Received February 17, 2021, accepted February 24, 2021, date of publication March 2, 2021, date of current version March 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063502

A Probabilistic Model to Predict Household Occupancy Profiles for Home Energy Management Applications

LUIS RUEDA[1], SIMON SANSREGRET[2], BRICE LE LOSTEC[2],
KODJO AGBOSSOU[1], (Senior Member, IEEE), NILSON HENAO[1],
SOUSSO KELOUWANI[3], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, Hydrogen Research Institute, University of Quebec at Trois-Rivières, Trois-Rivières, QC G8Z 4M3, Canada
[2]Laboratoire des Technologies de l'Énergie, Institut de Recherche Hydro-Québec, Shawinigan, QC G9N 7N5, Canada
[3]Department of Mechanical Engineering, Hydrogen Research Institute, University of Quebec at Trois-Rivières, Trois-Rivières, QC G8Z 4M3, Canada

Corresponding author: Luis Rueda (luis.rueda@uqtr.ca)

ABSTRACT Due to the impact of human lifestyle on building energy consumption, the development of occupants' behavior models is crucial for energy-saving purposes. In this regard, occupancy modeling is an effective approach to intend such a purpose. However, the literature reveals that existing occupancy models have limitations related to the representation of occupancy state duration and the integration of occupancy variability among individuals. Accordingly, this paper proposes an explicit differentiated duration probabilistic model to generate realistic daily occupancy profiles in residential buildings. The discrete-time Markov chain theory and the semi-parametric Cox proportional hazards model (Cox regression) are used to predict household occupancy profiles. The proposed model is able to capture occupancy states duration and integrate human behavior variability according to individuals' characteristics. Moreover, a parametric analysis is employed to investigate these characteristics' impact on the model performance and consequently, select the most significant input variables. A validation process is conducted by comparing the model performance with that of previous methods, presented in the literature. For this purpose, the k cross-validation technique is utilized. Validation results demonstrate that the proposed approach is highly efficient in generating realistic household occupancy profiles.

INDEX TERMS Occupancy, behavior, survival analysis, hazard rate, markov-chain, Cox regression.

I. INTRODUCTION

The electric grid faces a transformation in terms of growth and development due to global warming, huge electricity demand, and the impending depletion of fossil energy resources. As a result, the smart grid is promoted as a conceptual framework for improving energy efficiency and advanced management of available resources [1]. The building sector accounts for more than 30% of total world energy consumption that is expected to increase by an average of 1.5% per year between 2012-2040 [2]. Consequently, this sector receives significant attention particularly due to its energy-saving potentials, which can be up to 30% [3].

Energy consumption in buildings is influenced by different factors, including climate, building envelope, building services and energy systems, indoor environment quality, building operation and maintenance, and occupants' behavior [4]. Among these factors, human behavior can be considered as a major issue towards enhancing the performance, design, and simulation of buildings [3]. Occupant behavior refers to the interaction between occupants and buildings in order to preserve a healthy indoor environment and acquire desired comfort, and security [5], [6]. It can be described by occupants presence, and their interaction with appliances, control systems (e.g., heating, ventilation, and air conditioning), and building elements (e.g., doors and windows) [7].

In the last decades, occupant behavior modeling, occupancy detection, and activity recognition have attracted significant attention among researchers whose focus is building

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Anvari-Moghaddam.

design and operation [7]. These models can be used for applications like building performance simulation and load forecasting, which can help utilities anticipate peak power demand and make decisions on power management, load switching, network reconfiguration, and infrastructure development [8], [9]. Generally, human behavior is dealt with as a deterministic variable, for example occupancy static schedules [5]. Nonetheless, occupant behavior has a stochastic nature and it is influenced by different factors such as weather conditions, physical characteristics of the house, and socio-demographic characteristics of individuals. For this reason, the development of probabilistic occupant behavior models has attracted researchers' attention to integrating variability of individuals' behavior patterns.

Moreover, a better understanding of human-building interactions can improve building energy performance and reduce its energy consumption while maintaining occupants' comfort [10]. Indeed, several works have demonstrated that a significant amount of energy can be saved by performing occupancy-based and activity-based control strategies. For instance, Nguyen *et al.* [11] surveyed intelligent buildings research efforts, focused on energy saving and user activity recognition. The authors concluded that occupancy-based control strategies can result in up to 40% energy savings in HVAC system. Georgievski *et al.* [12] presented an approach based on Hierarchical Task Network (HTN) planning combined with activity recognition. Their proposed method, applied to control lights and workstations in an office building and a restaurant, provided up to 80% energy savings. Layered hidden Markov models (LHMMs) and deterministic finite state machines (FSMs) were used in [13] to identify desk-related activities and count people in an office room. The authors showed that by using activity-based control over lighting systems, it was possible to achieve overall energy savings of 78.5% per year compared to manual control. Additionally, by using a people count-based control strategy, it was possible to reduce annual HVAC energy consumption by 39.9%. Scott *et al.* [14] used PIR and RFID sensors' information to predict home occupancy patterns for heating control. The authors analyzed three different control techniques (based on scheduled, always-on, and PreHeat algorithms) and consequently, obtained energy savings of up to 35%. Boait and Rylatt [15] presented a prototype to control the heating system in a house in the UK. The prototype utilized Bayesian inference to identify the occupancy state based on electricity consumption data and hot water temperature. The authors achieved up to 14.1% energy savings in the gas, consumed by the boiler (19.2 kWh/day). Likewise, heterogeneous occupancy patterns were used to evaluate a simulation-based optimization approach for the design of energy management systems (EMS), applied to microgrids [16], [17]. Moreover, a zonal control strategy of thermostatic loads was presented in [18]. The authors demonstrated that their strategy could save more than 15% energy consumption while increasing the thermal comfort by more than 25%.

A. BACKGROUND

A comprehensive understanding of individuals' behavior inside and outside of the house is of great importance for sectors such as energy [2] and transportation [19]. Literature reveals that Time-use survey (TUS) and Travel survey (TS) are important sources of information to achieve such comprehension [20], [21]. Particularly, TUS data has been widely used for the development of probabilistic occupant behavior models, which primarily focus on individuals in the residential sector [22]. It can provide detailed information about the daily routine of a large number of respondents and capture the stochastic nature of occupant behavior and its variability according to individuals' characteristics. Therefore, TUS is suitable for representing a population [5]. Moreover, the literature reveals that independent from the country in which the survey is conducted, TUS provides a similar set of information about respondents, for instance their age, gender, marital status, and income. Accordingly, it allows the development of highly replicable methods regardless of the survey location. Walker and Pokoski [23] proposed a model of residential demand based on customer behavior, captured from surveys. The authors were one of the first to use TUS data to model human behavior. They proposed an "availability" and a "proclivity" function to estimate the number of people in a household and their probability of doing a certain activity at specific time periods. However, one of the main drawbacks of this model, which limits its application scope, is its insufficient flexibility to analyze behavioral variability related to users' socio-demographic characteristics.

Among the techniques, used to model human behavior by using TUS data, first-order Markov chains (FOMC) and higher-order Markov chains (HOMC) are the most popular ones. A first-order MC is a memoryless stochastic process according to its Markovian property. Therefore, the transition probability between states only depends on the current state. Richardson *et al.* [24] used this technique to generate active occupancy[1] daily profiles at a ten-minute resolution based on the United Kingdom 2000 time-use survey [25]. The approach was able to indicate the number of active occupants in the house. However, only the number of household residents and the weekdays were used as parameters to characterize occupancy variability. Furthermore, Richardson *et al.* extended their occupancy model in another study [26] to generate activity profiles in order to develop a domestic electricity demand model. Widen *et al.* [27] proposed a first-order Markov Chain Monte Carlo technique to generate domestic lighting demand data from occupancy patterns and daylight availability. The authors presented a simplified model in which all persons were characterized by four parameters related to dwelling (detached houses and apartments) and day type (weekdays and weekends). Subsequently, in [28], they refined their initial model to generate occupant activity

[1] Active occupancy refers to a moment when a user is in the house and performs an activity other than sleeping.

sequences and use them to estimate domestic electricity demand.

Likewise, Muratori *et al.* [29] presented a heterogeneous Markov chain to generate activity patterns. The authors used a method similar to the one presented by Widen and Wackelgard [28]. In addition, they included a small set of parameters to assess behavior differences among individuals, represented by working, non-working, male, female, and children classes. Collin *et al.* [30] developed a thirteen-state first-order Markov model, to create activity profiles, differentiated by household size as well as working and non-working individuals. Afterwards, the authors applied the model to generate demand profiles according to user activities. They employed a probabilistic function to identify electrical appliances, shared by users performing the same activity. Baptista *et al.* [31] proposed the utilization of the interactive Markov chain approach, presented by Conlisk, [32] to incorporate interaction between individuals in a household. For this purpose, the transition probability matrices were conditioned by the present activity of a leader. In this way, occupants were able to explicitly coordinate their activities. However, this approach did not consider a comprehensive analysis of behavior variability according to individuals' characteristics. The articles, presented above, emphasize that behavior variability has not been adequately explored. Moreover, they share the disadvantage of a memoryless property, yielded by the first-order Markov models. [33], [34] has revealed that first-order Markov models are not able to consistently predict state duration and thus, their generated occupancy or activity profiles often fail to reflect a realistic behavior.

Alternatively, some authors have proposed the use of higher-order Markov models. These probabilistic models do not possess the memoryless property of traditional Markov chain models. Therefore, the duration is not required to assume exponential or geometric distributions over continuous or discrete-time cases, respectively. Accordingly, Tanimoto [35], Wilke *et al.* [33], [36], and Vorger [37] captured occupants' behavior from a semi-Markovian process. In this process, the states are conditioned by their precedents through a Markovian transition, and the duration is conditioned by only the current state. In fact, the procedure is based on a decremental counter that forces a state transition every time it reaches "1". Once a new state is determined, the counter initial value is estimated by a duration distribution. For such an estimation, Tanimoto used a logarithmic distribution to generate the duration of activities. Wilke *et al.* and Vorger used a Weibull distribution to determine presence and activity duration. However, these distributions put limitations on capturing the multimodal character that often occurs in duration data.

Furthermore, Flett *et al.* [34], [38] presented a model to produce occupancy profiles. They evaluated the impact of relationships between cohabiting individuals on the overall active occupancy probability. The authors developed a higher-order Markov method where multiple probability transition matrices were generated according to the existing state duration. Aerts *et al.* [39], [40] presented a methodology to identify 'typical occupancy patterns' from Belgian time-use data by using hierarchical clustering. These patterns were used to calibrate a three-state probabilistic model (absent, asleep, and at home or awake) to generate individual daily and yearly occupancy sequences. However, none of the above methods include a detailed set of parameters to assess behavior differences. This, in turn, has limited the analysis of individuals characteristics to age, gender, income and employment status. Consequently, it can be deduced that although some higher-order Markov-based methods have been proposed, there are still lacks in the representation of state duration, mainly related to occupant-specific behavior variations and multimodal duration distributions generation.

In addition, existing occupant behavior models have other limitations that must be addressed to enable a more reliable residential occupancy and activity prediction. For the simulation of behavior in multi-person households, most of the approaches have treated users as independent agents, which does not reflect actual human behavior in dwellings. In real life, occupants interact and thus, they can perform activities together. Accuracy in modeling such forms of interaction can significantly impact the simulation of load curves since energy consumption changes according to whether or not occupants share appliances. It should be noted that seasonal effects and long absence periods due to holidays and illness are additional concerns with occupancy modeling that have not been adequately studied. This can be partially associated with limited information available in the surveys. Generally, surveys supply daily data of respondents and skip information of all inhabitants and longer periods.

B. CONTRIBUTION

This paper aims to address the aforementioned issues related to designing a model capable of describing occupancy state duration while incorporating occupancy variability among individuals. Literature shows that the hazard-based model is a suitable approach for dealing with these concerns since it is able to capture duration-dependent transition probabilities and incorporate exogenous variables [41]. In fact, this method has proved its potential for applications such as windows opening and closing behavior modeling [42], activity-travel modeling, as well as in-home and out-of-home activity generation [19]. Accordingly, we present an explicit-duration probabilistic model that combines the semi-parametric Cox proportional hazards model (Cox regression) and the discrete-time Markov chain theory in order to effectively handle the challenges, faced in this study. Cox regression is a technique that at each time step estimates a state chances of making transition according to its duration up to that time [43]. This method is normally used for survival analysis. Moreover, unlike parametric models, the proportional hazards approach does not restrict the hazard function shape to a particular distribution. Furthermore, this method allows the hazard function to consider several explanatory variables and thus, enables the model to include behavior variability.

Additionally, regarding the diversity of individuals' characteristics, a parametric analysis is performed to assess their impact on the model capability to represent occupancy variability. Besides, a variance-based analysis and a backward elimination technique are used to identify the most statistically significant parameters of the proposed model. Moreover, an improvement of the approach, proposed by Wilke *et al.*, [33] is presented. This improvement seeks to overcome the limitations of the Weibull distribution to fit the multimodal characteristic of duration data. In order to achieve this ambition, we utilize a Gaussian mixture model (GMM) along with a silhouette analysis. Hence, it is possible to automatically identify the number of clusters of each duration distribution in order to represent them as normally distributed sub-populations.

C. OUTLINE

In Section II, the time-use survey data, employed to calibrate the model, is described. Section III presents the probabilistic model, proposed to predict realistic occupancy profiles. Section IV explains the utilized parametric analysis. Section V describes the validation process and benchmarks the model performance with other approaches, proposed in the literature. In Section VI, the results and future prospects are discussed. Finally, the conclusions are presented in Section VII.

II. TIME-USE SURVEY DATA

This paper focuses on domestic occupancy modeling by using national time-use survey (TUS) data. In particular, the calibration process of the proposed model is performed based on the Canada TUS data from April 2015 to April 2016, supplied by Statistics Canada [44]. This survey comprises a sample size of 15,390 respondents of a target population that includes all persons, aged 15 years old or more in Canada. Each individual diary provides a detailed record of a wide variety of daily activities and their devoted time, location, and other participants (except for the respondent). In this survey, the diary starts at 4:00 AM with a list of 266 activities. The survey presents information of up to three simultaneous activities. In addition, it provides information about perceptions of time, unpaid work periods, well-being, paid work and education time, cultural and sports activities, transportation, as well as numerous socio-demographic characteristics.

A. OCCUPANCY PATTERNS DIFFERENTIATION

In order to represent variability and stochastic nature of the occupancy, a set of characteristics, defined by $\boldsymbol{x} = \{x_1, x_2, \ldots, x_M\}$ are used to describes each individual. As presented in TABLE 1, these characteristics include personal information about the respondent (including age, occupation, and marital status), the family composition, and the household.

Individuals' characteristics enable the model to capture the occupancy variability of populations with different socio-demographic features. As illustrated in FIGURE 1, the probability of an individual's presence at home (notwithstanding the activity he/she performs) is different regarding the targeted sub-population. Furthermore, the time-varying nature of behavior is observed according to the hour of the day, the weekday, and the month of the year.

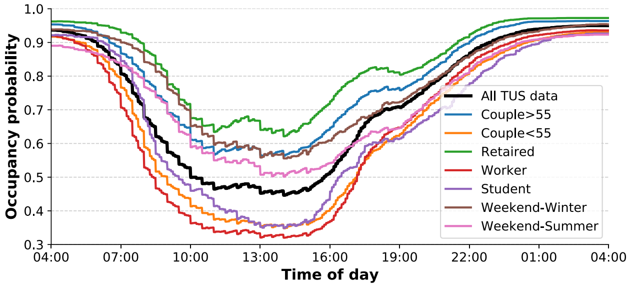


FIGURE 1. Occupancy patterns differentiation.

TABLE 1. Individuals characteristics.

Name	Description
x_1	Weekday
x_2	Type of day
x_3	Month
x_4	Sex of respondent
x_5	Age of respondent
x_6	Marital status of respondent
x_7	Respondent has a spouse/partner living in the household
x_8	Age difference between respondent and spouse/partner
x_9	Sex of respondent's spouse/partner living in the household
x_{10}	Type of partner respondent has living in the household
x_{11}	Child(ren) of the respondent living in the household
x_{12}	Age of respondent's youngest child in household
x_{13}	Age of youngest household member in respondent's household
x_{14}	Number of respondent's child(ren) in household
x_{15}	Number of respondent's child(ren) in household - 0 to 14 years
x_{16}	Age group of respondent's child(ren) in household
x_{17}	Number of respondent's parents in the household
x_{18}	Living arrangement of respondent's household (8 categories)
x_{19}	Living arrangement of respondent's household (11 categories)
x_{20}	Household size of respondent
x_{21}	Number of members in respondent's household ($\geq$ 15 years)
x_{22}	Three generations or more in the respondent's household
x_{23}	Main activity - Last 12 months
x_{24}	Main activity of spouse/partner - 12 months
x_{25}	Dwelling type of the respondent's household
x_{26}	Educational attainment - Highest degree
x_{27}	Income - Personal income group (before tax)
x_{28}	Household income - Household income group (before tax)

B. CENSORED DATA

Since the duration is taken from one-day diaries, information about the episodes that last until the end of the diary day (4 AM) is incomplete. This can be due to the fact that these episodes probably last for additional time and thus, their exact duration is unknown (right-censored observations). Accordingly, these events, considered as censored date, are identified and included in the estimation of the transition probabilities.

Different statistical methods such as complete-data analysis, imputation approach, and dichotomized data-bases analysis are commonly used to deal with censored data [45]. However, likelihood-based strategies like Kaplan-Meier estimator and Cox-regression have been promoted as effective methods that can handle all available information, whether it presents censored data.

III. MODELING APPROACH

The proposed approach to modeling residential occupant behavior consists of a non-homogeneous semi-Markov process (NHSMP)[2] to predict individuals' occupancy profiles according to temporal information (e.g., hour, weekday, and month) and socio-demographic characteristics.

A. EXPLICIT-DURATION PROBABILISTIC MODEL

Individual occupancy sequences are generated through a two-state probabilistic model that produces daily profiles with a time-resolution of 10 min. Occupancy profile is represented by the vector $\vec{\mathbf{z}} = [z_1, z_2, \ldots, z_N]$ with a length of $N = 144$. This profile can be represented by a finite-state machine (FSM) with two possible states $s = \{0, 1\}$. In discrete-time $k = 1, 2, \ldots, N$, the occupancy state, z_k is zero when the machine is at the "absence" state and one when it remains in the "presence"state. Therefore, by using the chain rule, the probability of the sequence $\mathrm{P}(\vec{\mathbf{z}})$ can be described in terms of conditional probabilities, expressed by:

$$\mathrm{P}(\vec{\mathbf{z}}) = \mathrm{P}(z_1)\,\mathrm{P}(z_2|z_1)\,\mathrm{P}(z_3|z_2, z_1)\ldots\mathrm{P}(z_N|z_{N-1}, \ldots, z_1) \tag{1}$$

Since equation 1 covers occupancy profile dynamic, its calculation becomes a complicated challenge if the current state of the machine has a long dependency on its previous state. Consequently, the dimension of the conditional laws becomes too large when the sequence lengthens. However, the development of approximations and the adoption of hypotheses that reduce the history of states sequence can bring the problem to an affordable dimension. Accordingly, a hypothesis of medium-term independence is introduced in order to truncate the calcul of $\mathrm{P}(\vec{\mathbf{z}})$. Such hypothesis implies that the probability of changing a state depends on the current state and its duration, explained by,

$$\mathrm{P}(z_{k+1}|z_k, z_{k-1}, \ldots, z_2, z_1) = \mathrm{P}(z_{k+1}|z_k, d_k) \tag{2}$$

By applying this hypothesis, the sequence of states, $\vec{\mathbf{z}}$ can be represented as a Semi-Markov or explicit duration process $\mathrm{P}(z_{k+1}|z_k, d_k)$, in which the duration, d_k is considered as an incremental counter, restarted ($d_k = 1$) when a transition is made ($z_{k+1} \neq z_k$). In this study, the system has only two states. Therefore, when a transition must be made, the new state, z_{k+1} is selected in a deterministic manner. By applying Bayes' theorem and chain rule to the expression on the right side of the equation 2, the transition probability can be defined as,

$$\mathrm{P}(z_{k+1} \mid z_k, d_k) = \frac{\mathrm{P}(d_k \mid z_{k+1}, z_k)\,\mathrm{P}(z_{k+1} \mid z_k)}{\mathrm{P}(d_k \mid z_k)} \tag{3}$$

When estimating the transition probability, two situations can occur. These cases account for state change, expressed by equation (4), or state preserve, defined by equation (5), as below.

$$\mathrm{P}(z_{k+1} \neq s|z_k = s, d_k = d') \tag{4}$$

$$\mathrm{P}(z_{k+1} = s|z_k = s, d_k = d') \tag{5}$$

These two situations are mutually exclusive. As a result, by analyzing the circumstance in which a change in state takes place (according to the literature [46], [47]), the expressions on the right side of equation 3 can be reformulated as,

$$\mathrm{P}(d_k = d' \mid z_{k+1} \neq s, z_k = s) = f(d' \mid \theta_s) \tag{6}$$

$$\mathrm{P}(z_{k+1} \neq s \mid z_k = s) = \frac{1}{\mathbb{E}\left[d_k = d' \mid z_{k+1} \neq s, z_k = s\right]} \tag{7}$$

$$\mathrm{P}(d_k = d' \mid z_k = s) = \frac{S\left(d' \mid \theta_s\right)}{\mathbb{E}\left[d_k = d' \mid z_{k+1} \neq s, z_k = s\right]} \tag{8}$$

where $f(d_k = d' \mid \theta_s)$ is the state duration probability, $S\left(d_k = d' \mid \theta_s\right)$ is the survival function, associated with state s, $\mathbb{E}\left[d_k = d' \mid z_{k+1} \neq s, z_k = s\right]$ is the mathematical expectation of state duration distribution, and θ_s is a set of parameters, related to state duration distribution.

Consequently, by combining equations (3) and (6)-(8), the probability of transition between states can be described by,

$$\mathrm{P}(z_{k+1}|z_k = s, d_k = d') = \begin{cases} h(d' \mid \theta_s) & \text{if } z_{k+1} \neq s \\ 1 - h(d' \mid \theta_s) & \text{otherwise} \end{cases} \tag{9}$$

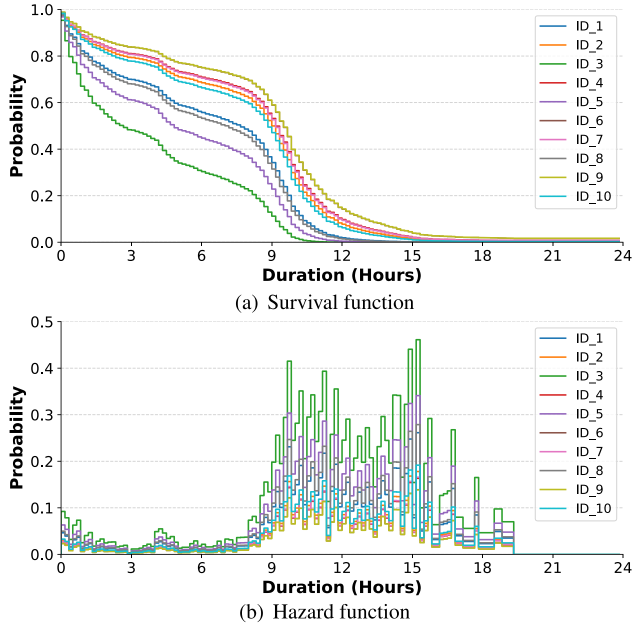
where $h(d' \mid \theta_s)$ is the discrete hazard function of state s, which can be expressed as,

$$h(d' \mid \theta_s) = \frac{f(d' \mid \theta_s)}{S(d' \mid \theta s)} = 1 - \frac{S(d' + 1 \mid \theta_s)}{S(d' \mid \theta s)} \tag{10}$$

Besides, time dependency is an important factor in human behavior modeling. For this reason, temporal information is included to the model by calculating the hazard function in hourly intervals $t \in [0, 1, \ldots, 23]$. Accordingly, the model parameters, θ are affected not only by the state s, but also by the time of day t. This results in a hazard function with the form $h(d' \mid \theta_s^t)$.

In order to include occupancy variability in the model, we must evaluate the impact of individuals' characteristics (covariates) on the hazard function. Accordingly, we use the Cox proportional hazard model (Cox regression). This semi-parametric method allows to estimate the relationship between the hazard rate and the explanatory variables without any assumptions about the shape of the baseline hazard

[2] A non-homogeneous semi-Markov process (NHSMP) is a generalization of Markov chains and the renewal process, where transition probability between states are time-varying and depend on sojourn times.



(a) Survival function

(b) Hazard function

FIGURE 2. Survival and hazard function variability.

function [43], based on,

$$h(d' \mid \boldsymbol{x}, \theta_s^t) = h(d' \mid \phi_s^{\boldsymbol{x},t}) = h_0(d' \mid t, s)e^{\boldsymbol{\beta}(\boldsymbol{x},t,s)\boldsymbol{x}'} \tag{11}$$

where $h_0(d' \mid t, s)$ is a non-parametric baseline hazard function, which relates to the value of the hazard when all covariates are zero ($\boldsymbol{x} = 0$), $e^{\boldsymbol{\beta}(\boldsymbol{x},t,s)\boldsymbol{x}'}$ is a duration-independent function to represent the effect of the individuals' characteristics, and ϕ represents the parameters of the hazard function $\phi = \{h_0, \boldsymbol{\beta}\}$. FIGURE 2 shows ten randomly selected individuals' survival and hazard functions for absence periods between 4 and 5 AM. It can be observed that individuals 3 and 9 are the ones with the highest and lowest transition probability, respectively. Periods of absence that begin in this time range are likely to last for 9 to 10 hours. Thereafter, their related transition probability increases considerably. In fact, this figure is an example to depict the heterogeneity of survival functions that promotes the computation of the hazard function of each individual. This calculation can be handled by Cox-regression methods since they can integrate the influence of exogenous variables (individuals characteristics) on transition probabilities and consequently, lead to more realistic occupancy profiles.

From (11), the conditional transition probability for the two-state FSM can be explained by matrix M:

$$M = \begin{bmatrix} 1 - h(d' \mid \phi_0^{\boldsymbol{x},t}) & h(d' \mid \phi_0^{\boldsymbol{x},t}) \\ h(d' \mid \phi_1^{\boldsymbol{x},t}) & 1 - h(d' \mid \phi_1^{\boldsymbol{x},t}) \end{bmatrix} \tag{12}$$

where $h(d' \mid \phi_0^{\boldsymbol{x},t})$ and $h(d' \mid \phi_1^{\boldsymbol{x},t})$ are the hazard rates when $s = 0$ and 1, respectively.

B. SIMULATION PROCESS

In order to illustrate the dynamic of the two-state machine, used for the generation of the individual occupancy profiles, the pseudo-code of the simulation process is presented in Algorithm 1. In this algorithm, the transition signal, $\tau_k \in \{0, 1\}$ is sampled in every time-step from a Bernoulli distribution with $\alpha = h(d' \mid \phi_{z_{k-1}}^{\boldsymbol{x},t_k})$ (see line 5 of the Algorithm 1). According to this transition signal, the algorithm (i) updates the occupancy state, z_k by using an XOR function, and (ii) restarts the counter, d_k, which determines the elapsed time in each state.

Algorithm 1 Occupancy Profile Generation

```
Input: Individuals characteristics: $x_1, x_2, \dots, x_M$
Output: Discrete state sequence: $z_1, z_2, \dots, z_N$
1 begin
2     sampling $z_1 \sim [\pi_0, \pi_1]$
3     initialize $d_1 = 1$
4     for $k$=2,3,...,$N$ do
5         sampling $\tau_k \sim \text{Bern}(h(d = d_{k-1} \mid \phi_{z_{k-1}}^{\boldsymbol{x},t_k}))$
6         update $z_k = z_{k-1} \oplus \tau_k$
7         update $d_k = (1 - \tau_k)d_{k-1} + 1$
8     end
9 end
```

Once the simulation process is completed, a general verification of the model performance is carried out.[3] Hence, the model capacity to describe the occupancy of the whole TUS population is analyzed. FIGURE 3(a) presents a comparison between overall presence probabilities that have been obtained from the survey and the simulation data. This Figure shows the proposed approach effectiveness to represent the population's behavior.

Furthermore, the model capability to describe the duration of the system states has been analyzed. FIGURE 3(b) indicates a comparison between cumulative probability functions (CDF) of absence duration at 6 am. Although in this case, the distribution of duration follows a bimodal shape, the proposed approach is able to adequately reproduce its related pattern.

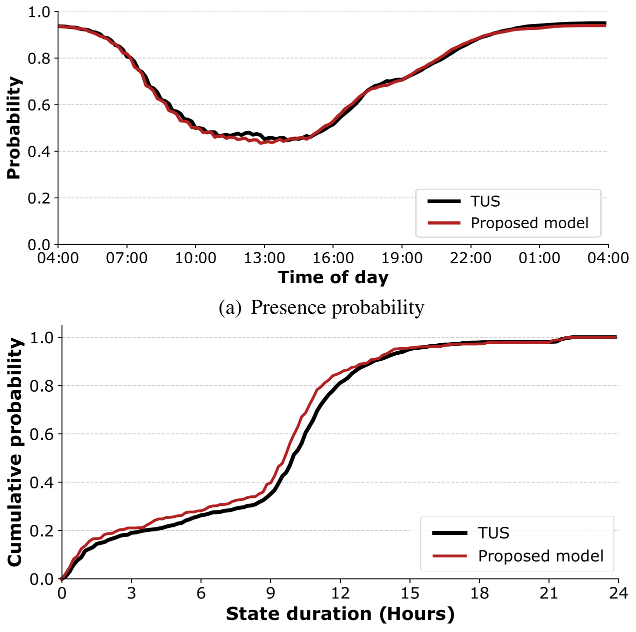
C. PERFORMANCE INDICATORS

In the following sections of the article, the performance of the model is evaluated and compared with other existing approaches in the literature by using the following metrics.

1) Mean absolute error (MAE): This metric is used to determine the model capability to represent the average occupancy state of different populations. The MAE can be expressed as,

$$MAE = \frac{1}{N}\sum_{k=1}^{N}\left|\bar{P}_s^{mod}(k) - \bar{P}_s^{tus}(k)\right| \tag{13}$$

[3] A detailed analysis of the results and their comparison with other approaches, presented in the literature, are discussed in Section V



(a) Presence probability

(b) CDF of absence duration at 6 am

FIGURE 3. **Simulation results for all the survey population.**

where $\bar{P}_s^{mod}(k)$ and $\bar{P}_s^{tus}(k)$ are average probabilities of the state s at time-step k, derived from the model simulation and the TUS data, respectively.

2) First-Wasserstein distance: Also known as Earth mover's distance (EMD), this metric assesses the model ability to represent the duration of the occupancy states. This indicator is defined by,

$$W_1 = \sum_{d=1}^{144} \left| D_s^{mod}(d) - D_s^{tus}(d) \right| \tag{14}$$

where, $D_s^{mod}(d)$ and $D_s^{tus}(d)$ are CDFs of duration of the state s, derived from the the model simulation and the TUS data, respectively.

IV. PARAMETRIC ANALYSIS

Considering the diversity of individuals' characteristics, provided by the survey, a global sensitivity analysis (GSA) is performed. GSA makes it possible to quantify each characteristic contribution to the model output variability. As a result, the most important input variables of the model can be identified. The GSA is conducted by exploiting variance-based and regression methods.

A. VARIANCE-BASED ANALYSIS

Variance-based methods evaluate the impact of model input variability on the uncertainty of its output. Therefore, given a model with the form $y = f(x_1, x_2, \ldots, x_M)$, the first-order Sobol' index is employed to quantify the relative contribution of x_i to the uncertainty of y, while excluding the interaction effect of other parameters. Additionally, the total effect index is used to evaluate the total effect of x_i considering its interaction with all others parameters. These two steps are carried out by (15) and (16), respectively,

$$S_i = \frac{V_{x_i}\left(E_{\mathbf{x}_{\sim i}}(y \mid x_i)\right)}{V(y)} \tag{15}$$

$$S_{Ti} = \frac{E_{x_{\sim i}}\left(V_{\mathbf{x}_i}(y \mid x_{\sim i})\right)}{V(y)} = 1 - \frac{V_{x_{\sim i}}\left(E_{\mathbf{x}_i}(y \mid x_{\sim i})\right)}{V(y)} \tag{16}$$

where V and E denote the variance and expected value operators, respectively, x_i stands for the i-th characteristic and the set $\mathbf{x}_{\sim i}$ contains all input variables except for x_i.

FIGURE 4 presents the result of the sensitivity analysis, performed on 10 trials of 45000 samples. The results, shown by FIGURE 4(a), evidence that most characteristics have first-order sensitivity between 4 AM and 5 PM. This coincides with the period of the day during which the variation in occupancy is the highest due to activities such as working or studying. Besides, it is observed that the characteristics between x_{20} and x_{28} have a slightly higher sensitivity index, which allows us to draw inferences about their significance.

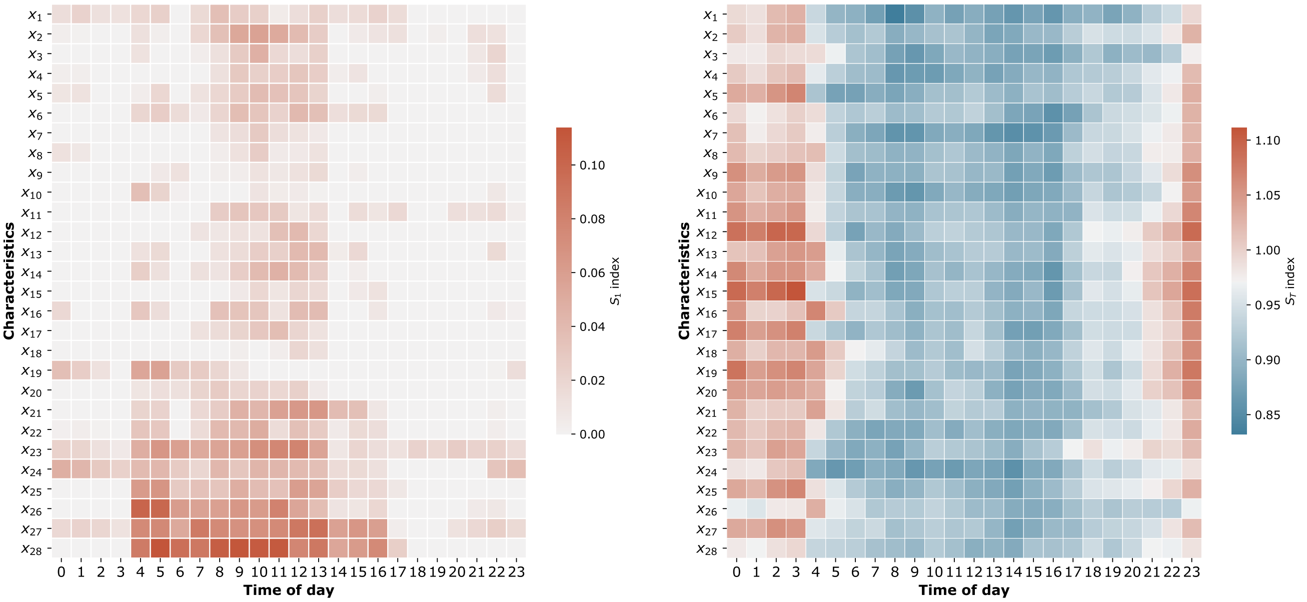
Moreover, FIGURE 4(b) depicts that the total-index is greater than the first-order index for all the characteristics. This reveals that higher-order interactions exist between the input variables. Although some characteristics do not have high importance individually, their interaction with other inputs can impact the model performance.

In addition, FIGURE 5 presents the results of the sensitivity analysis without differentiating between the time of day. It can be noticed that both first-order and total-order indexes can identify some important characteristics for modeling such as x_{23}, x_{27}, x_{28}, x_5 and x_{25}.

B. REGRESSION ANALYSIS

Regression methods assist with obtaining a model whose output is described by a linear combination of its input parameters. Therefore, the regression coefficient of a given characteristic can be considered as a sensitivity measure [48]. Backward elimination is the regression technique that has been used in this study. It is applied to each model through the following steps. Initially, each model is calibrated by using all the variables. Afterwards, the variable with the highest $p-value$ is identified and compared with a significance level, $\alpha = 0.05$. If $p-value > \alpha$, the variable is eliminated and the model is re-calibrated with the remaining characteristics. This process is repeated until all the input variables have $p-value$ lower than α.

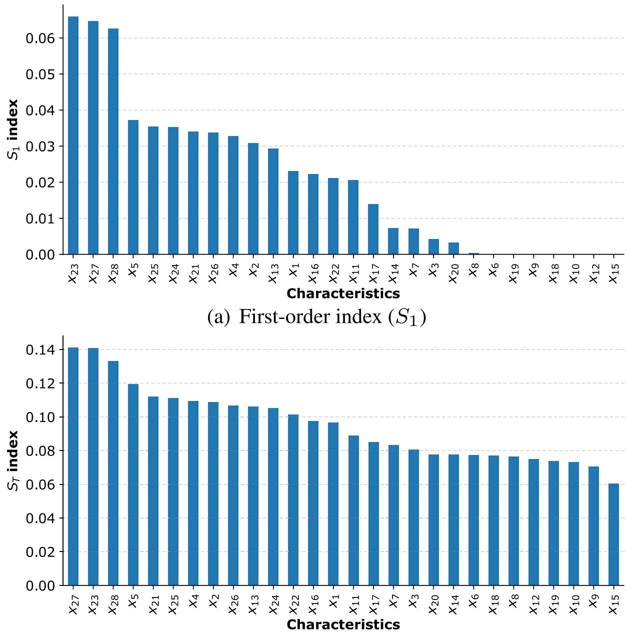
TABLE 2 summarizes the ten most important model characteristics according to the results of the feature selection process. It can be noted that these features are consistent with the ones, captured by the variance-based method. In fact, the age, (x_5) and the main activity of respondent, (x_{23}) can be presented as the two most important characteristics. Conversely, x_8 and x_{10} can be reported as the least significant input variables.



(a) First-order index (S_1) by hour of the day

(b) Total-order index (S_T) by hour of the day

FIGURE 4. Mean estimation with 45000 samples from 10 retrials of first-order and total-order Sobol' indexes.



(a) First-order index (S_1)

(b) Total-order index (S_T)

FIGURE 5. First-order and total-order Sobol' indexes without differentiation between time of day.

TABLE 2. Most significant individuals' characteristics.

	Description
x_5	Age of respondent
x_{23}	Main activity - Last 12 months
x_{26}	Educational attainment - Highest degree
x_1	Weekday
x_{16}	Age group of respondent's child(ren) in household
x_{13}	Age of respondent's youngest child in household
x_{28}	Household income
x_2	Type of day
x_{21}	Number of members in respondent's household ($\geq$ 15 years)
x_{27}	Personal income

In order to verify the impact of the feature selection on the model performance, a comparison has been made between the model, fitted with the full set of characteristics and the one, obtained from the feature selection process. In order to reduce the risk of overfitting, a k-fold cross-validation technique has been utilized. Accordingly, the TUS data set has been divided into $k = 10$ subsets with equal sizes. For each subset, the individuals have been chosen randomly based on a uniform probability distribution. These subsets have been divided into one validation set and $k - 1$ training sets to calibrate the model. This process has been repeated k times for a different validation set at each time.

The comparison results, presented in FIGURE 6, show that the backward elimination method has not substantially affected the model performance. Furthermore, it has reduced the number of model parameters from 1344 to 220 without jeopardizing the model effectiveness in explaining subpopulations' occupancy variability. It should be noted since two hazard functions (one for the presence and one for the absence) have been used for every hour, the resultant reduction is notable.

V. VALIDATION

The proposed model is validated by comparing its performance with the following methods, which have been studied in the literature.

1) First-order Markov model: This method has the Markov property and thus, it presents a system in which, the transition probability between the states only depends on the current state. In this model,

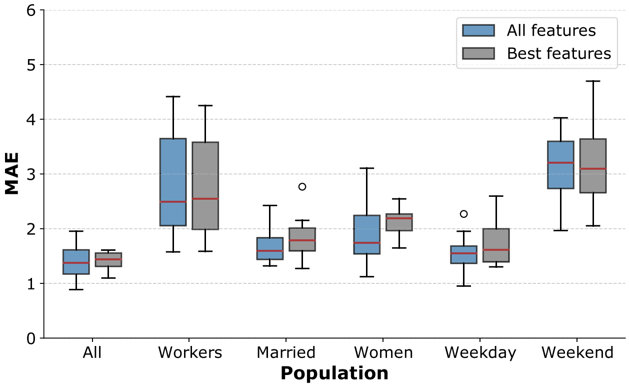


FIGURE 6. **Comparison of the model performance with and without the feature selection phase.**

the transition probabilities are calibrated at hourly intervals and behavior variability between individuals is integrated by using a logistic regression, defined by,

$$p(\boldsymbol{x}, t, s) = \frac{1}{1 + e^{-(\beta_0(\boldsymbol{x},t,s)+\beta_1(\boldsymbol{x},t,s)x_1+\cdots+\beta_M(\boldsymbol{x},t,s)x_M)}} \tag{17}$$

where $\boldsymbol{x}$ is the vector of the individuals' characteristics, t presents the hour of the day, and s stands for the system state.

Afterwards, to perform a simulation with a 10-minute time step, the Chapman-Kolmogorov equation is employed [36]. In this way, ten-minutes transition matrices can be computed based on the previously calculated hourly matrices ($M_{10\,\text{min}} = M_t^{1/6}$). Hence, the transition between the system states can be explained by the matrix M as,

$$M = \begin{bmatrix} 1 - p(\boldsymbol{x}, k, s = 0) & p(\boldsymbol{x}, k, s = 0) \\ p(\boldsymbol{x}, k, s = 1) & 1 - p(\boldsymbol{x}, k, s = 1) \end{bmatrix} \tag{18}$$

The simulation process, used for the first-order Markov model, is described by the pseudo-code in Algorithm 2. Therefore, the system state at time instant k is determined according to the transition signal τ_k. This signal is generated at every step of the simulation by sampling from a Bernoulli distribution with $\alpha = p(\boldsymbol{x}, k, z_{k-1})$.

2) Duration ranges model: This model is an improvement to the first-order Markov model based on Flett's approach [34], [38]. It is a time-dependent higher-order Markov model that generates multiple probability transition matrices in accordance with the duration, d_k of the existing state (equation (19)).
Therefore, the transition probability, used to evaluate the occurrence of a state transition, not only varies over time k but also depends on the duration of the current state. In this model, the system state, z_k and duration, d_k are updated according to the estimated transition signal, τ_k (see Algorithm 3).

Algorithm 2 First-Order Markov Model

```
Input: Individuals characteristics: $x_1, x_2, \dots, x_M$
Output: Discrete state sequence: $z_1, z_2, \dots, z_N$
begin
    sampling $z_1[\pi_0, \pi_1]$
    initialise $k = 1$
    for $k = 2, 3, \dots, N$ do
        sampling $\tau_k \sim \text{Bern}(p(\boldsymbol{x}, k, z_{k-1}))$
        update $z_k = z_{k-1} \oplus \tau_k$
    end
end
```

Algorithm 3 Duration Ranges Model

```
Input: Individual characteristics: $x_1, x_2, \dots, x_M$
Output: Discrete state sequence: $z_1, z_2, \dots, z_N$
begin
    sampling $z_1[\pi_0, \pi_1]$
    initialise $d_1 = 1$
    for $k = 2, 3, \dots, N$ do
        if $d_k \leq 12$ then
            sampling $\tau_k \sim \text{Bern}(p_{0\text{-}2}(\boldsymbol{x}, k, z_{k-1}))$
        else if $12 < d_k \leq 24$ then
            sampling $\tau_k \sim \text{Bern}(p_{2\text{-}4}(\boldsymbol{x}, k, z_{k-1}))$
        else if $24 < d_k \leq 36$ then
            sampling $\tau_k \sim \text{Bern}(p_{4\text{-}6}(\boldsymbol{x}, k, z_{k-1}))$
        else if $36 < d_k \leq 48$ then
            sampling $\tau_k \sim \text{Bern}(p_{6\text{-}8}(\boldsymbol{x}, k, z_{k-1}))$
        else
            sampling $\tau_k \sim \text{Bern}(p_{8+}(\boldsymbol{x}, k, z_{k-1}))$
        end
        update $z_k = z_{k-1} \oplus \tau_k$
        update $d_k = (1 - \tau_k)d_{k-1} + 1$
    end
end
```

$$p(\boldsymbol{x}, k, s \mid d_k) = \begin{cases} p_{0\text{-}2}(\boldsymbol{x}, k, s) & \text{if } d_k \leq 2 \text{ hrs} \\ p_{2\text{-}4}(\boldsymbol{x}, k, s) & \text{if } 2 < d_k \leq 4 \text{ hrs} \\ p_{4\text{-}6}(\boldsymbol{x}, k, s) & \text{if } 4 < d_k \leq 6 \text{ hrs} \\ p_{6\text{-}8}(\boldsymbol{x}, k, s) & \text{if } 6 < d_k \leq 8 \text{ hrs} \\ p_{8+}(\boldsymbol{x}, k, s) & \text{if } d_k > 8 \text{ hrs} \end{cases} \tag{19}$$

3) Event model: It is a probabilistic model, in which transitions between states are estimated by drawing a duration, Δk from the corresponding probability distribution function (PDF) of duration $f(d \mid \theta_s^{\boldsymbol{x},t})$. In order to estimate individual-dependent PDFs, a binary tree structure is used to split duration data according to a chosen set of individuals' characteristics. The selected characteristics explain a significant statistical

Algorithm 4 Event Model

```
Input: Individuals characteristics: $x_1, x_2, \dots, x_M$
Output: Discrete state sequence: $z_1, z_2, \dots, z_N$
1  begin
2      sampling $z_1[\pi_0, \pi_1]$
3      initialise $s = z_1$
4      initialise $k = 1$
5      while $k \leq N$ do
6          sampling $\Delta k \sim f(d \mid \theta_s^{\boldsymbol{x},t_k})$
7          update $z_{[k:k+\Delta k]} = s$
8          update $s = s \oplus 1$
9          update $k = k + \Delta k + 1$
10     end
11 end
```

Algorithm 5 Improved Event Model

```
Input: Individuals characteristics: $x_1, x_2, \dots, x_M$
Output: Discrete state sequence: $z_1, z_2, \dots, z_N$
1  begin
2      sampling $z_1[\pi_0, \pi_1]$
3      initialise $s = z_1$
4      initialise $k = 1$
5      while $k \leq N$ do
6          update $i = \min_i \left(\sum_{i'=1}^{K} p_{i'}\right) \geq r$
7          sampling $\Delta k \sim \phi_i \mathcal{N}\left(d \mid \varphi_{i,s}^{x,t_k}\right)$
8          update $z_{[k:k+\Delta k]} = s$
9          update $s = s \oplus 1$
10         update $k = k + \Delta k + 1$
11     end
12 end
```

difference between the subgroups' average duration, obtained in the branches at the end of the tree (more details can be found in [33], [36]). As a result, each individual corresponds only to one of the distinct subsets at the bottom of the tree. Afterwards, the PDFs are fitted at hourly intervals ($t \in [0, 1, \dots, 23]$) by using a Weibull distribution, defined by,

$$f(d \mid \theta_s^{\boldsymbol{x},t}) = e^{-\left(\frac{d}{\alpha(\boldsymbol{x},t,s)}\right)^{\beta(\boldsymbol{x},t,s)}} - e^{-\left(\frac{d+1}{\alpha(\boldsymbol{x},t,s)}\right)^{\beta(\boldsymbol{x},t,s)}} \tag{20}$$

where $\theta_s^{\boldsymbol{x},t}$ represents the scale and shape parameters of the Weibull distribution.

As illustrated in Algorithm 4, once a transition occurs at time-step k, the duration Δk of the new state is estimated (row 6). Therefore, a deterministic process takes place between k and $k + \Delta k$, where the state of the system remains unchanged (row 7). Once the simulation reaches the time-step $k+\Delta k$, the system is forced to change the state (row 8) and update the k value.

4) Improved event model: As discussed above, the event model use a binary tree structure to divide the duration data in order to estimate individual-dependent PDFs. However, as stated in the literature [37], the Weibull distribution has limitations in capturing the multimodal character of duration data. Accordingly, we propose the utilization of Gaussian Mixture Model (GMM) instead of Weibull distribution. GMM assists with a better representation of duration with multimodal shape. Therefore, the PDFs can be represented as the summation of several normal distributions $\mathcal{N}(d \mid \mu_i, \sigma_i)$, expressed by,

$$f(d \mid \theta_s^{\boldsymbol{x},t}) = \sum_{i=1}^{K} \phi_i \mathcal{N}(d \mid \mu_i(\boldsymbol{x}, t, s), \sigma_i(\boldsymbol{x}, t, s)) \tag{21}$$

$$= \sum_{i=1}^{K} \phi_i \mathcal{N}\left(d \mid \varphi_{i,s}^{\boldsymbol{x},t}\right) \tag{22}$$

where K is the number of clusters, estimated by the silhiutte method,[4] $\varphi_{i,s}^{\boldsymbol{x},t}$ represents the mean $\mu_i(\boldsymbol{x}, t, s)$ and the variance $\sigma_i(\boldsymbol{x}, t, s)$ of the i^{th} component, and ϕ_i presents the mixture component weight, which must satisfy the condition $\sum_{i=1}^{K} \phi_i = 1$.

Additionally, as a part of the simulation process (see Algorithm 5), it is necessary to determine the probability that an individual belongs to a given cluster. This probability can be estimated from (22) through,

$$p_i = \frac{\phi_i \mathcal{N}\left(d \mid \varphi_{i,s}^{x,t}\right)}{\sum_{i=1}^{K} \phi_i \mathcal{N}\left(d \mid \varphi_{i,s}^{x,t}\right)}, \quad \text{for } i \in [1, K] \tag{23}$$

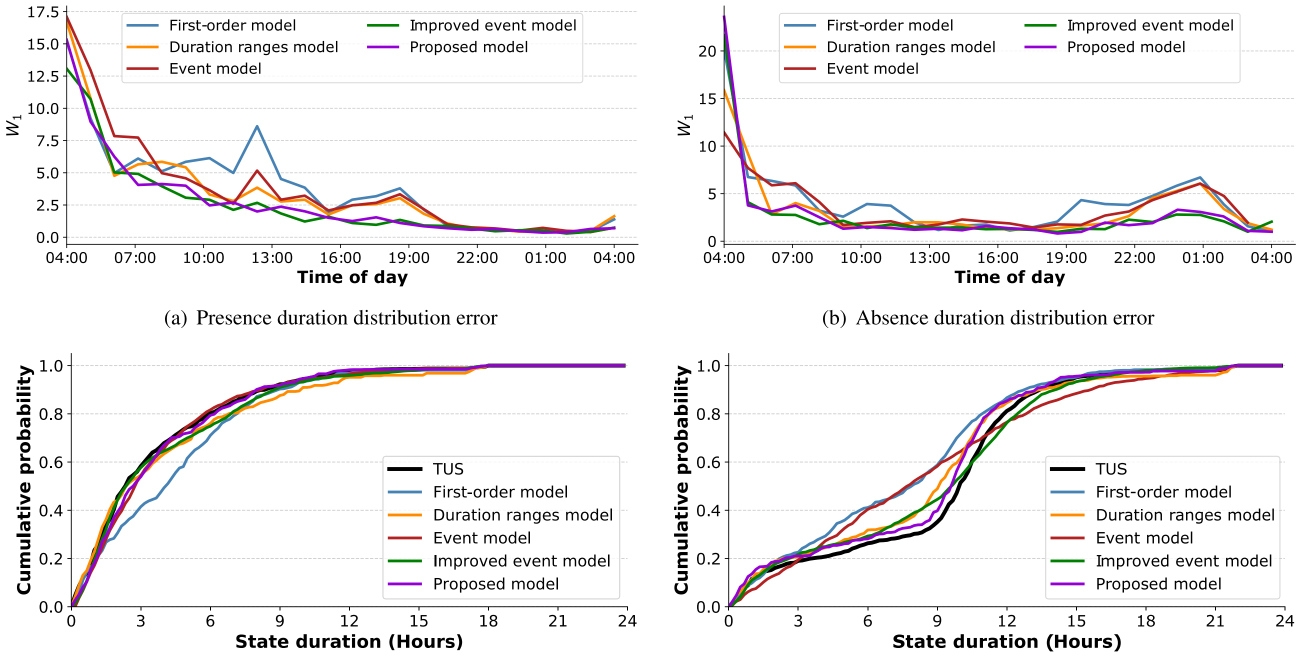
where p_i is a vector of dimension K in which each element indicates the probability that an individual belongs to a cluster.

As illustrated in Algorithm 5, once a transition occurs at time-step k, each individual is assigned to one of the clusters, defined by the GMM. In order to do this, the vector p_i is compared with the uniformly distributed random variable $r \in [0, 1]$ (row 6). Subsequently, according to the duration probability distribution ($\phi_i \mathcal{N}\left(d \mid \varphi_{i,s}^{x,t}\right)$), associated to the selected cluster, the duration Δk of the new state is estimated (row 7). Afterwards, the current occupancy state is maintained from k to $k + \Delta k$ (row 8). Once this simulation step is reached, the occupancy state changes and the simulation step k is updated.

A. WHOLE POPULATION

To validate the proposed approach, a comparison is initially made by evaluating the models performance in estimating the overall presence probability. The results, presented in

[4] The silhiutte method [49] allows to identify the appropriate number of distributions (clusters) in a clustering analysis.



(a) Presence duration distribution error

(b) Absence duration distribution error

(c) Geometric distribution.

(d) Multimodal distribution.

FIGURE 7. Comparison of duration distribution representation.

TABLE 3. Summary of the performance comparison of whole population occupancy predictions by the utilized models.

	MAE	W_1
First order model	3.07 ± 0.99	4.03/4.11
Duration ranges model	3.24 ± 0.59	3.39/3.44
Event model	3.30 ± 0.78	3.68/3.71
Improved event model	1.60 ± 0.94	2.47/2.78
Proposed model	1.44 ± 0.71	2.64/2.72

W_1 values corresponding to presence/absence states.

TABLE 3, show that among the baseline approaches, the first-order model has the best accuracy in estimating the presence probability of individuals. However, the duration range and event models can improve the first-order model performance in explaining the duration distribution of the occupancy states by up to 20%.

TABLE 3 indicates that the proposed model can outperform other approaches and reduce the error in estimating both the presence probability and the duration distribution by up to 56% and 34%, respectively. Furthermore, it can be observed that the recommended improvement to the event model, presented by Wilke *et al.* [33], has yielded a reduction of up to 50% in the MAE. This reduction has resulted in a performance close to that achieved by the proposed explicit duration model.

Moreover, FIGURE 7 presents a comparison of the duration distribution errors according to the time of day. FIGURE 7(a) and FIGURE 7(b) show that compared to other techniques, the proposed model and the improved event model have achieved better performances in both presence and absence duration probability estimation. This proves the capability of these methods to efficiently evaluate events duration even in cases with multimodal distribution (FIGURE 7(d)). It should be noted that during different times of day, the performance of all models is similar. In fact, this similarity occurs when distributions have an exponential shape (FIGURE 7(c)).

B. SUB-POPULATION

Moreover, the models performance has been compared to realize their ability to explain the behavior of different individuals' sub-populations. The results of this comparison are presented in TABLE 4. It can be observed that the proposed approach and the improved event model result in a more accurate representation of sub-populations' occupancy patterns. Particularly, the proposed approach can accomplish an improvement of up to 55% and 29% in representing both the presence probability and the duration distribution of sub-populations, respectively.

FIGURE 8 provides an example that shows the capability of the model to represent the variability of occupancy profiles among different sub-populations. For this purpose, four groups of individuals with significantly different occupancy patterns have been selected.

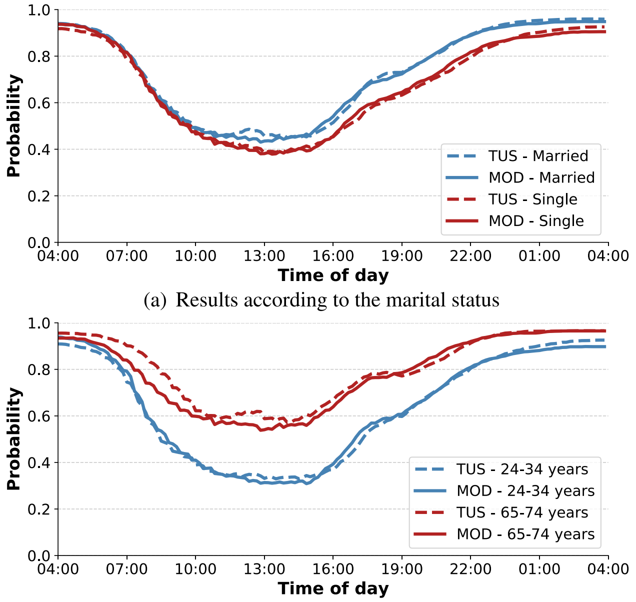
VI. DISCUSSION AND FUTURE PROSPECTS

A new probabilistic approach to modeling household occupancy has been proposed. It has been validated through a comparison with other techniques, reported in the literature. The results, obtained in Section V, demonstrate that the hazard-based model provides an appropriate methodology for modeling household occupancy. The comparative

TABLE 4. Performance comparison in sub-populations' occupancy prediction.

Population	First order model		Duration ranges model		Event model		Improved event model		Proposed model	
	MEA	W_1	MAE	W_1	MAE	W_1	MAE	W_1	MAE	W_1
Workers	3.85 ± 1.02	5.06/4.78	4.44 ± 1.40	5.02/4.17	4.28 ± 0.96	5.07/4.64	2.42 ± 1.08	4.08/3.87	2.76 ± 1.32	4.19/4.07
Retirees	3.90 ± 1.58	7.05/7.14	5.18 ± 1.60	6.63/6.39	3.98 ± 2.14	6.53/5.26	2.67 ± 1.48	6.05/5.51	3.49 ± 1.16	6.22/6.14
Married	3.46 ± 1.20	5.35/5.52	3.75 ± 1.10	5.12/5.16	3.79 ± 1.12	5.11/5.00	2.23 ± 0.74	4.36/4.35	1.80 ± 0.82	4.29/4.27
Single	3.52 ± 1.74	6.92/6.25	3.31 ± 1.38	7.19/5.74	3.71 ± 1.28	7.21/5.35	2.81 ± 1.44	6.72/5.44	2.73 ± 1.56	6.81/5.49
Men	3.54 ± 1.02	5.22/5.28	3.41 ± 1.44	4.81/4.40	3.52 ± 0.94	5.07/4.89	2.09 ± 1.06	4.63/4.09	1.87 ± 0.56	4.95/4.39
Women	3.30 ± 1.14	5.40/5.15	3.62 ± 1.00	4.96/4.91	3.47 ± 0.58	5.40/4.42	2.22 ± 0.96	4.27/4.55	1.82 ± 0.90	4.82/4.15
Weekday	3.59 ± 1.28	4.85/5.16	3.70 ± 0.66	4.57/4.51	4.04 ± 0.88	4.58/4.88	1.91 ± 1.10	3.71/3.68	1.60 ± 1.16	3.59/3.68
Weekend	3.06 ± 1.22	6.69/6.07	3.46 ± 1.62	6.67/5.88	3.22 ± 1.36	6.32/5.45	2.08 ± 0.82	6.22/5.42	2.91 ± 1.34	6.33/5.52

W_1 values correspond to the results of presence/absence states.



(a) Results according to the marital status

(b) Results according to the age range

FIGURE 8. Simulation results of the occupancy patterns of some individuals sub-populations.

study validates that the proposed method is more accurate in examining the duration distribution of occupancy states. Furthermore, it has higher efficiency in representing occupancy variability among sub-populations. It should be noted that the cross-validation analysis has been also employed since it can help reduce the overfitting risk and preserve the model generalization capability.

Additionally, the event model proposed by Wilke *et al.* [33] has been improved by adding a GMM. In fact, the inclusion of this clustering method is advantageous since it has improved the MAE performance by up to 50% with respect to the initial model. Moreover, the improved event model has yielded a performance comparable to the proposed explicit duration approach.

Besides, the results of the parametric analysis, performed in Section IV, show that the parameter selection for model calibration is important to achieve an adequate representation of the occupancy variability. It should be mentioned that some of the selected parameters such as age, employment status (main activity), dwelling type, income, and weekdays (TABLE 2) are consistent with those mentioned in previous studies [33], [34], [39]. Indeed, the utilized feature selection procedure can reduce the model complexity and result in a baseline for other researchers. This baseline can help i) formulate the most relevant questions to conduct a survey and ii) select the most appropriate variables for occupancy model calibration. Nevertheless, it should be noted that the parameters of interest can change according to the application context and the accuracy requirements.

Moreover, the robustness of the model is another topic to study in order to provide a more rigorous and detailed performance analysis. For this goal, the consistency of the model outcomes under significant variations of input data can be examined. The Monte Carlo simulation method can be used to perform such practice. Furthermore, the boundary of the prediction variance can be explored since input variables, explaining human behavior, have stochastic nature. Estimating the error of every single occupancy profile can be another concern to study. This is due to the fact that performance indicators usually explain the model ability to present the average behavior of a population. However, the primary output of occupancy models (including the one presented in this paper) is a time-series profile per individual. According to this issue, the consistency of the generated profiles considering the original ones can be also evaluated.

The number of the states of the proposed model can vary from two to N. This is beneficial to applications such as occupant activities prediction as an exercise that can significantly influence the performance of bottom-up models for electrical energy consumption and generation. Our approach can be also used to model household members interactions. For this purpose, the development of a multi-agent system with a leader has been advised in literature [31]. In such a system, the leader is an independent agent who conditions other agents' behavior according to their current state. On the other side, the integration of holidays, vacations, and seasons into the occupancy model is another aspect that requires significant consideration in order to generate annual profiles.

Moreover, the analysis of prolonged absences and interactions between occupants should be considered, albeit limitations related to surveys data. The model adaptability is another concern with regard to this type of data. Although TUS are useful methods for explaining occupants behavior, the models, offered by them, are not adoptable due to the static nature of the data. Accordingly, the use of in-situ measurements based on the fusion of environmental sensors has been promoted as an efficient and cost-effective alternative to model real-life occupants behavior [10]. This assists with analyzing changes in occupants' habits and lifestyles as well as the influence of external factors such as weather. Furthermore, it can be used to evaluate the impact of behavioral information on actual energy management applications. Nevertheless, a future research direction can use surveys information as prior knowledge to speed up the convergence of algorithms and improve their accuracy.

VII. CONCLUSION

This paper presents an explicit-duration probabilistic model to predict individuals' occupancy profiles in their dwellings. The purpose of this work is to provide an approach that allows a more accurate representation of residential occupancy, which is significantly important to enhance the performance, design, and simulation of buildings. In this study, the influence of socio-demographic characteristics on model performance has been analyzed through a parametric method. As a result, the most significant individuals' characteristics has been identified that allows for using a reduced set of variables to develop the model. In this way, it is possible to reduce the complexity of the model without affecting its accuracy in predicting occupancy profiles. Besides, the effectiveness of the proposed model has been thoroughly examined within a comparative study. This study demonstrates that the recommended approach is able to reduce the error in describing both the presence probability and the duration distribution by up to 56% and 34%, respectively. The utilized method and the parametric analysis are applicable to other TUS and indoor activities prediction exercises. Furthermore, an event model, studied in the literature, has been improved in this study by using GMM. The suggested improvement has reduced the MAE of this model by up to 50% and made it competitive with the proposed explicit-duration probabilistic model. Literature has highlighted occupancy and occupant behavior as crucial factors in improving building energy performance. Indeed, several works have demonstrated that a significant amount of energy can be saved by adopting occupancy and activity-based control strategies. Therefore, a comprehensive analysis of residential energy-saving potentials based on these control strategies can bring about interesting subjects for future research. In this regard, providing realistic occupancy profiles becomes a primary concern to analyze the behavior diversity among people. The explicit-duration probabilistic model, presented in this article, is favored as an efficient method to address this issue.

ACKNOWLEDGMENT

The authors would like to thank the Laboratoire des technologies de l'énergie d'Hydro-Québec, the Natural Science and Engineering Research Council of Canada, the Foundation of Université du Québec à Trois-Rivières, the Mitacs accelerate program, and Dr. Sayed Saeed Hosseini.

REFERENCES

[1] S. Aman, Y. Simmhan, and V. K. Prasanna, "Energy management systems: State of the art and emerging trends," *IEEE Commun. Mag.*, vol. 51, no. 1, pp. 114–119, Jan. 2013.

[2] Y. Zhang, X. Bai, F. P. Mills, and J. C. V. Pezzey, "Rethinking the role of occupant behavior in building energy performance: A review," *Energy Buildings*, vol. 172, pp. 279–294, Aug. 2018.

[3] D. Yan, T. Hong, B. Dong, A. Mahdavi, S. D'Oca, I. Gaetani, and X. Feng, "IEA EBC annex 66: Definition and simulation of occupant behavior in buildings," *Energy Buildings*, vol. 156, pp. 258–270, Dec. 2017.

[4] H. Yoshino, T. Hong, and N. Nord, "IEA EBC annex 53: Total energy use in buildings—Analysis and evaluation methods," *Energy Buildings*, vol. 152, pp. 124–136, Oct. 2017.

[5] B. F. Balvedi, E. Ghisi, and R. Lamberts, "A review of occupant behaviour in residential buildings," *Energy Buildings*, vol. 174, pp. 495–505, Sep. 2018.

[6] E. Delzendeh, S. Wu, A. Lee, and Y. Zhou, "The impact of occupants' behaviours on building energy analysis: A research review," *Renew. Sustain. Energy Rev.*, vol. 80, pp. 1061–1071, Dec. 2017.

[7] B. Dong, D. Yan, Z. Li, Y. Jin, X. Feng, and H. Fontenot, "Modeling occupancy and behavior for better building design and operation—A critical review," *Building Simul.*, vol. 11, no. 5, pp. 899–921, Oct. 2018.

[8] F. Amara, K. Agbossou, Y. Dubé, S. Kelouwani, A. Cardenas, and S. S. Hosseini, "A residual load modeling approach for household short-term load forecasting application," *Energy Buildings*, vol. 187, pp. 132–143, Mar. 2019.

[9] S. Sansregret, K. Lavigne, B. Le Lostec, L. Francois, and F. Guay, "High resolution bottom-up residential electrical model for distribution networks planning," in *Proc. 16th Conf. IBPSA*, 2019, pp. 3540–3547.

[10] L. Rueda, K. Agbossou, A. Cardenas, N. Henao, and S. Kelouwani, "A comprehensive review of approaches to building occupancy detection," *Building Environ.*, vol. 180, Aug. 2020, Art. no. 106966.

[11] T. A. Nguyen and M. Aiello, "Energy intelligent buildings based on user activity: A survey," *Energy Buildings*, vol. 56, pp. 244–257, Jan. 2013.

[12] I. Georgievski, T. A. Nguyen, F. Nizamic, B. Setz, A. Lazovik, and M. Aiello, "Planning meets activity recognition: Service coordination for intelligent buildings," *Pervas. Mobile Comput.*, vol. 38, pp. 110–139, Jul. 2017.

[13] M. Milenkovic and O. Amft, "An opportunistic activity-sensing approach to save energy in office buildings," in *Proc. 4th Int. Conf. Future Energy Syst.*, 2013, pp. 247–258.

[14] J. Scott, A. J. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, and N. Villar, "PreHeat: Controlling home heating using occupancy prediction," in *Proc. 13th Int. Conf. Ubiquitous Comput.*, 2011, pp. 281–290.

[15] P. J. Boait and R. M. Rylatt, "A method for fully automatic operation of domestic heating," *Energy Buildings*, vol. 42, no. 1, pp. 11–16, Jan. 2010.

[16] C. D. Korkas, S. Baldi, I. Michailidis, and E. B. Kosmatopoulos, "Intelligent energy and thermal comfort management in grid-connected microgrids with heterogeneous occupancy schedule," *Appl. Energy*, vol. 149, pp. 194–203, Jul. 2015.

[17] C. D. Korkas, S. Baldi, I. Michailidis, and E. B. Kosmatopoulos, "Occupancy-based demand response and thermal comfort optimization in microgrids with renewable energy sources and energy storage," *Appl. Energy*, vol. 163, pp. 93–104, Feb. 2016.

[18] S. Baldi, C. D. Korkas, M. Lv, and E. B. Kosmatopoulos, "Automating occupant-building interaction via smart zoning of thermostatic loads: A switched self-tuning approach," *Appl. Energy*, vol. 231, pp. 1246–1258, Dec. 2018.

[19] S. Srinivasan and C. R. Bhat, "Modeling household interactions in daily in-home and out-of-home maintenance activity participation," *Transportation*, vol. 32, no. 5, pp. 523–544, Sep. 2005.

[20] J. Torriti, "A review of time use models of residential electricity demand," *Renew. Sustain. Energy Rev.*, vol. 37, pp. 265–272, Sep. 2014.

[21] M. J. Roorda, E. J. Miller, and K. M. N. Habib, "Validation of TASHA: A 24-H activity scheduling microsimulation model," *Transp. Res. A, Policy Pract.*, vol. 42, no. 2, pp. 360–375, Feb. 2008.

[22] G. Buttitta, O. Neu, W. J. Turner, and D. Finn, "Modelling household occupancy profiles using data mining clustering techniques on time use data," in *Proc. Building Simulation*, San Francisco, CA, USA, Aug. 2017, pp. 1788–1797.

[23] C. F. Walker and J. L. Pokoski, "Residential load shape modelling based on customer behavior," *IEEE Trans. Power App. Syst.*, vol. PAS-104, no. 7, pp. 1703–1711, Jul. 1985.

[24] I. Richardson, M. Thomson, and D. Infield, "A high-resolution domestic building occupancy model for energy demand simulations," *Energy Buildings*, vol. 40, no. 8, pp. 1560–1566, Jan. 2008.

[25] Office for National Statistics and Ipsos-RSL. (2003). *United kingdom time use survey, 2000*. [Online]. Available: http://doi.org/10.5255/UKDA-SN-4504-1

[26] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: A high-resolution energy demand model," *Energy Buildings*, vol. 42, no. 10, pp. 1878–1887, Oct. 2010.

[27] J. Widén, A. M. Nilsson, and E. Wäckelgård, "A combined Markov-chain and bottom-up approach to modelling of domestic lighting demand," *Energy Buildings*, vol. 41, no. 10, pp. 1001–1012, Oct. 2009.

[28] J. Widén and E. Wäckelgård, "A high-resolution stochastic model of domestic activity patterns and electricity demand," *Appl. Energy*, vol. 87, no. 6, pp. 1880–1892, Jun. 2010.

[29] M. Muratori, M. C. Roberts, R. Sioshansi, V. Marano, and G. Rizzoni, "A highly resolved modeling technique to simulate residential power demand," *Appl. Energy*, vol. 107, pp. 465–473, Jul. 2013.

[30] A. J. Collin, G. Tsagarakis, A. E. Kiprakis, and S. McLaughlin, "Development of low-voltage load models for the residential load sector," *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2180–2188, Sep. 2014.

[31] M. L. Baptista, H. Prendinger, R. Prada, and Y. Yamaguchi, "A cooperative multi-agent system to accurately estimate residential energy demand," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, 2014, pp. 1405–1406.

[32] J. Conlisk, "Interactive Markov chains," *J. Math. Sociol.*, vol. 4, no. 2, pp. 157–185, 1976.

[33] U. Wilke, F. Haldi, J.-L. Scartezzini, and D. Robinson, "A bottom-up stochastic model to predict building occupants' time-dependent activities," *Building Environ.*, vol. 60, pp. 254–264, Feb. 2013.

[34] G. Flett and N. Kelly, "An occupant-differentiated, higher-order Markov chain method for prediction of domestic occupancy," *Energy Buildings*, vol. 125, pp. 219–230, Aug. 2016.

[35] J. Tanimoto, A. Hagishima, and H. Sagara, "A methodology for peak energy requirement considering actual variation of occupants' behavior schedules," *Building Environ.*, vol. 43, no. 4, pp. 610–619, Apr. 2008.

[36] U. Wilke, "Probabilistic bottom-up modelling of occupancy and activities to predict electricity demand in residential buildings," Ph.D. dissertation, École Polytechnique Fédérale De Lausanne, Lausanne, Switzerland, 2013.

[37] E. Vorger, "Étude de l'influence du comportement des habitants sur la performance énergétique du bâtiment," Ph.D. dissertation, l'École nationale supérieure des mines de Paris, Paris, France, 2014.

[38] G. Flett, "Modelling and analysis of energy demand variation and uncertainty in small-scale domestic energy systems," Ph.D. dissertation, Dept. Mech. Aerosp. Eng., Univ. Strathclyde, Scotland, U.K., 2017.

[39] D. Aerts, J. Minnen, I. Glorieux, I. Wouters, and F. Descamps, "A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison," *Building Environ.*, vol. 75, pp. 67–78, May 2014.

[40] D. Aerts, "Occupancy and activity modelling for building energy demand simulations, comparative feedback and residential electricity demand characterisation," Ph.D. dissertation, Vrije Univ. Brussel, Brussels, Belgium, 2015.

[41] H. Hou, J. Pawlak, A. Sivakumar, B. Howard, and J. Polak, "An approach for building occupancy modelling considering the urban context," *Building Environ.*, vol. 183, Oct. 2020, Art. no. 107126.

[42] F. Haldi and D. Robinson, "Interactions with window openings by office occupants," *Building Environ.*, vol. 44, no. 12, pp. 2378–2395, Dec. 2009.

[43] S. P. Jenkins, "Survival analysis," Unpublished manuscript, Inst. for Social Econ. Res., Univ. Essex, Colchester, U.K., Tech. Rep., 2005, pp. 54–56.

[44] S. Canada. *General Social Survey—Time Use (GSS)* Accessed: May 1, 2020. [Online]. Available: https://www.statcan.gc.ca

[45] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, vol. 793. Hoboken, NJ, USA: Wiley, 2019.

[46] M. Dewar, C. Wiggins, and F. Wood, "Inference in hidden Markov models with explicit state duration distributions," *IEEE Signal Process. Lett.*, vol. 19, no. 4, pp. 235–238, Apr. 2012.

[47] V. Barbu and N. Limnios, "Empirical estimation for discrete-time semi-Markov processes with applications in reliability," *J. Nonparametric Statist.*, vol. 18, nos. 7–8, pp. 483–498, Oct. 2006.

[48] A. Saltelli, S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, vol. 1. Hoboken, NJ, USA: Wiley, 2004.

[49] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

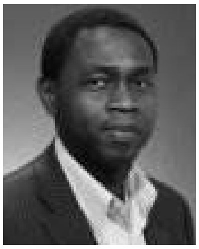
LUIS RUEDA received the B.S. degree in electronic engineering from the Universidad Industrial de Santander, Bucaramanga, Colombia, in 2013, and the master's degree in electronic engineering from the Universidad Industrial de Santander, Bucaramanga, Colombia, in 2017. He is currently pursuing the Ph.D. degree in electrical engineering with the Smart Energy Research and Innovation Laboratory, Université du Québec à Trois-Rivières, QC, Canada. His research interests include residential energy management systems, artificial intelligence applications for smarts grids, embedded systems, renewable energies, and occupant behavior modeling in residential buildings.

SIMON SANSREGRET received the B.A.Sc. and M.A.Sc. degrees in mechanical engineering with specialization in energy from the University of Sherbrooke. He is a member of the Ordre des ingénieurs du Québec. He has been a Researcher with the laboratoire des technologies de l'énergie (LTE), Hydro-Québec Research Institute, since 2001. His expertise is related to energy efficiency and demand respond in building sector. In recent years, he has been devoted to the development of simulation tools in order to improve the energy efficiency of commercial and institutional buildings. He was responsible for the development of simulation software called SIMEB, an interface to EnergyPlus Simulation Engine. He has published several scientific papers in connection with the building energy simulation, model calibration, and visualization of building performance data. He also contributed to various projects related to energy consumption and demand response in the residential sector. He was on the board of directors of the Canadian Chapter of International Building Performance Association (IBPSA-Canada) from 2010 to 2016.

BRICE LE LOSTEC received the M.S. degree in science and technologies from the Université de Savoie, France, in 2005, and the Ph.D. degree in mechanical engineering from Sherbrooke University, Canada, in 2010. His research interests include building stock modeling, solar energy, power generation from low grade heat, electrically driven compression heat pump, absorption heat pump, and refrigeration.

KODJO AGBOSSOU (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic measurements from the Université de Nancy I, France, in 1987, 1989, and 1992, respectively. He is currently the Hydro-Québec Research Chairholder on Transactive Management of Power and Energy in the Residential Sector, and the Chair of the Smart Energy Research and Innovation Laboratory of Université du Québec à Trois-Rivières (UQTR). He was the Head of Engineering School, UQTR, from 2011 to 2017. He was the Head of the Department of Electrical and Computer Engineering Department, UQTR, from 2007 to 2011. He was also the Director of Graduate Studies in Electrical Engineering, UQTR, from 2002 to 2004. He was a Postdoctoral Researcher (1993–1994) with the Electrical Engineering Department, UQTR, and was a Lecturer (1997–1998) at the same department. He is the author of more than 325 publications and has four patents and two Patent Pending. His present research activities are in the areas of renewable energy, the use of hydrogen, Home demand side management (HDSM), integration of energy production, storage and electrical energy generation system, connection of electrical vehicle to the grid, control and measurements. He is a member of the Hydrogen Research Institute and Research group "GREI" of UQTR. Since 2015, he has been the Sub-Commitee Chair on Home and Building Energy Management of Smart Grid Technical Committee," IEEE Industrial Electronics Society (IES).

NILSON HENAO received the B.S. degree in electronics engineering from the Universidad de los Llanos, Villavicencio, Colombia, in 2010, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Quebec at Trois Rivières (UQTR), Trois-Rivières, QC, Canada, in 2013 and 2018, respectively. His research interests include statistical and machine learning methods with applications to energy management in the residential sector, distributed optimization, multi-agent control, smart grid technologies, intelligent energy planning, smart energy storage, and load monitoring.



SOUSSO KELOUWANI (Senior Member, IEEE) received the Ph.D. degree in robotics systems from Ecole Polytechnique de Montreal, in 2011, and completed a postdoctoral internship on fuel cell hybrid electric vehicles at the Université du Québec à Trois-Rivières (UQTR), in 2012.

He is currently a Full Professor of Mechatronics with the Department of Mechanical Engineering since 2017 and a member of the Hydrogen Research Institute. He holds four patents in USA and Canada, in addition to having published more than 100 scientific articles. He is the Holder of the Canada Research Chair in Energy Optimization of Intelligent Transport Systems and the Holder of the Noovelia Research Chair in Intelligent Navigation of Autonomous Industrial Vehicles. He developed expertise in the optimization and the intelligent control of vehicular applications. In 2019, his team received the 1st Innovation Prize in partnership with DIVEL, awarded by the Association des Manufacturiers de la Mauricie et Center-du-Québec for the development of an autonomous and natural navigation system. In 2017, he received the Environment Prize at the Gala des Grands Prix d'excellence en transport from the Association québécoise du Transport (AQTr) for the development of hydrogen range extenders for electric vehicles. He was the Co-President and President of the technical committee of the IEEE International Conferences on Vehicular Power and Propulsion in Chicago (USA, 2018) and in Hanoi (Vietnam, 2019). His research interests focus on optimizing energy systems for vehicle applications, advanced driver assistance techniques, and intelligent vehicle navigation taking into account Canadian climatic conditions. He is a member of the Order of Engineers of Quebec. He is the Winner of the Canada General Governor Gold Medal, in 2003.