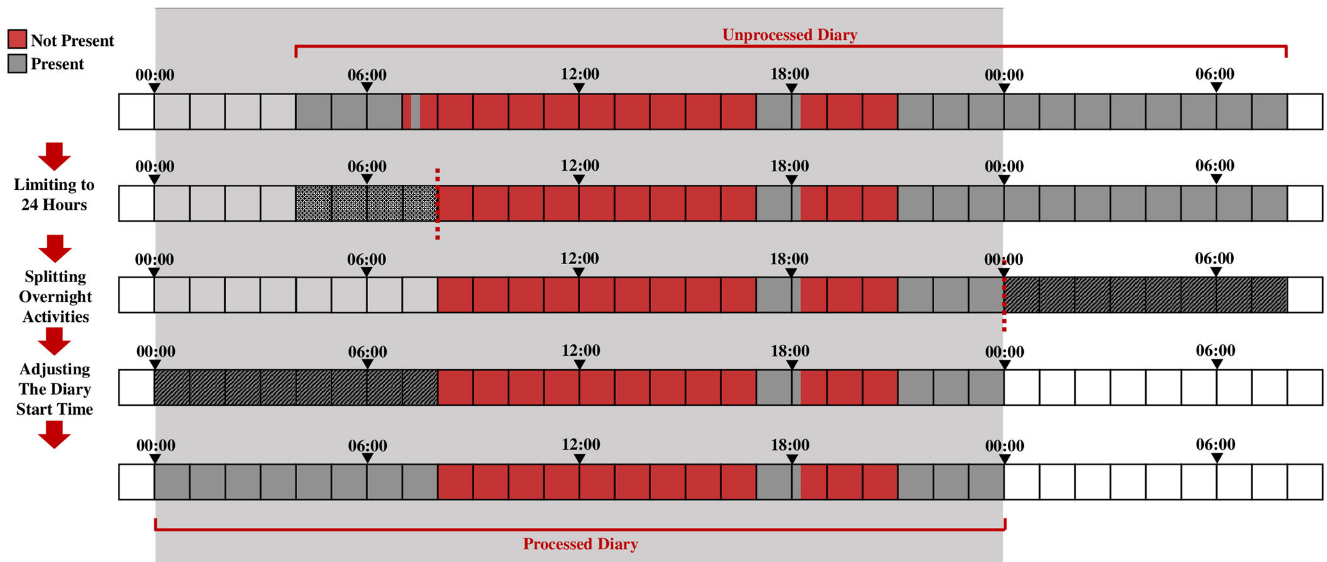

Fig. 1. An example of the Step 1-2 process for modifying the respondents' diaries.

activity recorded is defined as the stop time for the entire diary. Then, the diaries that are longer than 24 hours are cut to be no more than 24 hours long. To do so, if an activity exists that its TUSTOPTIME occurs less than 24 hours away from the defined stop time for the entire diary, then that activity is broken into two parts, at 24 hours before the defined stop time. The remaining portion after 24 hours is then used as the initial activity entry.

It is now necessary to make sure that each diary starts at midnight and finishes at a second midnight, 24 hours later. This means that any activity entries that either occur on the second day or overnight need to be modified. To do so, any activities that start after midnight on the second day are moved to the beginning of the diary. Then, if the second midnight occurs between the TUSTARTTIM and the TUSTOPTIME of the last diary entry, that entry is broken into two parts at midnight. Then, the second part (the one that starts at the second midnight) is moved to the beginning of the diary while the first part is kept in its original spot at the end of the diary. Figure 1 shows how a sample diary is processed according to this procedure.

Step 1-3: Regulating the time steps for all recorded activities
While some time-use surveys are recorded in predefined time intervals, the ATUS time steps are determined by the duration of activities (Centre for Time Use Research 2017; The French National Institute of Statistics and Economic Studies (INSEE), 2010). Moreover, the temporal resolution of the diary recordings in the ATUS is in minutes, while previous studies have suggested that a 10-minute temporal resolution is sufficient for the purposes of building energy use studies (Mahdavi and Tahmasebi 2016; Richardson, Thomson, and Infield 2008; Yan et al. 2015). Therefore, in this step, each activity entry is broken into one or multiple sequential 10-minute activity entries. The process begins at midnight on the first night and the first activity's duration is modified to be exactly 10 minutes long. Next, if the original TUACTDUR for the first activity is more than 10 minutes long, a duplicate of that activity is redefined so that it starts right after the first one and ends 10 minutes later (e.g. minute 20 of the entire diary). This process is repeated until the TUSTOPTIME of the last duplicate goes beyond the TUSTOPTIME of the original activity. Then, that TUSTOPTIME marks the TUSTARTTIM of the next activity in the diary and the same process is repeated for this activity and its succeeding activities until the TUSTOPTIME for an activity marks the midnight for the second night (Figure 2).

At this stage, the 24-hour diary inputs developed in the last step (Step 1-2) are divided into 10-minute time slots and are ready to be used in the occupancy presence model which is explained in the next step (Step 3).

Step 2: Data clustering procedure

In Step 2, the diaries that were cleaned and processed in Step 1 are clustered into homogeneous groups based on select respondent and diary characteristics. These characteristics are selected to be available to the average energy modeler or other possible users of this model and include day of the week, month of the year, meteorological season, housing type, household size, and geographic region/division. It is important to note that in situations where number of occupants are unknown to the energy modeler, previous studies (e.g. Agthe and Billings 2002) have suggested that the number of bedrooms has a high correlation with the number of occupants in a residence and thus can be used as a proxy to calculate the number of residents in the planning and design stage.

First, it is common with occupant behavior models to account for the differences between weekdays and weekend days as occupancy patterns can vary significantly by the day of the week. While some studies have only accounted for differences between weekdays and weekends, others have explored the possibility of separate occupancy schedules for

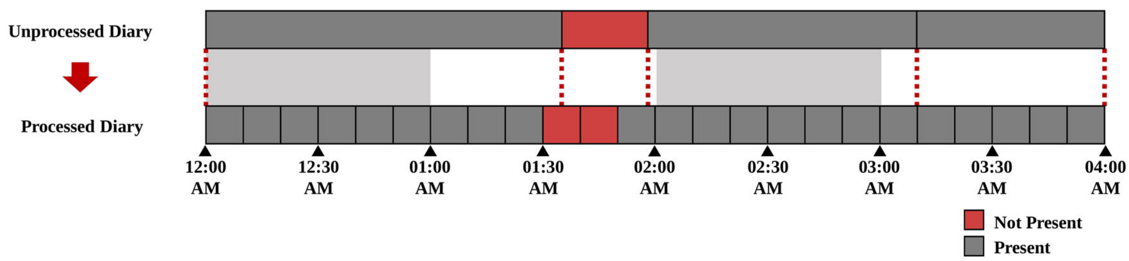

Fig. 2. An example of the Step 1-3 process for modifying the respondents' diaries.

Saturdays and Sundays. In some cases, every day of the week is assigned a unique schedule. Figure 3 shows these three different ways that days of the week have been clustered in previous studies, using the 2019 ATUS database. Each graph shows the percentage of ATUS population in a specific presence state across the 24-hour period that diaries is recorded in. As can be seen, the differences between occupancy presence schedules among different working days of the week are negligible. However, the presence schedule between weekdays and weekend days is clear. As for the variation in presence schedules between Saturdays and Sundays, both of these days share the same general profile for the sleeping state, while their active and away schedules are not identical and show substantial differences in the afternoon and evening hours of the day. Hence, the proposed model in this study captures this diversity of presence and activity profiles between weekdays, Saturdays, and Sundays.

Figures 4 and 5 show the ATUS 2019 databased clustered based on the months and seasons of the year. The season delineations are based on the meteorological season definitions provided by U.S. National Oceanic and Atmospheric Administration, which considers spring in the Northern Hemisphere to include March, April, and May; summer to include June, July, and August; fall to include September, October, and November; and winter to include December, January, and February (The National Oceanic and Atmospheric Administration (NOAA), 2021). Looking at the difference between different state profiles in different months and seasons showed that the variety in occupancy presence and activity schedules during different times of the year is negligible. Therefore, the proposed model does not take monthly and seasonal changes in occupancy presence and activity schedules into consideration.

Another set of clustering efforts focused on the type of housing. This variable is recorded under the "HEHOUSUT" label in the ATUS-CPS File and can take any whole numerical value between 1 and 12. Table 2 below includes the definitions of each of these values and shows the distribution of this variable across the 2019 ATUS Respondent database.

All respondents that have indicated a HEHOUSUT variable equal to 1 are clustered into one group, while all other respondents are clustered into another group. As can be seen in Figure 6, the housing type does not appear to affect the presence profiles significantly and is thus not considered as a clustering variable in the developed model.

Another variable is the location of the respondents' residents. In the ATUS database, this variable includes two values which indicate if a respondents' place of residence is located in either a "metropolitan" or a "non-metropolitan" area (based on the delineation provided by the (U.S. Census Bureau 2021). The results of this clustering analysis (Figure 7) show that the choice of metropolitan or non-metropolitan housing type does not substantially affect occupancy presence schedules. Therefore, this variable is also not considered in the developed model.

Previous studies suggest that data on active occupancy along with household size are the most important source of information when assessing energy load profiles (Abu-Sharkh et al. 2005). Therefore, the next set of clustering efforts in this study is focused on household size and its effect on occupancy schedules (Figure 8). It is evident that while the sleep state profiles are generally the same between different household sizes, the active and away profiles seem to depend on the household size. Accordingly, respondents from households with a higher number of members appear to spend more time away from their place of residence during the daytime hours. Therefore, the study's proposed model considers household size.

The final clustering considerations focused on the geographic location of the respondents. The variables GEDIV and GEREG, which store the data related to the ATUS respondents' locations in terms of geographic division and region respectively, are used for this analysis (U.S. Bureau of Labor Statistics (BLS), 2020). Looking at the distribution of different respondents in various geographic regions (Figure 9) and divisions (Figure 10) showed little to no difference in presence states. Thus, the proposed model does not account for the geographic location of the respondents and considers all respondents' schedules to be the same, regardless of their location of residence.

Overall, the clustering analysis conducted showed that only two variables show meaningful variation in occupancy presence states across the 2019 ATUS database, including household size and day of the week type.

Step 2-1: Creating group clusters based on the respondents' household size

Previous studies suggest that data on active occupancy along with household size are the most important source of information when assessing energy load profiles (Abu-Sharkh et al. 2005). Therefore, as a first step of the clustering process, the respondents' household sizes are determined, and all respondents are grouped based on this variable. In the 2019 ATUS Respondent File, this information is available in

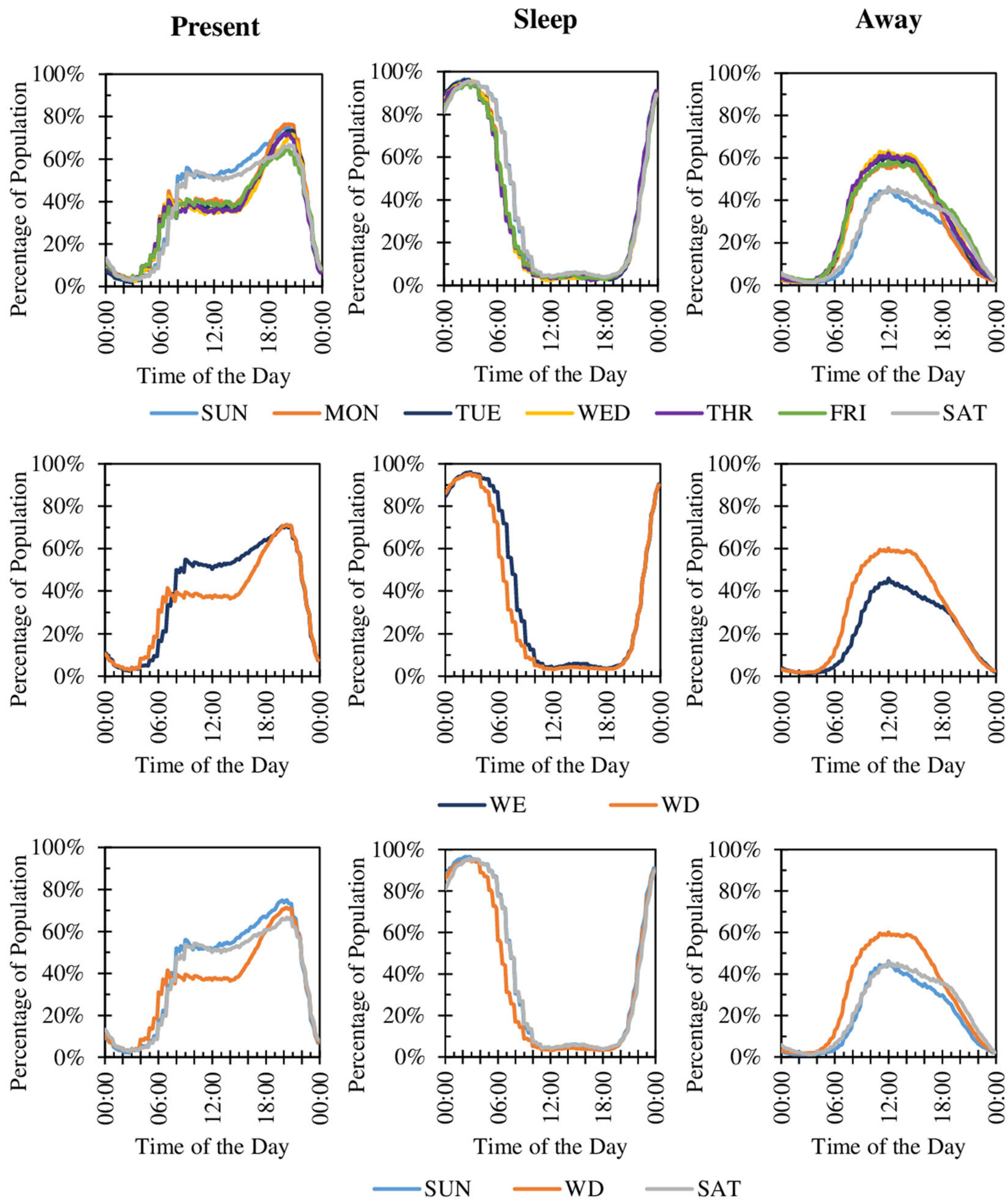

Fig. 3. Presence Status of ATUS respondents based on their diary recording day type.

the "TRNUMHOU." Table 3 shows the distribution of this variable across the 2019 ATUS Respondent database.

Step 2-2: Creating subgroups based on day type

Next, it is common with occupancy behavioral models to account for differences between weekdays and weekends. Therefore, the respondent groups created in the last step (Step 2-1) were divided into subgroups based on the type of the day. The goal of this step was to reveal the variance of the occupancy over a typical week. In the 2019 ATUS Respondent File (and also the Activity Summary File), "TUDIARYDAY" is the variable that holds this information and is defined as the "day of the week of diary day (day of the week about which the respondent was interviewed)" (U.S. Bureau of Labor Statistics (BLS), 2019). Valid entries for this variable are 1 through 7 for Sunday through Saturday, respectively (U.S. Bureau of Labor Statistics (BLS), 2019). Table 4 shows the distribution of this variable within the defined groups across the 2019 ATUS database.

Step 3: Model development procedure

Step 3 is the model development procedure that uses the Markov chain technique to generate ATUS based occupancy presence schedules with the cleaned data prepared and clustered from the previous steps. For probabilistic models, Markov chains are among the more common methods used to stochastically model occupancy and predict occupancy profiles (Mitra et al. 2020), allowing occupancy status to be determined based only on the status at the previous time (Flett and Kelly 2016). Therefore, in order to generate the

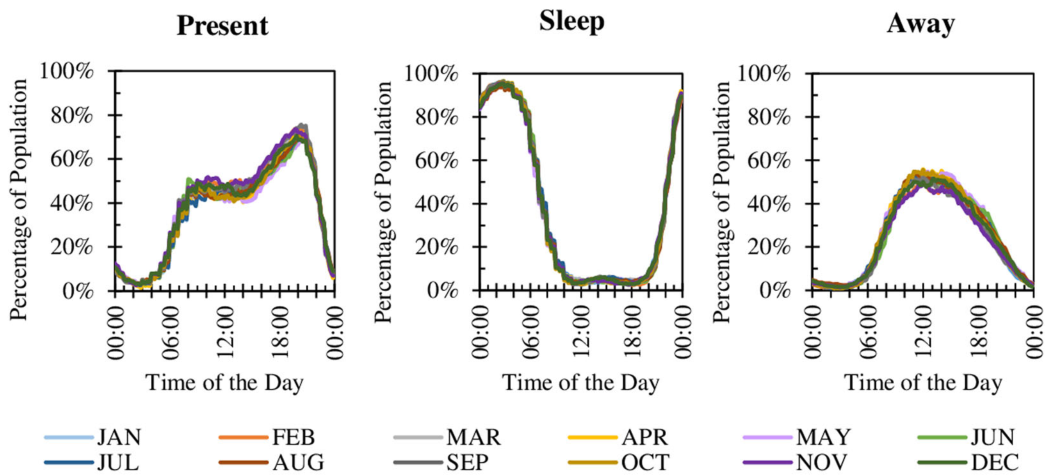

Fig. 4. Presence Status of ATUS respondents by diary recording month.

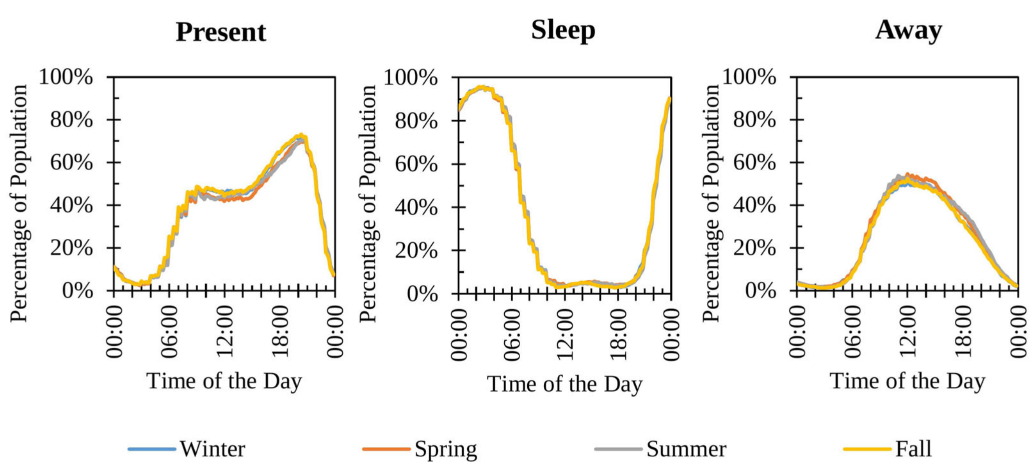

Fig. 5. Presence Status of ATUS respondents by diary recording season.

Table 2. HEHOUSUT variable and its value descriptions.

HEHOUSUT	Description	% of ATUS data
1	House, apartment, flat	95.99%
2	Housing unit in non-transient hotel, motel, etc.	0.04%
3	Housing unit permanent in transient hotel, motel	0.03%
4	Housing unit in rooming house	0.00%
5	Mobile home or trailer with no permanent room added	3.31%
6	Mobile home or trailer with 1 or more rooms added	0.46%
7	Housing unit not specified above	0.11%
8	Quarters not housing unit in rooming/boarding house	0.00%
9	Unit not permanent in transient hotel/motel	0.00%
10	Unoccupied tent site or trailer site	0.00%
11	Student quarters in college dorm	0.00%
12	Other unit not specified above	0.06%

synthetic data, a random number (uniform from 0 to 1) is picked at each timestep and used, together with the appropriate transition probability matrix and with the state at the current timestep, to determine the state at the next timestep. The following steps describe the necessary calculations for developing both the start state probabilities and the transition probability matrices (Richardson, Thomson, and Infield 2008). A first-order Markov chain approach to predict changes in occupancy was used in this research. While the use of higher-order models has been previously shown to be slightly more effective in predicting occupancy with accuracy, the added benefits of such techniques when compared to a first-order model were minimal given the added complexity (Flett and Kelly 2016).

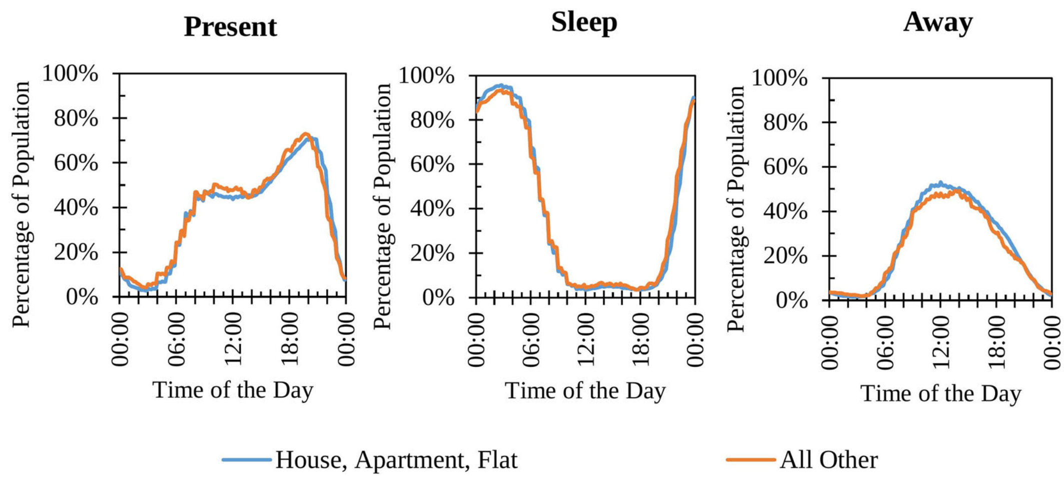

Fig. 6. Presence Status of ATUS respondents based on their housing type.

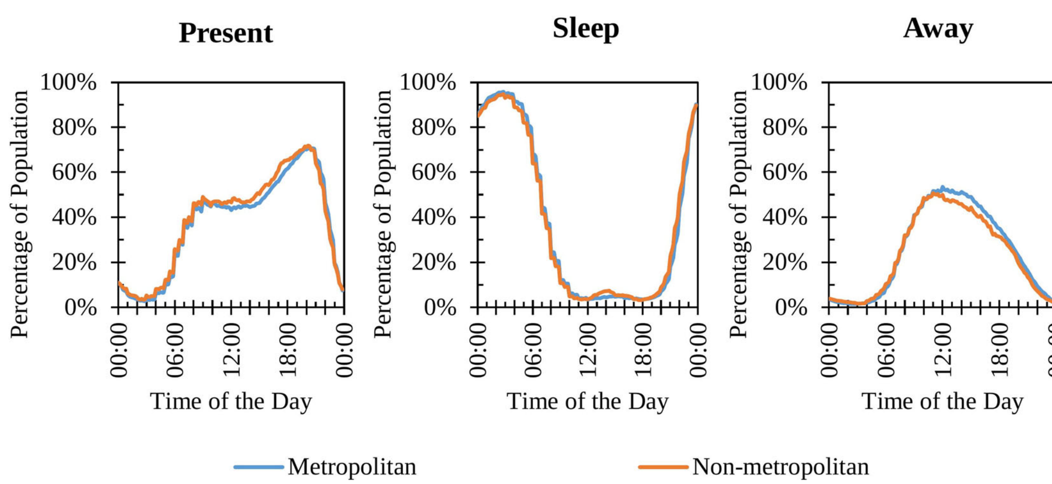

Fig. 7. Presence Status of ATUS respondents based on their housing location type.

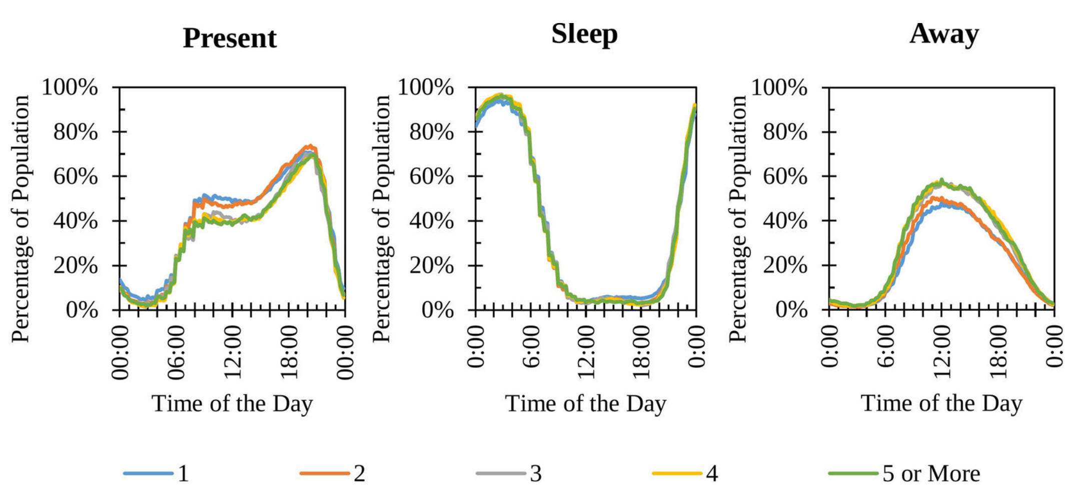

Fig. 8. Presence Status of ATUS respondents based on household size.

Step 3-1: Calculating the start time probability distributions
In order to generate the Markov chain, a start state is needed which describes how probable it is for an individual to be present in their home at midnight at the beginning of the 24-hour period. This should match the probabilities found in the original ATUS data. For instance, out of the 704 entries from three-person households in the weekdays subgroup, 62 indicated "P" (present and active) as their presence state at 00:00, 612 of them indicated "S" (present and sleeping), and the remaining 30 were absent, i.e. "A." Accordingly, the

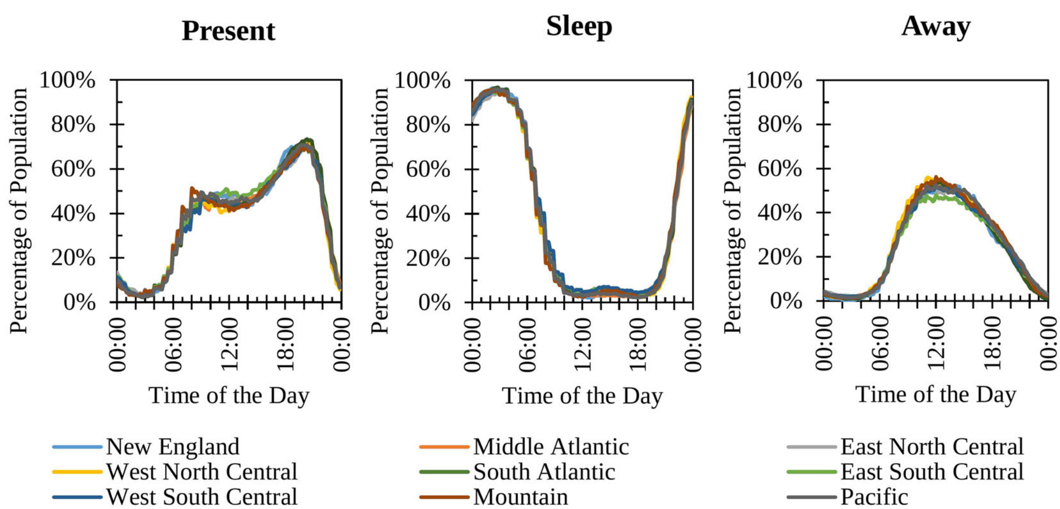

Fig. 9. Presence Status of ATUS respondents based on their housing geographic division.

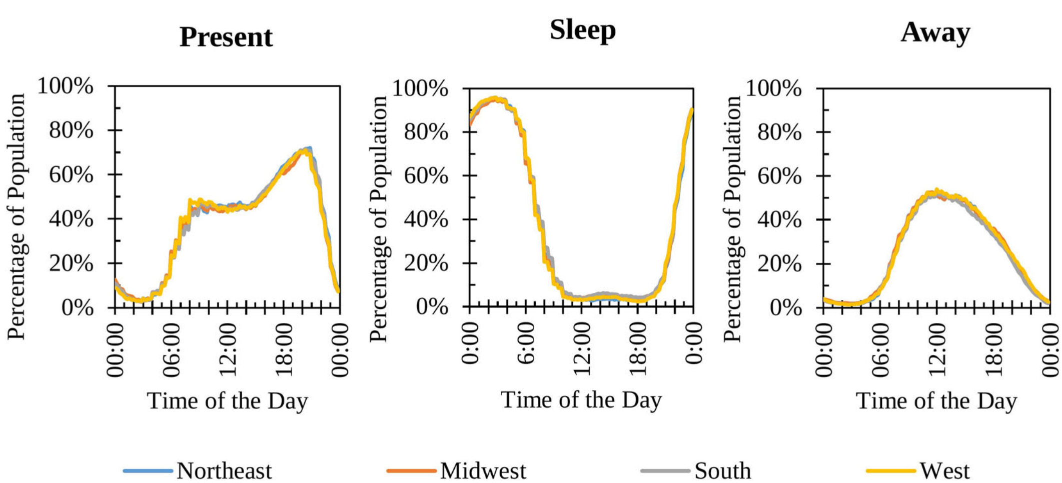

Fig. 10. Presence Status of ATUS respondents based on their housing geographic region.

Table 3. Household size based groups' distribution across the 2019 ATUS database.

Household size	Number of cases	% of total
1	2575	27.29%
2	2972	31.50%
3	1428	15.14%
4	1463	15.51%
5 or More	997	10.57%

Table 4. Day type based subgroups' distribution across the 2019 ATUS database.

	Weekday		Saturday		Sunday	
Household size	#	%	#	%	#	%
1	1247	13.22%	657	6.96%	671	7.11%
2	1530	16.22%	666	7.06%	776	8.22%
3	704	7.46%	345	3.66%	379	4.02%
4	698	7.40%	378	4.01%	387	4.10%
5 or More	463	4.91%	235	2.49%	299	3.17%

chance of a respondent from a three-person household being present and active at home at 00:00 on a weekday night was set to be 8.81% ($62/704 = 0.09$). Accordingly, in each subgroup, the probability inputs are calculated (Equation 1), where i can be "A," "S" or "P." These sets of calculations were repeated for all the subgroups and then organized into corresponding start time probability matrices (see the Appendix).

$$P_i = \frac{\#\text{ of cases where start state } = i \text{ in the subgroup}}{\#\text{ of cases in the subgroup}} \tag{1}$$

Step 3-2: Developing the transition probability matrices

The first-order Markov chain technique assumes each state is dependent only on the previous state together with the probabilities of that state changing. This set of probabilities is held in "transition probability matrices" and are directly derived from the ATUS data. As such, in each subgroup, using Equation 2, 9 probability inputs are calculated for each of the 144 defined 10-minute time steps, where i and j can be the variables "A," "S" or "P":

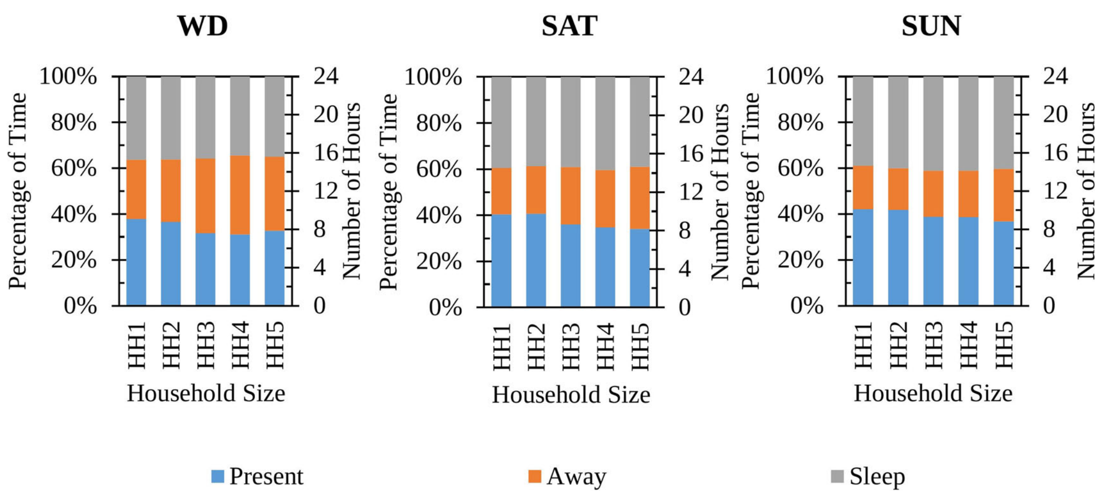

Fig. 11. State durations in the 2019 ATUS data based on household size and weekday type.

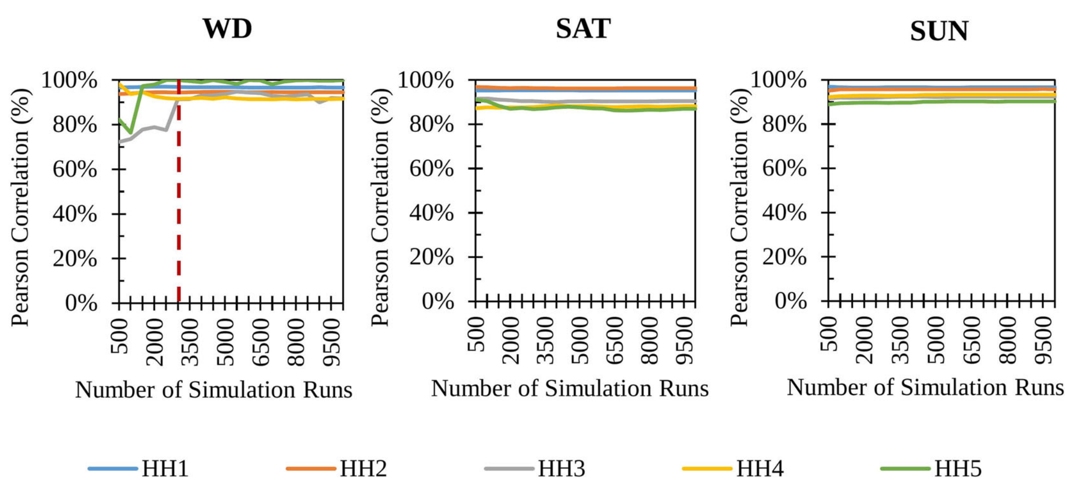

Fig. 12. Correlation between synthetic data and ATUS 2019 data in terms of state durations.

$$T_{i,j} = \frac{\#\ \text{of cases where start state} = i\ \&\ \text{end state} = j}{\#\ \text{of cases where start state} = i} \tag{2}$$

For example, for the three-person household weekday subgroup example, of all the 30 respondents absent in the house at 00:00, 1 reported that they were present and active in their home at 00:10, and 0 reported being present and sleep in their houses at 00:10. This means that T_{AP} for this subgroup was 3.33% ($1/30 = 0.03$), T_{AS} was 0% ($0/30 = 0$), and T_{AA} was 96.67% ($29/30 = 0.97$) at this timestep. These sets of calculations were repeated for all the subgroups and then organized into corresponding transition probability matrices (included in the supplemental data).

Step 3-3: Generating the occupancy presence schedules
In order to generate occupancy presence schedules, first, the start state was chosen by chosing a random number of present occupants from the appropriate probability distribution as calculated in Step 3-1. Subsequent states in the chain were determined by selecting a random number for each timestep and using this number with the appropriate transition probability matrix as defined in Step 3-2. The transition probability matrices and the start state distributions of this model (see the Appendix) were implemented in an R script that uses the *markovchian* package (Spedicato et al. 2015). In the following section, multiple example runs of the model are presented, validated, and discussed.

Results

State durations

Figure 11 provides the duration of each of the three states for all household sizes, according to their weekday type cluster. The sleep state duration generally does not change much by household size, while weekday type seems to play a larger role in the mean duration of hours in the sleep state in comparison. This is consistent with the findings of previous studies that have suggested that weekdays and weekend days do not usually share similar sleeping patterns (Basner, Spaeth, and Dinges 2014; Khajehzadeh 2017). On the other hand, the duration of present and away states seems to depend on both clustering criteria. Figure 11 generally shows that as the household size becomes larger, occupants tend to spend less time being

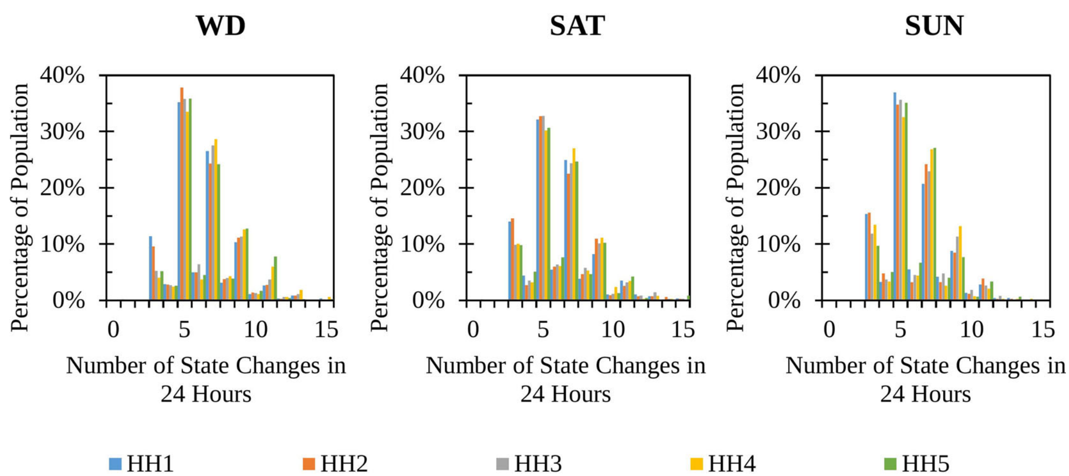

Fig. 13. Number of state changes in the 2019 ATUS data based on household size and weekday type *(Note: HHx = household size, where X is the # of people in the household)*.

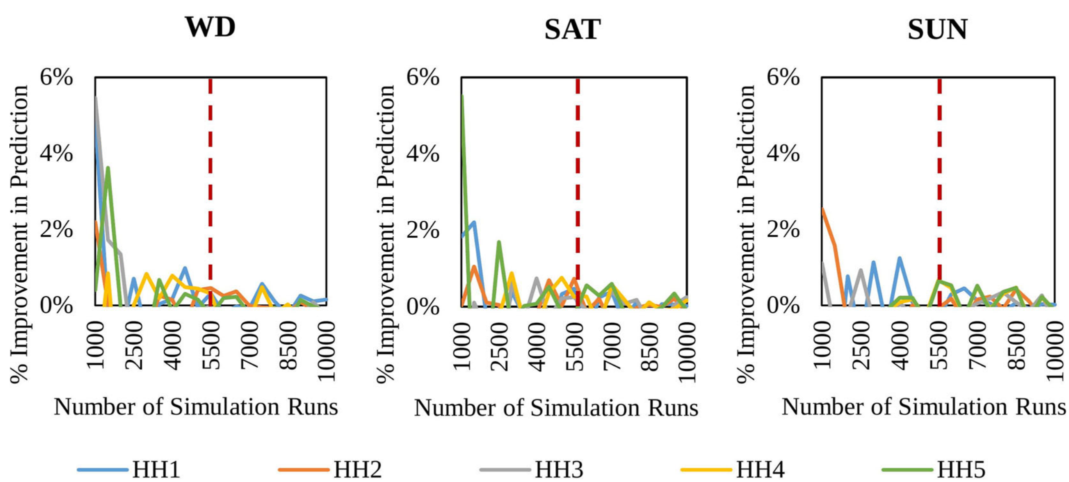

Fig. 14. Improvement in the Pearson Correlation values between synthetic data and ATUS 2019 data in terms of number of state changes *(Note: HHx = household size, where X is the # of people in the household)*.

present and active in their place of residence and spend more time being away instead. This is not surprising, given that the ATUS diaries are filed by a designated person (DP) from each household being represented that is at least 15 years of age (U.S. Census Bureau 2012). Therefore, younger household members that might spend more time at home are not included in the ATUS database.

Figure 12 compares the states duration of occupancy states for the synthetic and original data for all possible three states. In these figures, the Pearson Correlation values for mean state durations among different household sizes are compared between the original 2019 ATUS data and the synthetic data produced by the model. Overall, the synthetic data appears to capture the variations in durations shown by the original data with 3500 runs being the optimal number of simulation runs for the model.

Number of state changes

The principle aim of the model is to provide a basis for energy performance simulations and studies. Therefore, there is a particular need therefore to ensure the model accurately accounts for the proportion of time that dwellings are occupied or occupied by non-active occupants. Figure 13 shows the probability distribution of the number of times occupants in each household size cluster have changed their presence state in 24 hours. The data shown represent the means across original 2019 ATUS dataset, clustered by weekday type and household size. The majority of changes in presence state for all clusters occur in odd number during a 24-hour period which corresponds to waking up in the morning and going back to bed at night in addition to other activities throughout the day. Accordingly, those with 5 and 7 state changes were the most common type of schedules across all represented clusters.

As for the comparision of the number of state changes in the original ATUS data with synthethic data generated by the Markov chain model proposed, synthetic data appeared to capture the variations in durations shown by the original data relatively well (Pearson Correlation > 80% for 5000 runs per cluster). Moreover, it is suggested that a higher-order Markov chain model should be able to improve these

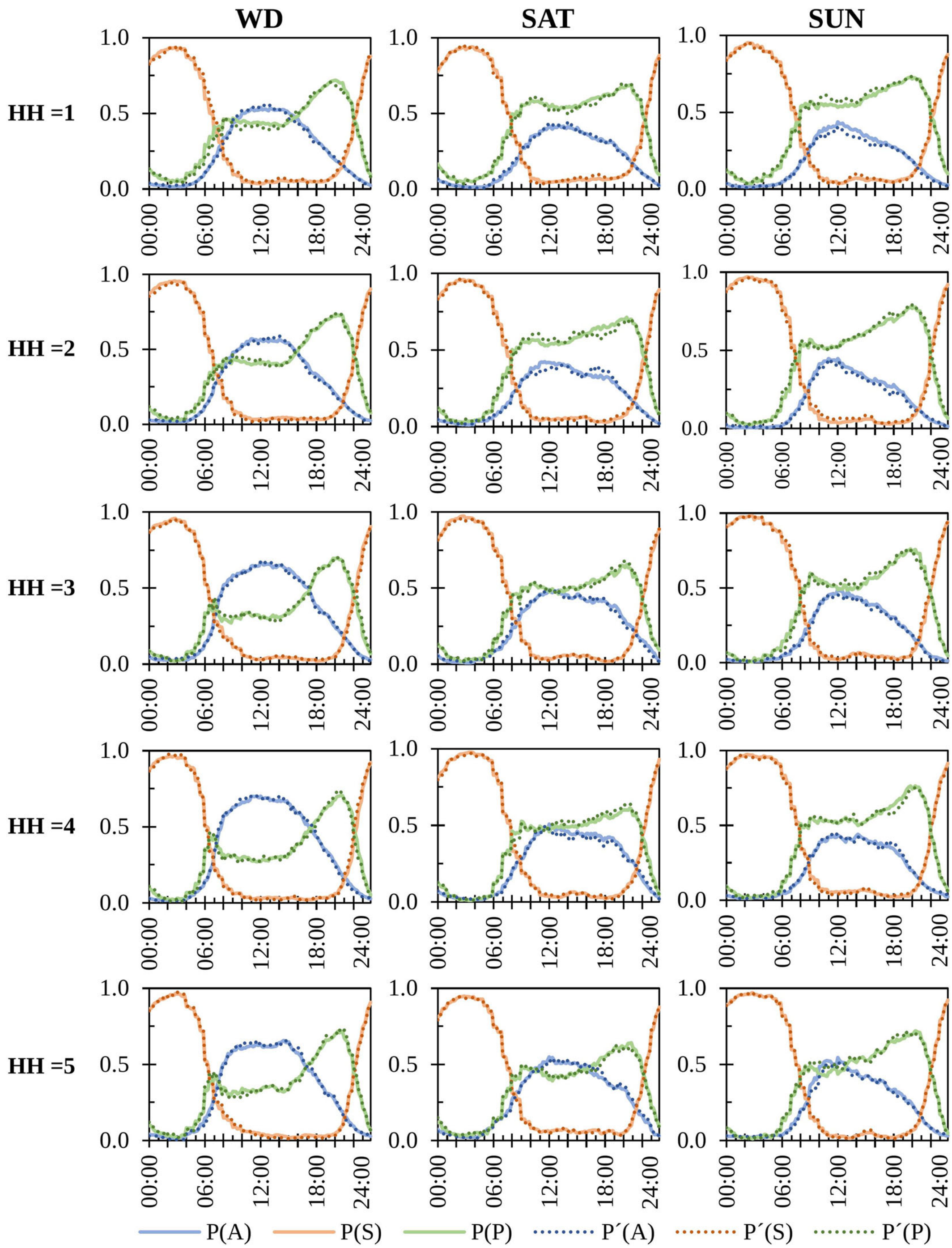

Fig. 15. Verification of state probabilities (y-axis) for different subgroups by comparison of the original ATUS data (P) with synthetic data (P′) (Note: HHx = household size, where X is the # of people in the household; P = present and active, A = absent, S = present and inactive).

correlation values in future studies. Figure 14 here shows the improvements in the model's predictions of the number of state changes based on the number of simulation runs performed. As shown, after 4500 runs, the improvements being made are minimal in all clusters (less than 1%). These improvements are measured by calculating the Pearson correlation for the mean number of state changes between the two dataset (synthetic and original data) for different number of simulation runs.

State probabilities

Figure 15 compares the state probabilities for differenet household size clusters for weekdays, Saturdays, and

Sundays for the time-use survey and the synthetic occupancy data generated by the model. The time-use survey data is based on the average across each cluster, while the synthetic data is based on 5,000 runs of the model for that cluster. Overall, the profiles show expected features including a low proportion of activity at night, a tendency for people to be out of their home during the day, and peaks in activity around meal times. The original and synthetic data shown in Figure 15 are generally in close agreement. The root-mean-square error in state probabilities has been checked for all combinations of number of residents and day types and all combinations show the same trend according to which errors are in the range of 1% to 5%. In general, therefore, it can be said the first-order Markov chain technique accurately reproduces the state probabilities found in the original data.

Conclusions

Occupancy schedules are often recognized as a leading source of uncertainty in building energy simulations and this uncertainty results in an undesired gap between actual performance and predictions. To address this issue and with the aim of increasing the reliability and accuracy of occupancy behavioral inputs of building energy models, in the recent years, many stochastic building occupancy models have been proposed as a replacement for the conventional generic static schedules. Such stochastic models commonly use time-use surveys (TUS), which are large nationally representative surveys of how people use their time, as their input. This paper describes the development of a three-state stochastic occupancy model based on the 2019 American Time-Use Survey (ATUS) data that takes account of differences between weekdays, Saturdays, and Sundays. In this stochastic occupancy model, survey respondents are clustered based on the number of residents in their households. A first-order inhomogeneous Markov chain technique is used to generate occupancy presence schedules such that it has the same overall statistics as the original ATUS data, notably in terms of state probabilities and state durations. The high-resolution representative occupancy data that this model generates can be used as input to any residential energy modeling tool that uses occupancy time-series as a base variable. The main advantages of this model compared to those that have been previously proposed are related to the simplicity and practicality of the model. The availability of the required inputs as well as the compatibility of the resulting outputs' formatting for implementation in conventional energy simulation tools make it optimal for use in future building energy modeling efforts instead of the commonly used static generic schedules proposed by standards and guidelines.

One limitation of this study is that ATUS data was utilized which is self-reported data and human error can influence self-reported data. Several recent papers have discussed the comparison between other nationwide datasets such as the Residential Energy Consumption Survey (RECS) data and ATUS data (Mitra et al. 2019, 2020). Thus, to further validate these findings related to occupancy schedules, it would be beneficial to compare the results with the Residential Energy Consumption Survey data or other national level datasets. However, due to unavailability of occupancy scheduling data in this and other datasets, the comparison cannot be completed at this time. Preliminary studies of the incorporation of the developed model's outputs in building energy simulation has shown promising results in terms of practicality and ease of use(Malekpour Koupaei et al. 2019a, 2019b, Malekpour Koupaei, Geraudin, and Passe 2020; Passe et al. 2020). Future work will include the optimization of this model for compatibility with co-simulation tools such as the Building Controls Virtual Test Bed (BCVTB). This will allow users to evaluate the impact of the developed occupancy schedules on the energy performance of residential buildings.

Nomenclature

HH_x = household size
P_i = probability of occupants' start time being *I in the original ATUS data*
P'_i = probability of occupants' start time being *I in the synthetic data*
$T_{i,j}$ = transition probability between state *i* and state *j*

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was partially funded by the 2016 Iowa State University Presidential Interdisciplinary Research Initiative (PIRI) on Data-Driven Science with additional support from the Iowa Energy Center grant "20-IEC-004."

ORCID

Diba Malekpour Koupaei http://orcid.org/0000-0002-2760-1792
Kristen S. Cetin http://orcid.org/0000-0003-2662-8480
Ulrike Passe http://orcid.org/0000-0003-2242-3082

References

Abu-Sharkh, S., R. Li, T. Markvart, N. Ross, P. Wilson, R. Yao, K. Steemers, J. Kohler, and R. Arnold. 2005. *Microgrids: Distributed on-site generation*. Manchester, UK: Tyndall Centre for Climate Change Research.

Adamopoulou, A. A., A. M. Tryferidis, and D. K. Tzovaras. 2016. A context-aware method for building occupancy prediction. *Energy and Buildings* 110:229–44. doi:10.1016/j.enbuild.2015.10.003

Agthe, D. E., and R. B. Billings. 2002. Water price influence on apartment complex water use. *Journal of Water Resources*

Planning and Management 128 (5):366–9. doi:10.1061/(ASCE)0733-9496(2002)128:5(366)

Basner, M., A. M. Spaeth, and D. F. Dinges. 2014. Sociodemographic characteristics and waking activities and their role in the timing and duration of sleep. *Sleep* 37 (12):1889–906. doi:10.5665/sleep.4238

Brémaud, P. 2013. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Vol. 31. Cham, Switzerland: Springer Science & Business Media.

Centre for Time Use Research. 2017. *United Kingdom Time Use Survey (UKTUS), 2014–2015*.

Diao, L., Y. Sun, Z. Chen, and J. Chen. 2017. Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. *Energy and Buildings* 147:47–66. doi:10.1016/j.enbuild.2017.04.072

Flett, G., and N. Kelly. 2016. An occupant-differentiated, higher-order Markov Chain method for prediction of domestic occupancy. *Energy and Buildings* 125:219–30. doi:10.1016/j.enbuild.2016.05.015

Grandjean, A., J. Adnot, and G. Binet. 2012. A review and an analysis of the residential electric load curve models. *Renewable and Sustainable Energy Reviews* 16 (9):6539–65. doi:10.1016/j.rser.2012.08.013

Happle, G., J. A. Fonseca, and A. Schlueter. 2018. A review on occupant behavior in urban building energy models. *Energy and Buildings* 174:276–92. doi:10.1016/j.enbuild.2018.06.030

Hoes, P., J. L. M. Hensen, M. Loomans, B. de Vries, and D. Bourgeois. 2009. User behavior in whole building simulation. *Energy and Buildings* 41 (3):295–302. doi:10.1016/j.enbuild.2008.09.008

Hong, T., S. C. Taylor-Lange, S. D'Oca, D. Yan, and S. P. Corgnati. 2016. Advances in research and applications of energy-related occupant behavior in buildings. *Energy and Buildings* 116: 694–702. doi:10.1016/j.enbuild.2015.11.052

Hong, T., D. Yan, S. D'Oca, and C-f. Chen. 2017. Ten questions concerning occupant behavior in buildings: The big picture. *Building and Environment* 114:518–30. doi:10.1016/j.buildenv.2016.12.006

Khajehzadeh, I. 2017. An investigation of the effects of large houses on occupant behaviour and resource-use in New Zealand.

Mahdavi, A., and F. Tahmasebi. 2016. The deployment-dependence of occupancy-related models in building performance simulation. *Energy and Buildings* 117:313–20. doi:10.1016/j.enbuild.2015.09.065

Malekpour Koupaei, D., M. Geraudin, and U. Passe. 2020. *Lifetime energy performance of residential buildings: A sensitivity analysis of energy modeling parameters*. The Symposium on Simulation for Architecture and Urban Design (SimAUD), 585–91.

Malekpour Koupaei, D., F. Hashemi, V. Tabard-Fortecoëf, and U. Passe. 2019a. A technique for developing high-resolution residential occupancy schedules for urban energy models. 2019 Iowa State Research Day.

Malekpour Koupaei, D. F. Hashemi, V. Tabard-Fortecoëf, and U. Passe. 2019b. *Development of a modeling framework for refined residential occupancy schedules in an urban energy model*. Building Simulation 2019: 16th Conference of IBPSA. doi:10.26868/25222708.2019.210553

Mitra, D., Y. Chu, N. Steinmetz, P. E. Kristen Cetin PhD, P. Kremer, and J. Lovejoy. 2019. Defining typical occupancy schedules and behaviors in residential buildings using the American Time Use Survey. *ASHRAE Transactions* 125:382–90.

Mitra, D., N. Steinmetz, Y. Chu, and K. S. Cetin. 2020. Typical occupancy profiles and behaviors in residential buildings in the United States. *Energy and Buildings* 210:109713. doi:10.1016/j.enbuild.2019.109713

Nijhuis, M., M. Gibescu, and J. F. G. Cobben. 2016. Bottom-up Markov Chain Monte Carlo approach for scenario based residential load modelling with publicly available data. *Energy and Buildings* 112:121–9. doi:10.1016/j.enbuild.2015.12.004

Osman, M., and M. Ouf. 2021. A comprehensive review of time use surveys in modelling occupant presence and behavior: Data, methods, and applications. *Building and Environment* 196:107785. doi:10.1016/j.buildenv.2021.107785

Passe, U., M. Dorneich, C. Krejci, D. Malekpour Koupaei, B. Marmur, L. Shenk, J. Stonewall, J. Thompson, and Y. Zhou. 2020. An urban modelling framework for climate resilience in low-resource neighbourhoods. *Buildings and Cities* 1 (1):453–74. doi:10.5334/bc.17

Ramírez-Mendiola, J. L., P. Grünewald, and N. Eyre. 2019. Residential activity pattern modelling through stochastic chains of variable memory length. *Applied Energy* 237:417–30. doi:10.1016/j.apenergy.2019.01.019

Richardson, I., M. Thomson, and D. Infield. 2008. A high-resolution domestic building occupancy model for energy demand simulations. *Energy and Buildings* 40 (8):1560–6. doi:10.1016/j.enbuild.2008.02.006

Spedicato, G. A., T. S. Kang, S. Bhargav, and M. G. A. Spedicato. 2015. *Package 'markovchain'*. https://cran.r-project.org/web/packages/markovchain/markovchain.pdf

Swan, L. G., and V. I. Ugursal. 2009. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews* 13 (8): 1819–35. doi:10.1016/j.rser.2008.09.033

The French National Institute of Statistics and Economic Studies (INSEE). 2010. *The 2009–2010 French Time Use Survey*.

The National Oceanic and Atmospheric Administration (NOAA). 2021. *Meteorological versus astronomical seasons*. https://www.ncei.noaa.gov/news/meteorological-versus-astronomical-seasons

Torriti, J. 2014. A review of time use models of residential electricity demand. *Renewable and Sustainable Energy Reviews* 37:265–72. doi:10.1016/j.rser.2014.05.034

U.S. Bureau of Labor Statistics (BLS). 2007. *American Time Use Survey Questionnaire*. https://www.bls.gov/tus/

U.S. Bureau of Labor Statistics (BLS). 2019. *American Time Use Survey (ATUS) data dictionary: 2019 interview data*.

U.S. Bureau of Labor Statistics (BLS). 2020. *American Time Use Survey (ATUS) Data Dictionary: 2019 Interview Data*. https://www.bls.gov/tus/atusintcodebk19.pdf

U.S. Census Bureau. 2012. *American Time Use Survey user's guide: Understanding ATUS 2003 to 2011*.

U.S. Census Bureau. 2021. *Metropolitan and micropolitan*. https://www.census.gov/programs-surveys/metro-micro/guidance.html

U.S. Energy Information Administration (EIA). 2019. *Annual Energy Outlook 2019 (with projections to 2050)*.

U.S. Energy Information Administration (EIA). 2021. Monthly energy review. https://www.eia.gov/totalenergy/data/monthly/

Yan, D., T. Hong, B. Dong, A. Mahdavi, S. D'Oca, I. Gaetani, and X. Feng. 2017. IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings. *Energy and Buildings* 156: 258–70. doi:10.1016/j.enbuild.2017.09.084

Yan, D., W. O'Brien, T. Hong, X. Feng, H. B. Gunay, F. Tahmasebi, and A. Mahdavi. 2015. Occupant behavior modeling for building performance simulation: Current state and future challenges. *Energy and Buildings* 107:264–78. doi:10.1016/j.enbuild.2015.08.032

Yang, J., M. Santamouris, and S. E. Lee. 2016. Review of occupancy sensing systems and occupancy modeling methodologies for the application in institutional buildings. *Energy and Buildings* 121: 344–9. doi:10.1016/j.enbuild.2015.12.019

Yoshino, H., T. Hong, and N. Nord. 2017. IEA EBC annex 53: Total energy use in buildings—Analysis and evaluation methods. *Energy and Buildings* 152:124–36. doi:10.1016/j.enbuild.2017.07.038

Zhang, Y., X. Bai, F. P. Mills, and J. C. v. Pezzey. 2018. Rethinking the role of occupant behavior in building energy performance: A review. *Energy and Buildings* 172:279–94. doi:10.1016/j.enbuild.2018.05.017

Zhou, Y., Z. J. Yu, J. Li, Y. Huang, and G. Zhang. 2017. The effect of temporal resolution on the accuracy of predicting building occupant behaviour based on Markov Chain Models. *Procedia Engineering* 205:1698–704. doi:10.1016/j.proeng.2017.10.359

Appendix

Tables A1–A3 below include start time probabilities for different household size clusters on weekdays, Saturdays, and Sundays. It should be noted that the presented values in this section are rounded up to two decimal values for readability and thus, in some cases might not add up to be 100% exactly.

Table A1. Start time probabilities of different household sizes in weekdays based on the ATUS 2019 data.

	Household size				
	1	2	3	4	5 or more
P(A)	3.29%	2.61%	4.26%	3.01%	4.10%
P(S)	83.48%	87.52%	86.93%	87.25%	86.39%
P(P)	13.23%	9.87%	8.81%	9.74%	9.50%

Table A2. Start time probabilities of different household sizes on Saturdays based on the ATUS 2019 data.

	Household size				
	1	2	3	4	5 or more
P(A)	5.94%	4.35%	6.09%	6.61%	7.23%
P(S)	77.47%	83.48%	81.74%	81.48%	81.28%
P(P)	16.59%	12.16%	12.17%	11.90%	11.49%

Table A3. Start time probabilities of different household sizes on Sundays based on the ATUS 2019 data.

	Household size				
	1	2	3	4	5 or more
P(A)	3.43%	1.68%	2.37%	2.84%	3.01%
P(S)	84.95%	88.40%	91.29%	87.60%	88.96%
P(P)	11.62%	9.92%	6.33%	9.56%	8.03%