Building and Environment 121 (2017) 277–290

Contents lists available at ScienceDirect

Building and Environment

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

A new modeling approach for short-term prediction of occupancy in residential buildings

Zhaoxuan Li, Bing Dong, Ph.D *

Department of Mechanical Engineering, The University of Texas at San Antonio, San Antonio, United States

ARTICLE INFO

Article history:
Received 7 February 2017
Received in revised form
21 April 2017
Accepted 8 May 2017
Available online 10 May 2017

Keywords:
Occupancy
Residential building
Statistical modeling

ABSTRACT

Occupancy models are necessary towards design and operation of smart buildings. Developing an appropriate algorithm to predict occupancy presence will allow a better control and optimization of the whole building energy consumption. However, most previous studies of development of such model only focus on commercial buildings. The occupancy model of residential houses are usually based on Time User Survey data. This study focuses on providing a unique data set of four residential houses collected from occupancy sensors. A new inhomogeneous Markov model for occupancy presence prediction is proposed and compared to commonly used models such as Probability Sampling, Artificial Neural Network, and Support Vector Regression. Training periods for the presence prediction are optimized based on change-point analysis of historical data. The study further explores and evaluates the predictive capability of the models by various temporal scenarios, including 15-min ahead, 30-min ahead, 1-hour ahead, and 24-hour ahead forecasts. The spatial-level comparison is additionally conducted by evaluating the prediction accuracy at both room-level and house-level. The final results show that the proposed Markov model outperforms the other methods in terms of an average 5% correctness with 11% maximum difference in 15-min ahead forecast of the occupancy presence. However, there is not much differences observed for 24-hour ahead forecasts.

1. Introduction

Nowadays, buildings are increasingly expected to meet higher and potentially more complex standards on the energy efficiency, sustainability, comfort, and yet to be maintained economically. The latest smart building can yield substantial savings on the energy consumptions and help maintain thermal comfort [1]. Occupancy plays a key role to provide information for smart building controls. Occupancy in offices is less varied comparing to homes, which has a large uncertainty in residential environment [2,3]. Meanwhile, the occupancy states (e.g. the presence and the number), rather than actions are more relevant to building automation system, especially when buildings are equipped with occupancy sensors [4,5]. However, occupancy detection in residential homes is usually difficult because of privacy issues. Hence, it is interesting to investigate occupancy models in residential buildings and how to utilize such model to predict future changes of behavior patterns. In addition, this model should be able to predict in different temporal resolutions (e.g. 15-min to 24-hour window), spatial scales (e.g. a single person room or multiple people house), and occupant types (e.g. the occupancy presence or the number of occupants), which is missing in current literature.

The majority of previous models are used to generate stochastic daily occupancy profiles as a specific estimation problem for building energy performance simulation tools rather than real-time optimal controls. They are considered to be “validated” if one aspect of the interested information (e.g. mean and variance) of the occupancy is simulated in a statistically reasonable way. In other words, the models are matching the similarities between the simulated occupancy profiles and the monitored data in terms of the average arrival time, the average occupied rates, and the average departure time. They are not evaluated by an exact “one to one” matching between the prediction and ground truth at each time step, which is necessary in the advanced control applications such as ventilation control of an occupied space.

* Corresponding author.
E-mail address: bing.dong@utsa.edu (B. Dong).

http://dx.doi.org/10.1016/j.buildenv.2017.05.005
0360-1323/

This paper is organized as following: the authors review and provide insights on the shortcomings of current studies on modeling approaches of the occupancy in Section 2; Section 3 introduces the new developed Markov chain model for predictive control purpose, and three commonly used modeling tools, which are probability sampling, Artificial Neural Network and Support Vector Regression models; Section 4 provides the results of each model on occupancy prediction of four residential houses and temporal and spatial differences between models' predictive performance are further assessed; Section 5 discusses importance and concludes this study.

2. Current state-of-the art

The occupancy modeling is gaining more attentions in both building design and operation related research studies. Many modeling techniques have been developed. The most commonly used approaches are the probability sampling and the Markov process [6–8]. These two models are developed based on probability theories and suitable for both estimation and prediction. Statistical learning is another approach. It utilizes the statistical mining, such as decision tree branching [9], to cluster the occupancy schedules. Machine learning is an enhanced approach that can further be used to predict occupancy. Three basic types of occupancy information can be modeled for the aforementioned approaches [10,11]: 1) occupied status at a space level, which refers to whether or not a space is occupied at a particular time; 2) number of occupants at a space level, which refers to how many occupants are in a space at a particular time; and 3) occupant tracking, which refers to individual movement and behavior tracking. In this section, we focus on the review of sampling and Markov approaches.

Traditionally, building occupancy simulation is based on predefined static schedules. With newly developed co-simulation platform, it is possible to replicate relatively realistic occupancy schedules by integrating energy simulation software with stochastic occupancy models [12]. The probability sampling approach is one popular model that estimates the stochastic occupancy presence [12–15]. The presence profile is generated from a sampling process of the fitted distributions from historical data. The training data containing occupancy information is normally collected from multiple rooms of the same space type in one building. Then, the average occupied hour per day, the average vacancy ratio per day, the average departure and arrival time per day, and their standard deviations can be compared and modeled for multiple days in a cross-sectional analysis [14,15]. In essence, the modeling of these key information using probabilistic distributions is through the cumulative distribution functions (CDFs), describing the first arrival time and the last departure time. Meanwhile, several probability distribution functions (PDFs) are used to fit the intermediate departure time, intermediate absences for morning, lunch, afternoon, and overtime period. To generate the daily occupancy profile from the fitted CDFs and PDFs, the first arrival time and last departure time are firstly identified using the inverse sampling method. Then, intermediate activities can be determined by comparing the pseudo random number against the intermediate departure PDFs. The durations of the intermediate activities can be obtained by the inverse sampling from the duration PDFs [13].

Markov chain is another stochastic process that has been applied in the occupancy presence modeling. Different algorithms have been explored using both homogeneous and inhomogeneous Markov chains. Those studies include reproducing the Time-User Survey data through integrated Markov Montel Carlo technique [16], modeling the inhomogeneous Markov chain by utilizing the inverse function method [17], and developing a hierarchical approach that combines homogeneous model with the occupants' movement at different locations of offices [18]. There are also several efforts to extend the basic structures of Markov models by integrating more sophisticated statistical techniques such as logistic transformation to a generalized linear model [19], defining the closest distance between two states [20] or blending the transition matrixes linearly using the approximated coefficients by a slot function [20]. For all first-order Markov chain models, the key assumption is that only previous time-step state influences current time-step state. However, in an occupancy prediction scenario, there may be also a habitual sequence that connecting several previous occupancy states together. The studies carried out by Wkike [21], using a higher-order Markov model, provided a preliminary research results on habitual sequence problems. The transitional matrixes were calculated by an integration of the several previous occupancy states that are assumed to contribute to the current transition probabilities. A similar study was conducted to generate the presence distribution in residential buildings based on an integration of the discrete histogram from the presence duration [22]. As previous research studies show that a Markov chain of higher order has more complex calculations and non-bimodal characters than the first-order model.

In summary, the successful implementation of the aforementioned models to accurately capture the occupancy patterns depends on model inputs and level of complexity. Endogenous inputs (e.g. CO_2) other than the occupancy information required by certain type of models could be a solution for commercial buildings (data could be from building automation system) but almost impossible for residential buildings. A complex model could have a limitation on computation time, such as a high order Markov chain model. It normally takes tremendous time and efforts to develop the model for one specific scenario. In this study, authors avoid using a complex model. Instead, authors propose a hybrid approach that integrates basic Markov process with a optimal moving-window for occupancy presence prediction.

3. Methodology

Given the current-state-of-the-art, authors have developed a new approach of short-term predictions of the occupancy profile of residential buildings for specific advanced control purpose. The overall approach is described in Fig. 1. Here, “short-term” means within 24-hrs ahead prediction. Specifically, a new Markov process model has been developed. It is compared to a modified probability sampling method and two machine learning methods, Artificial Neural Network (ANN) and Support Vector Regression (SVR). Occupancy presence data from different residential houses are collected in 5-min interval. The data log for motion sensors entails a sequence of time stamped from the presence (value of 1) to the absence (values of 0). A low pass filter is utilized to generate a time series data in 15-min, 30-min and 1-hour intervals for different prediction window. The occupancy model, developed based on these data, is designed to deliver the necessary inputs for the predictive control design. In the following sections, the formulation of the inhomogeneous chain is described in Section 3.1. The development of new Markov process model is discussed in Section 3.2. Section 3.3 demonstrates a probability sampling model adopted for the residential house case and section 3.4 presents two commonly used machine learning approaches.

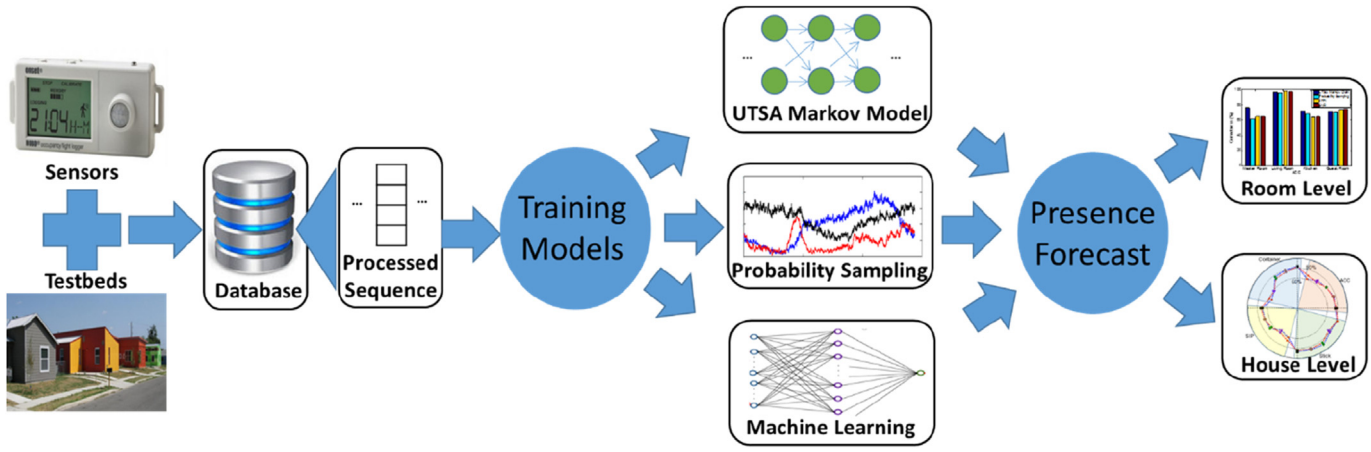

Fig. 1. The overview of the methods and approaches in this study.

3.1. Traditional inhomogeneous Markov chain model

The Markov process, developed for building energy simulation tools or other building applications, is usually in a discrete form with fixed time steps. The states of the Markov chain are not necessarily correlated with environmental variables of the building (e.g. indoor air temperature). The endogenous information collected from the occupancy sensors will be enough as the training inputs. The key assumption to model the first-order Markov chain is that only the instant previous state has the influence on the present state of the occupant (called Markovian property).

Let a Markov chain X at time step k be a sequence containing variables $x_1, x_2, ..., x_k$ and the observed set of occupancy states is $S = \{s_1, s_2, ..., s_n\}$ where $n \leq k$. The chance of the chain to move from the state s_i to the state s_j at time step $k+1$ is decided by the transitional probability defined as:

$$P_k^{ij} = p(x_{k+1} = s_j | x_k = s_i) \tag{1}$$

where $p(x_{k+1} = s_j | x_1, x_2, ..., x_k) = p(x_{k+1} = s_j | x_k)$.

Usually, time inhomogeneity implies changes in the underlying probability of the transition between the same pair of the states as time goes on:

$$p(x_{k+1} = s_j | x_k = s_i) \neq p(x_k = s_j | x_{k-1} = s_i) \tag{2}$$

The transition probabilities between states for more than one step is more easily to be calculated by a transition matrix. Let $P = (p_k)_{i,j}$ denote the matrix where each element at index (i,j) represents the probability defined in Equation (1). Suppose that the probabilities are fixed when they are not influenced by any other factors in the current time step, the transitional matrix is defined as:

$$P = \left(p_{ij}\right)_{n\times n} = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0n} \\ p_{10} & p_{11} & \cdots & p_{1n} \\ \vdots & \vdots & & \vdots \\ p_{n0} & p_{n1} & \cdots & p_{nn} \end{bmatrix} \tag{3}$$

where $\sum_{j=1}^{n} p_{ij} = 1$ for any $0 \leq i \leq n$.

The transition matrix trained for the Markov chain in this study uses the maximum likelihood estimation (MLE) with a moving window optimization. In essence, the moving window is a trimmed window covering the sequence $W = \{x_t, x_{t+1}, ..., x_s\}$ in the Markov chain X where $0 \leq t < s \leq k$ for the prediction of state x_{s+1}. Given one moving window where there are n_{ij} pairs of the states' sequence $\{s_i, s_j\}$ in the all pairs of the sequences $\{s_i, s_l\}$ for $t \leq l \leq s$, the transition probability estimated by MLE is:

$$\widehat{p}_{ij} = \frac{n_{ij} + \alpha}{\sum_{l=1}^{k}(n_{il} + \alpha)} \tag{4}$$

where α is a smooth factor $(0 < \alpha < 0.1)$. Owing to the limitation of the window size, the occupancy may enter a "sink" state with an extremely small probability of transition. The smooth factor is used to avoid the "sink" state to appear in the stages of estimations.

3.2. New Markov chain model

In this study, the authors propose to integrate the moving-window strategy to estimate the transitional probability matrix and utilize the model for prediction with a change-point analysis. The authors try to answer two questions: 1) at which time step should be the horizon of the moving window change; and 2) how long should the historical data be chosen in one horizon of the window. Let $D = \{d_1, ..., d_{24\times z}\}$ represents the all selectable historical data before the state that need to be predicted. Here, if the occupancy presence state to be predicted is in a working day, the selection of D only contains the available profiles of z working days. Regardless of the occupancy level, D is processed into a data set containing only the presence and absence as 1 and 0. A discrete profile of the presence probability in daily scale is generated by:

$$P_m = \frac{\sum_{j=1}^{z}\left(\lambda^{z-j} \cdot d_{(j-1)\times 24+m}\right)}{z} \tag{5}$$

where $1 \leq m \leq 24$ if the occupancy data is in hourly scale and λ is an exponential forgetting factor, which is below 1. Without forgetting effect, the data of all periods are treated equally with regards to the information that such data contains and to generate the distribution of presence. An exponential forgetting could maximize the penalty on the older information of the data and allow the presence probability to retain the most recent information only.

Change point detection is implemented to check at which time step in a daily profile of the set D that an occupancy presence distribution is changed. The assumption is that a change of the moving window should happen based on the change of presence distribution on a daily scale. The detection algorithm in this study used relative density-ration estimation with the Pearson divergence as a divergence measure to score the possible change points.

For a subsample m selected from the distribution n, the symmetrized divergence score is defined as follows [23]:

$$\int p_\alpha(m)\left[\frac{p(n)}{p_\alpha(m)}-1\right]^2 d(m)+\int p(n)\left[\frac{p_\alpha(m)}{p(n)}-1\right]^2 d(n) \quad (6)$$

where $p_\alpha(m)=\alpha p(n)+(1-\alpha)p(m)$, p is the probability density function of the corresponding variables, and the factor α is a smooth factor to the plain density ration.

Four main patterns can by basically extracted and recognized from occupancy presence data: long absence, low presence rate, high presence rate and long presence. They can be attributed to each of the classified windows. During prediction, each of the state in the next day after the set D is assumed to have the same changing points that are estimated from the presence distribution of the set D. The tuning of the training data, as an answer to the second question, in a moving window depends on the prediction horizon. In an extremely short-term forecast case, models are commonly trained with within one moving window (e.g. 15-min, 30-min, and 1-hour ahead cases). The size of the window is determined by a contingency table with a leave-one-out validation. The validation is performed on two sub sets of data in five most recent working days. Commonly used 10-folder validation is not suitable in this case due to the limited length of window size. Each tuning length within the horizon is assigned a score that added the true positive value and true negative value from the contingency table. The highest score represents a suitable candidate of the tuning length of one moving horizon. In 24-hour ahead prediction, it is impossible to access the intraday information to calibrate the inhomogeneity. Hence the assumption is that two consecutive days have a similar pattern. The predictions thus are simulated in a daily scale where the full horizon of each moving window classified by Eq. (6) is used as the tuning length in that moving window. A more detailed algorithm to predict the occupancy is shown in Fig. 2.

Prediction set $S \leftarrow \{0,\ldots,0\}$ in a length of the day duration
Predict step $t \leftarrow 1$
Historian Training Set H in a length of n historical days
for day i **do**
  $P \leftarrow$ the presence profile calculated by **Eq.(5)** from H
  Score $\leftarrow$ calculated score from **Eq.(6)** by looping P in fixed length
  change points $\leftarrow$ the local maximum points in *Score*
  windows $\leftarrow$ the classified windows decide by *change points*

  for t in one moving window from *windows* **do**
    if Predwindow=1 **then**
      validation set $\leftarrow$ most recent five working days before day i
      contingency table $\leftarrow$ perform leave-one-out validation of 2 folders in each day of the *validation set*
      $l \leftarrow$ the length of the trained set with best prediction score calculated by the *contingency table* for the leave-out set.
      $m \leftarrow$ the transitional matrix calculated by MLE of **Eq.(4)** from t-l to t-1 steps
    else if Predwindow>1 **then**
      $l \leftarrow$ the length of the current moving window.
      $m \leftarrow$ the transitional matrix by MLE of **Eq.(4)** from t-l-24 to t-24 steps (same period from yesterday)
    end

    $r1 \leftarrow$ random sampling form a uniform distribution
    if Predwindow=1 **then**
      $s \leftarrow$ the occupancy state from t-1
      $S(t) \leftarrow$ comparing $r1$ with the corresponding probability that describe the transition from s to other state in m
    else if Predwindow>1 **then**
      if t=l **then**
      $s \leftarrow$ the occupancy state from t-1
      $S(t) \leftarrow$ comparing $r1$ with the corresponding probability that describe the transition from s to other state in m
      else if t>1 **then**
      $s \leftarrow$ the occupancy state from $S(t\text{-}1)$
      $S(t) \leftarrow$ comparing $r1$ with the corresponding probability that describe the transition from s to other state in m
    end
end

Fig. 2. The simulation diagram of the first-order inhomogeneous Markov chain.

3.3. Probability sampling model

This approach applies the random sampling process on the data by assuming that it does not have the Markov properties. Currently, the probability sampling method has been validated in certain level to predict the states of the presence [24,25]. However, these studies are limited to the office buildings. In this study, the authors will adopt the popular methodologies designed for office environments to predict the occupancy presence in the residential samples.

The model is developed mainly based on the probability profile of the historical presence. The prediction is made by inverse sampling during the presence periods. The simulation algorithm only depends on the profile of the presence probabilities generated by Eq. (5). For each time step of the day to be predicted, the occupancy state is decided by comparing the presence probability at that time step from the profile with a random number drawn from the uniform distribution. The room is considered to be occupied only if the number is smaller than the presence probability. The algorithm to predict the presence using the random sampling is shown in Fig. 3.

3.4. Machine learning approaches

Machine learning is a black box approach. It is usually compared to the “white” model, such as the stochastic model, where each probability can be interpreted by the occupancy presence rate. It utilizes advanced computational learning algorithms as the “artificial intelligence” to learn patterns from the data set. Various empirical and theoretical studies have proven the capability of this approach for different kinds of applications [26,27].

Two common methods are applied and tested here: Artificial Neural Network (ANN) and Support Vector Regression (SVR). For ANN, feed forward neural network (FFNN) of a single layer and a double layer configuration are explored. However, due to the overfitting problem, good forecasts are not found from the double hidden layer structure [26,28]. Neurons calculate the weights sum of the inputs and produce the output by transfer functions as follows:

$$f(x)=\sum_{j=1}^{N} w_j \varphi_j\left[\sum_{i=1}^{M} w_{ij}x_i+w_{io}\right]+w_{jo} \quad (7)$$

where w is the weights for input, hidden, and output layers, x is the training input, N represents the total number of hidden neurons, M represents the total number of inputs, and M represents the transfer function for each hidden neuron. In this paper, FFNN is modeled as 1 hidden layer with 20 neurons, 1 output neuron (the prediction of presence state), and multiple input neurons (several time lagged values of historical occupancy states depending on the forecasting window). The transfer functions are the hyperbolic tangent sigmoid functions for the input layer, and the linear transfer function for the output layer. Hidden layer weights in

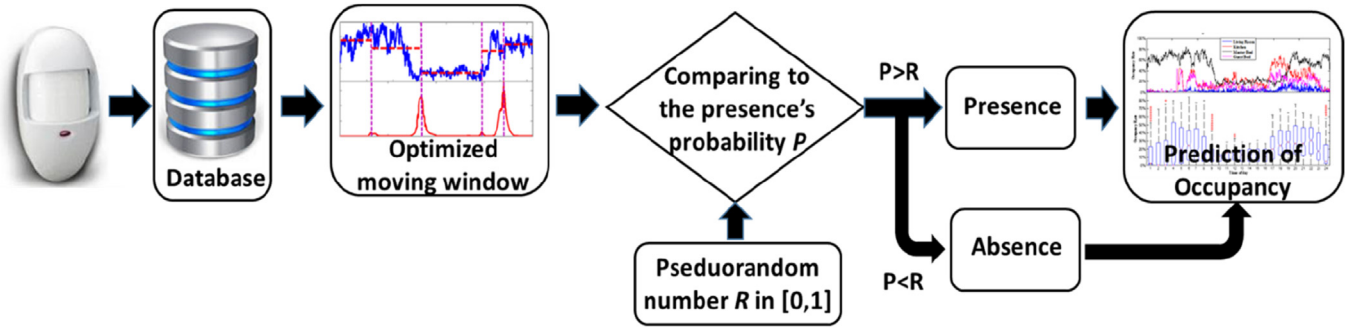

Fig. 3. Predictions of the occupancy using the probabilistic model.

Equation (7) are learned from Levenberg-Marquardt back-propagation algorithm [29]. Model validation is performed on a holdout set of the data using the criteria of the mean squared error, where the training set contains at most 70% of the input set.

For support vector regression (SVR), the SVR approximates the inputs and outputs using the following form.

$$f(x) = w\phi(x) + b \tag{8}$$

where $\phi(x)$ represents the transfer function and parameters w and b are estimated by minimizing the regularized risk function:

$$\min \frac{1}{2} w^T \cdot w + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right)$$

$$\text{s.t. } y_i - w^T \phi(x_i) - b \le \varepsilon + \xi_i \tag{9}$$

$$w^T \phi(x_i) - b - y_j \le \varepsilon + \xi_i^*$$

where n represents the total number of training samples, ξ is the error slacks guaranteeing the solutions, C is the regularized penalty, and ε defines the desired error tolerance. LibSVM [30] is used in our study. A radial basis function of 3° with weighting factor γ is used. The authors find the SVR model is relatively insensitive to the value of ε smaller than 0.01 whereas both C and γ necessitate independent tuning. These parameters are determined by 10 fold cross-validation based on mean square error. The grid search scale for C and γ is maintained among the range from 10^3 to 10^{-3}.

The training process of the two methods is facilitated by testing different configurations of the inputs from the historical presence information. The input set for the 15-min, 30-min and 1-hour ahead windows, defined as H1, is a Markov order 4 sequence:

$$H1 : f(O_{t-1}, O_{t-2}, ..., O_{t-4}) \tag{10}$$

where O_{t-1} represents the occupancy presence from the previous one time step back, …, and O_{t-4} represents the occupancy presence from the previous four time steps back.

Input set H2 is used to forecast the next 24-hour ahead occupancy presence of the current time step O_t:

$$H2 : f(O_{t-24}, O_{t-48}, O_{t-72}, .., O_{t-168}) \tag{11}$$

For one time step ahead forecast (the 15-min, the 30-min and 1 h ahead forecast), inputs include the historical occupancy presence from 1 to 4 time steps back. By comparison, 24-hour ahead case needs the historical occupancy at the same time from yesterday, the day before yesterday, …, and the day before one week. These features are selected based on an exhaustive search by minimizing the coefficient of determination.

4. Results

4.1. Description of testbeds

The occupancy data in this research are collected from four houses in west side of San Antonio, as shown in Fig. 4. The four samples are single-family dwellings around 110 m^2 each. Houses are named according to construction materials: SIP (Structure Insulated Panel), ACC (Autoclaved Aerated Concrete), Container (Steel Container), and Stick (Wood). They are leased and operated mostly by part-time workers and low-income people. The presences of occupancy are detected at 5 min intervals from over 30 sensors for all the rooms including kitchen, bathroom, living and bedroom areas during the year of 2014. The occupancy detection sensors are passive infrared sensors. Sensors are attached under ceiling in the middle of each room. Monitoring data are stored in the on-board memories. All the collected data are further exported to SQL database. To pre-process 5-minute data to other time intervals, presence counts are processed using moving average filters of Savitzky-Golay algorithm. To retain a realistic pattern, only 0.95 factor is used in each time interval. The following testing periods are selected for occupancy modeling and prediction: ACC was modeled from Sep 17th to Oct 31st. Container was modeled from May 21st to July 31st. SIP was modeled from Jan.1st to Apr.30th. Stick was modeled from Jan.1st to March.31st. All periods are in Year 2014 and only include weekdays.

Fig. 4. Four test houses.

The average rate of the presence was plotted for the monitored rooms of the four samples, as shown in the upper part of each figure in Fig. 5. All four houses demonstrate significant differences of the presence pattern at the room level. However, the similarity is revealed in a cross-sectional comparison. Master bedrooms are mostly occupied during the night. Living rooms or kitchens are occupied mostly around afternoons and evenings. The variances from all rooms' presence rate of the individual house are presented in the lower part of each figure in Fig. 5 using 1.5 interquartile range

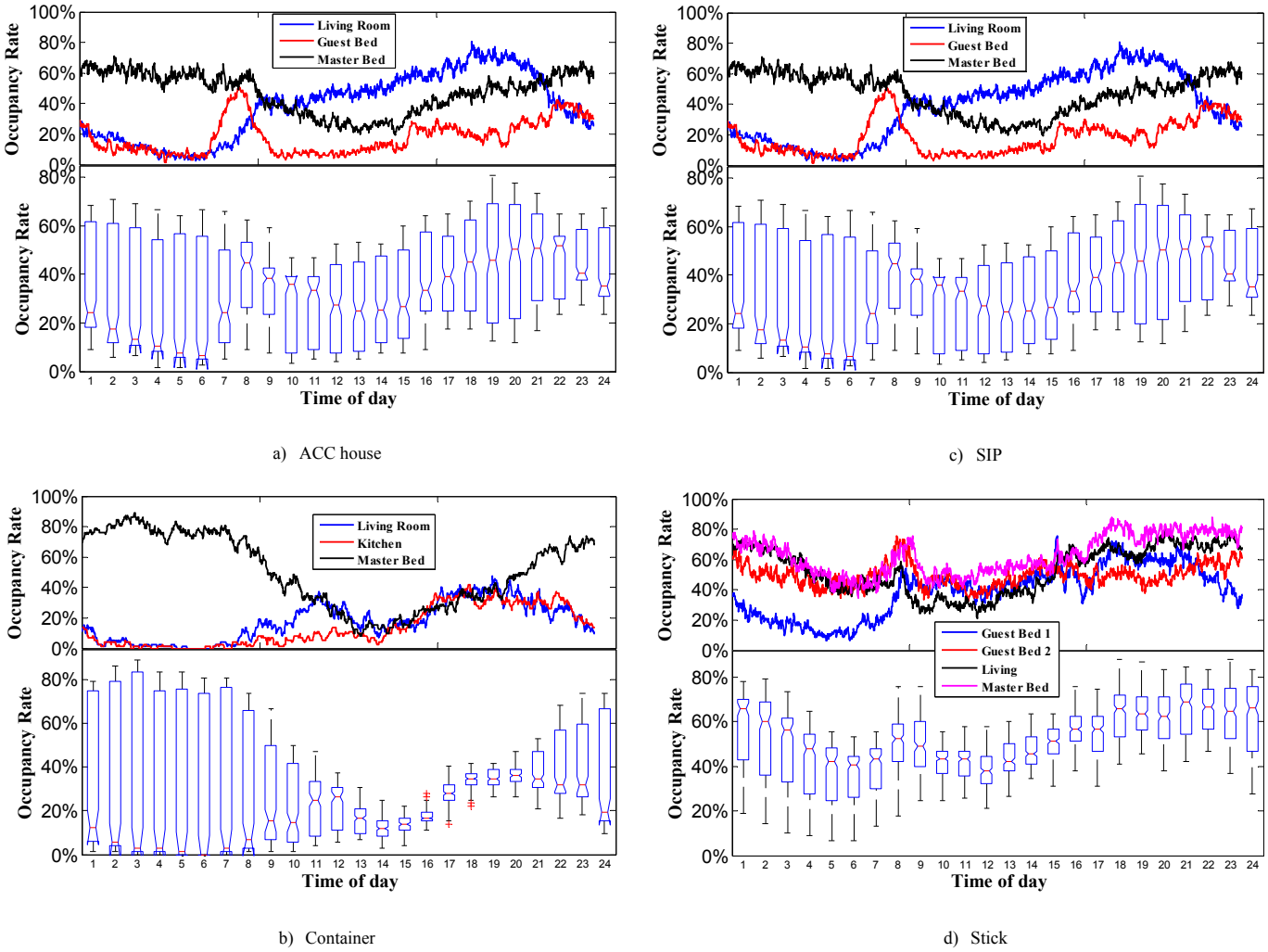

Fig. 5. The presence rate of the rooms in the four houses.

(99% confidence interval) in box-and-whisker plots. Large variances are observed in first three houses while the fourth one has a less varied pattern. More details can be referred to the study of those samples [8].

Further analyses reveal the consistent patterns at the house level, as shown in Fig. 6. It shows that until 10 a.m., most of the residents in the houses (except SIP) have the presence probabilities close to 100% of sleeping, where there are large variances of occupancy presence during the daytime. The ACC's occupant has a probability of 70% to leave the house between 10 a.m. and 4 p.m. and back after 5 p.m. A more gradual ramp-up and ramp-down is observed in the Container house between 10 a.m. and 6 p.m. SIP's resident is a part-time worker who works or leaves during the night explaining the lack of presence during the night while Stick's family has dependents at home all day, explaining the high occupancy during the daytime. It is also clear that the variances of the presence rate at the house level in Fig. 6 is significantly smaller compared to those at the room level in Fig. 5. Based on these analyses, two assumptions are made for modeling this specific data set: 1) for each day of working days in Fig. 6, less variance (mostly below 20%) is observed and thus training does not need to differentiate individual day types such as Monday or Friday; 2) the modeling from the house level rather than the room level will be acceptable for occupancy presence forecast if they maintain a similar prediction accuracy.

4.2. Model performance at the room level

The key to utilize the stochastic models depends on the optimization of the moving window as mentioned in Section 3.2. One example of the changing points between the windows is shown in Fig. 7 for the prediction of occupancy presence at ACC's master room on Oct 15th. The normalized score is calculated based on a forgetting factor of 0.8 with a span of all the historical records before the date. Based on the analysis of historical data, there should be five windows for prediction of that specific day. The first one starts from 12 a.m. until 7 a.m., the next one ends around 10 a.m., the third one ends around 6 p.m., the fourth one ends around 8 p.m., and the last one lasts till the end of the day (12 a.m.). The predictive performance of the models is evaluated based on the correctness of the occupancy predictions in terms of the occupied and unoccupied states. The correctness is used in previous studies [13,19]. In summary, there are two predicted classes: presence and absence. Of the total l predictions, if there are m predicted presence when the observations of the rooms are occupied and n predicted absence when the observations of the rooms are not occupied, the overall accuracy is thus calculated as a percentage: $100 \times (m+n)/l$. All results of the stochastic models' predictions for the individual rooms are presented in Figs. 8–12 for 15-min ahead, 30-min ahead, 1-hour ahead and 24-hour ahead respectively.

Comparing the plots with Fig. 5, the presence can be more

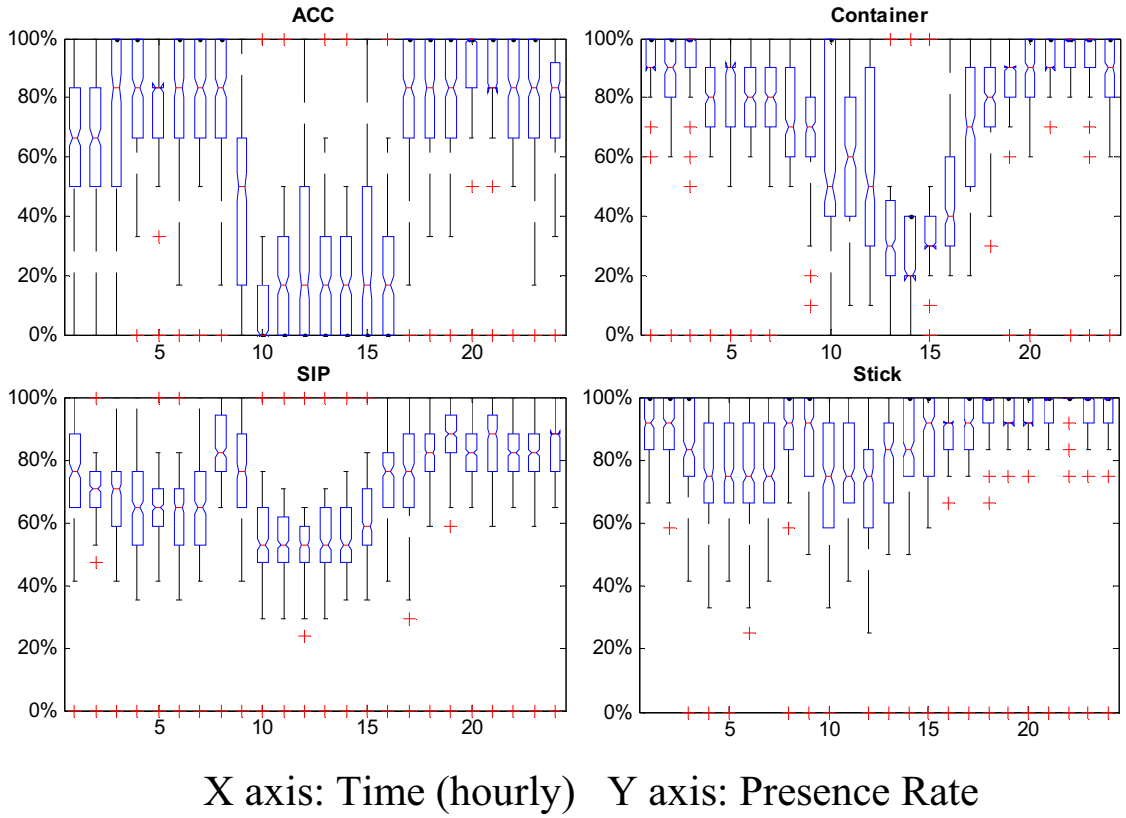

Fig. 6. Presence rate of the house profiles during weekdays.

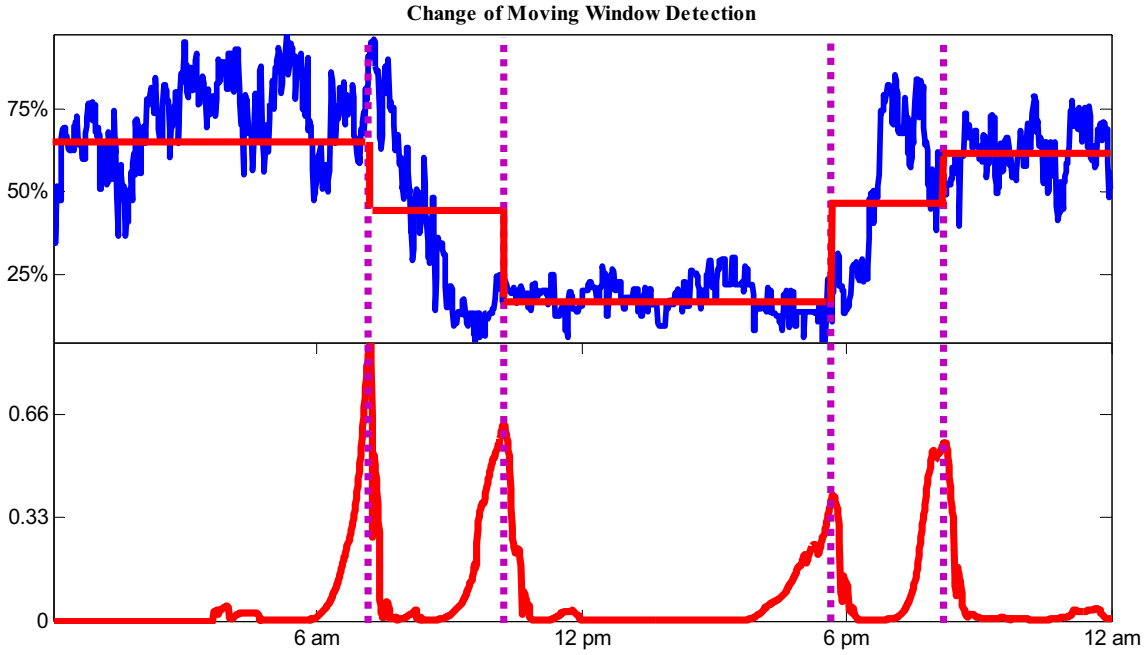

Fig. 7. Change points detection for the moving windows.

accurately predicted (>75%) in the extremely short-term forecast (e.g. 15-min till 1-hour) for the Markov model if the presence rate is smooth enough. For example, the presence rate of Container compared to other samples does not have the small spikes observed consistently. The predictive power of the model is also correlated with the variances. The examples are the living room of ACC (blue line) in (Fig. 5a) and the guest bed 2 of Stick (red line) in Fig. 5d). The living room may be a special case owing to the extremely low presence rate (<20%) which represents an absence dominated pattern. In contrast, the guest bedroom 2 with a

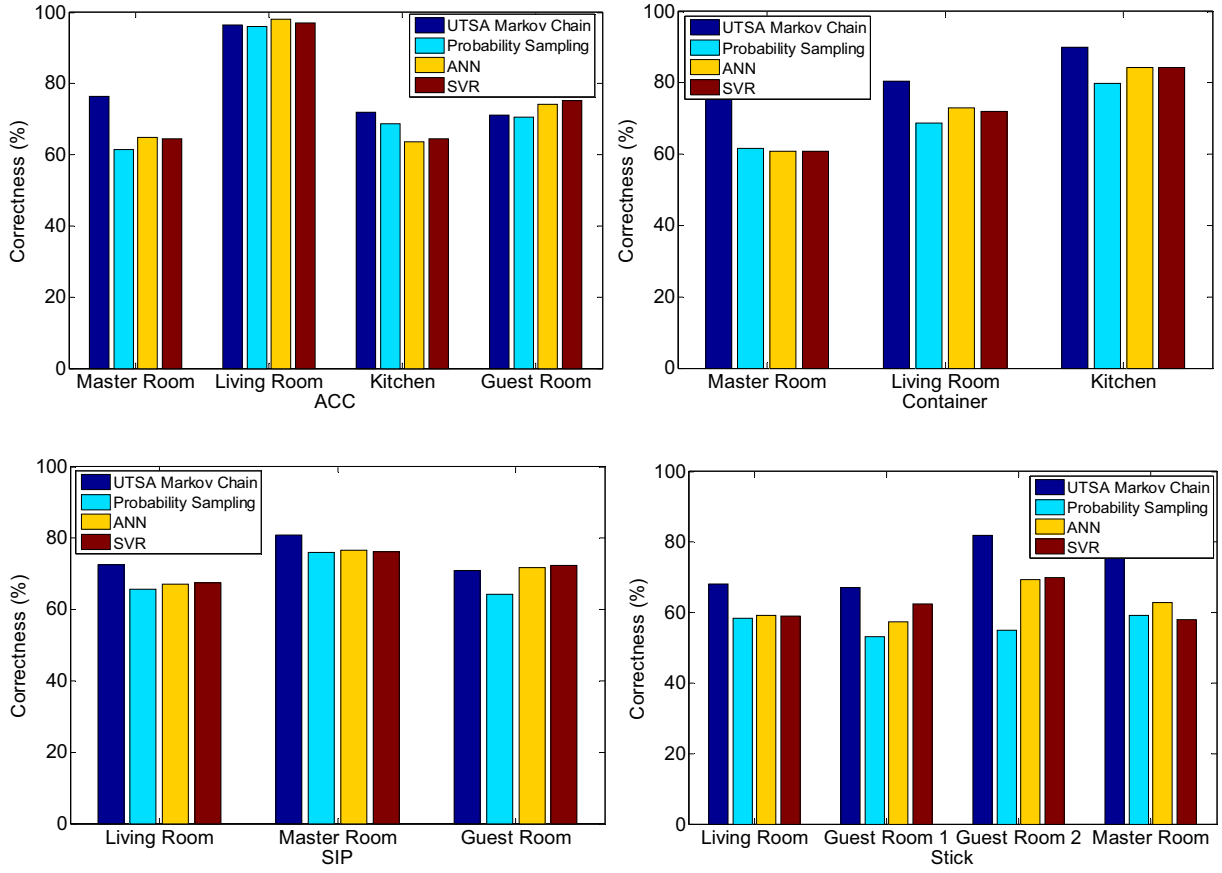

Fig. 8. Presence predictions for the residential houses (15-min ahead in 15-min resolution).

persistent presence (between 40% and 60%) can be interpreted as a stationary type, where the resident leaves or enters their room with a regular schedule. Another case that is a combination of reasonable smoothness and variance is the guest room of SIP (red line) in Fig. 5c) that still can reach 80% of accuracy. The only variance is observed between 6 a.m. and 9 a.m. which is up to 50%. Although the similar findings can be claimed for the probability sampling models, the average accuracy of prediction for each room is much lower than the Markov model. In contrast, ANN and SVR tend to provide comparable performances and even better in some cases (e.g. one guest room of ACC). For 24-hour ahead predictions, there are no significant differences in terms of accuracy among all four models. It is mainly because all models are based on the assumption that each day's presence pattern should be similar. This kind of assumption actually could be a drawback for a more stochastic sample. Only a few exceptions existed in Fig. 11 and Fig. 12 where they have more than 75% correct predations. This is one of the limitations of this study which is discussed in the conclusion part.

4.3. Model performance at the house level

The model performance at the house level is more important for applications such as smart control on thermostats. In this case, occupancy presence can be predicted and derived in two ways: 1) aggregates the room-level predictions to generate the prediction for the house-level, and 2) processes the data to a house-level first and then directly predict the occupancy status. The results of both ways are presented in Fig. 13 for all samples. Regardless of models, forecasts for the individual house have not much differences from the room level. The blue lines (the house level) and the red lines (the room level) are very similar. However, individual house does have different predictive potentials, although they are bounded within 60–80% correctness (two circles bounded the blue and red lines in Fig. 13). Another error criteria is called receiver operating characteristic (ROC) scores which is based on the true positive rate against the false positive rage. It is further presented in the Appendix.

The probability sampling model is improved in this case. This could be explained by fewer noises in datasets at the house level compared to the room level. Meanwhile, the Markov model is still expected to have a promising performance from 15-min to 1-hour forecast (the square, the round and the diamond shape labels in Fig. 13). Fig. 6 shows that the samples can be categorized as four different types: the single-square shape (ACC), the single-valley shape (Container), the twin-valley shape (SIP) and the flat shape (Stick). By ranking the overall accuracy of the individual house's predictions from Fig. 13 (the red and blue dashed lines), it can be concluded that the shape of the presence rate (Fig. 6) does not necessarily correlated to the predictive capabilities of the models (Fig. 13). The best case is Stick house, where most predictions are more than 80% of accuracy (the blue and red dashed curves at all the lower right quarters of each error polar plot in (Fig. 13). The next case with a similar mean and variance (50%–90% in Fig. 6) is SIP. The Container house has more variety, which is from 20% to 100% as shown in Fig. 6. In general, results from a Markov process model are similar to the probability sampling. However, Container house has the worst performance by comparing the accuracy curves (the blue and red dashed lines) in the left-upper quarter of the polar plots of Fig. 13a) and b). As shown in Fig. 13a), the prediction accuracy can achieve near 80% (meaning correctness of 80%), which demonstrates the accurate predictions made by the proposed Markov model.

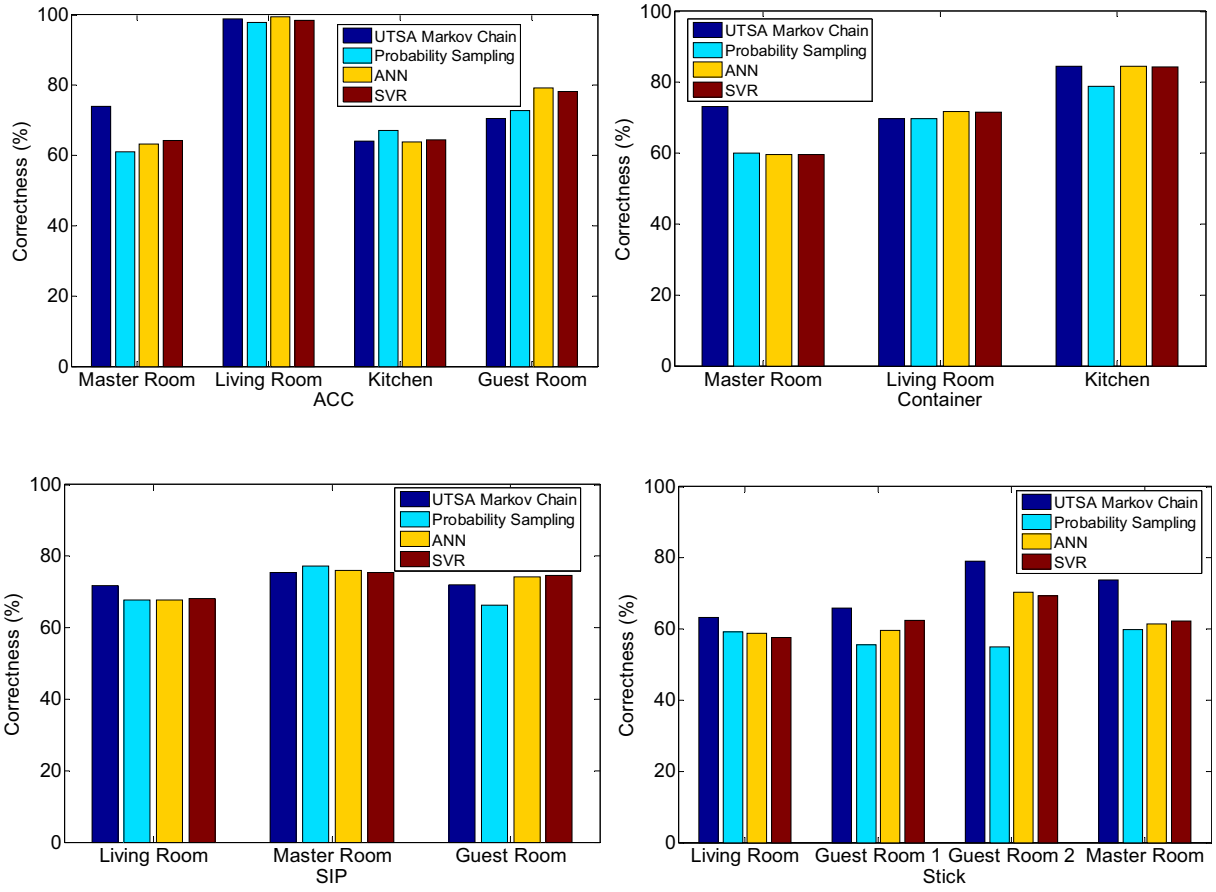

Fig. 9. Presence predictions for the residential houses (30-min ahead in 30-min resolution).

A further comparison among all accuracy curves (blue and red dashed lines) of Fig. 13b–d) can conclude the similar performances for all samples forecasted by the probability sampling, Artificial Neural Network and Support Vector Regression. The probability sampling and the machine learning approaches have stable forecasting performance regardless of the forecasting windows, as shown in b), c) and d) of Fig. 13 comparing to a) of Fig. 13, with more rounded and smoothed closed curves. However, the total area of the closed curves (both blue and red ones) in b), c) and d) of Fig. 13 for the other approaches are smaller compared to the proposed Markov chain, which shows the irregular but bigger closed curve in a) of Fig. 13. This indicates the better performances of the proposed Markov process. However, ANN and SVR perform slightly better in 24-hour ahead cases, as shown in Fig. 13 with the triangle labels. 24-hour ahead forecasting are conducted based on two different time step resolutions (15-min and 1-hour). The reason is that other inputs such as weather, electricity price, and load forecasting may have varied sampling frequency during predictive control design.

In summary, for extremely short term forecast from 15-min to 1-hour ahead, the Markov model is recommended while the machine learning approaches are suggested only for 24-hour ahead forecasts. The probability sampling model needs further improvement to improve the performance. It is also noticed that the house-level modeling is more convenient compared to the room-level modeling since there are not much difference between the accuracies for all the four methods in different spatial resolution forecasts shown in Fig. 13. The room-level modeling not only bring more samples (each room occupancy) to be processed, but also contribute more stochastic patterns (as shown in Figs. 5 and 6) need to be modeled.

5. Discussions

Currently, only a few studies in residential buildings focus on occupancy models [31–34]. They provide estimations of occupancy profiles using the Time Use Survey data. Individual occupancy profile at building level can be derived from the national survey and used for single houses [31]. However, studies based on such data represents an averaged stochastic pattern because TUS data are usually reported in terms of the average occupancy in a specific social-economic group of the population [32]. In addition, most models used in such studies solely depend on a standard Markov modeling process that integrate with Monte Carlo technique or Cross Validation to enhance the performances [33,34]. As discussed in Section 2, the current state of the art for more accurate occupancy modeling requires hybrid or improved model rather than basic Markov process. In this study, the authors use real-time measured data and develop a new method to predict occupancy presence in residential buildings. Advantages of the proposed model comparing to other approaches are: 1) more accurate forecast for one-time step ahead (up to 1 h) of the occupancy presence, 2) competitive performances to the current-state-of-the art day-ahead occupancy modeling, and 3) the ability to adapt to the large variance change of the occupancy pattern in both the room level and the house level.

As discussed in section 4, results of various prediction performance for each residential house stem from the fact that every presence profile of an occupant in houses is fundamentally different. There is not a single method could be the best among all possible cases (Fig. 13). One popular model commonly used for an office environment, the probability sampling, presents difficulties

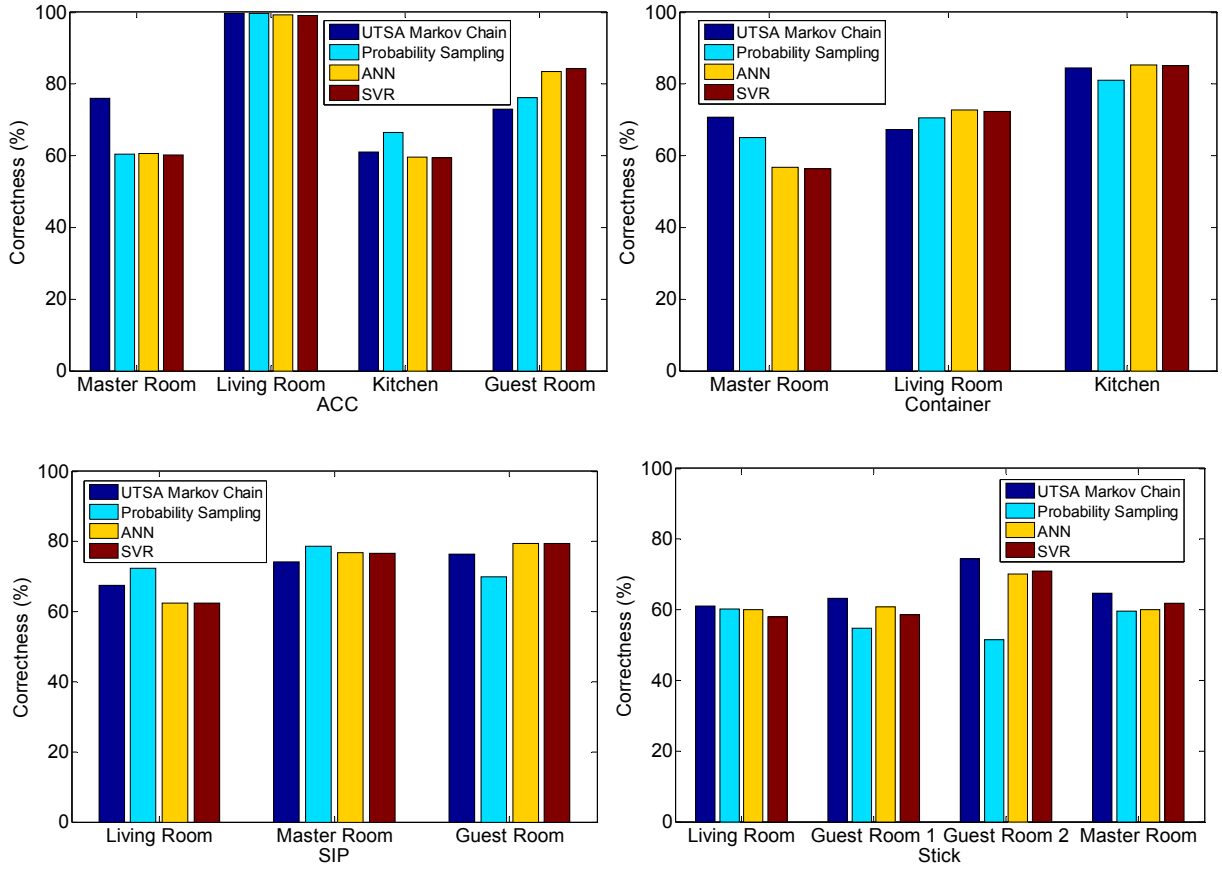

Fig. 10. Presence predictions for the residential houses (1-hour ahead in 1-hour resolution).

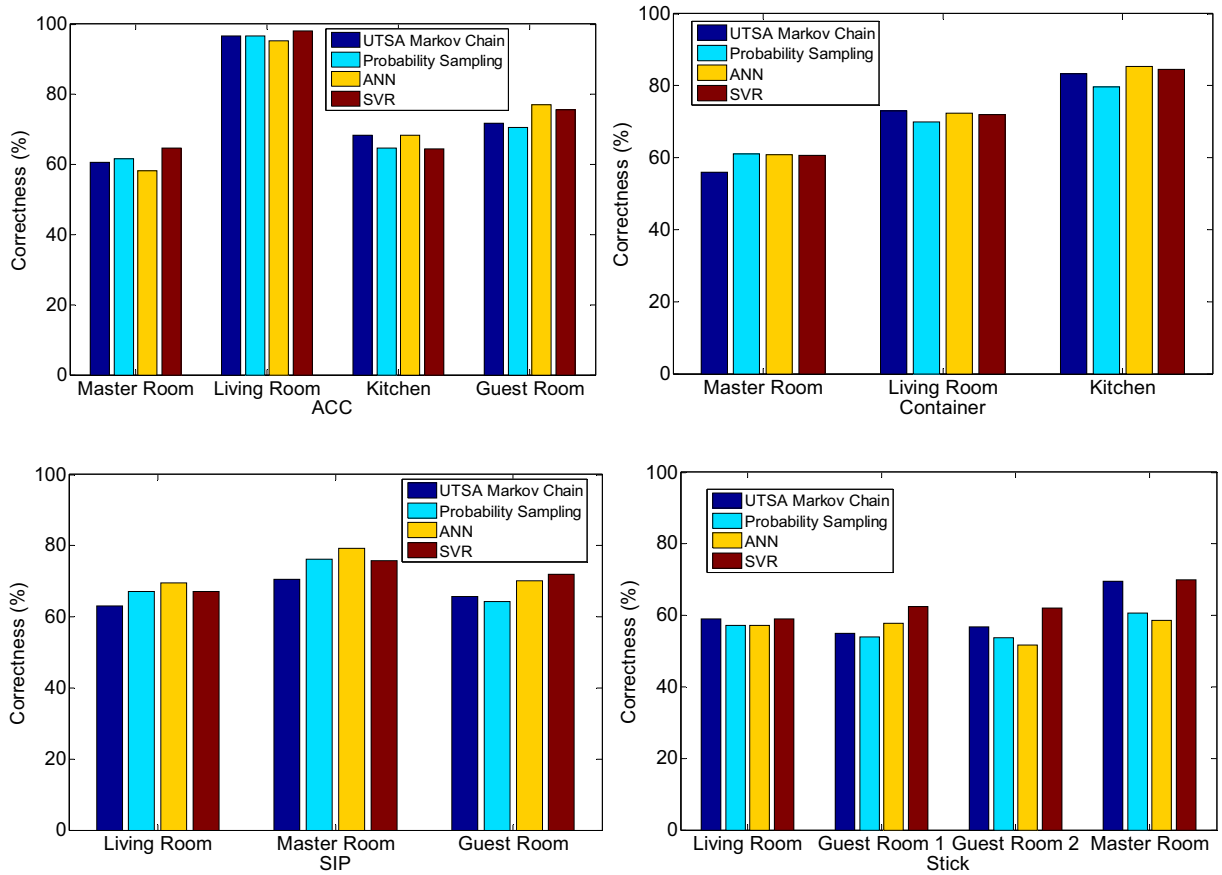

Fig. 11. Presence predictions for the residential houses (24-hour ahead in 15-min resolution).

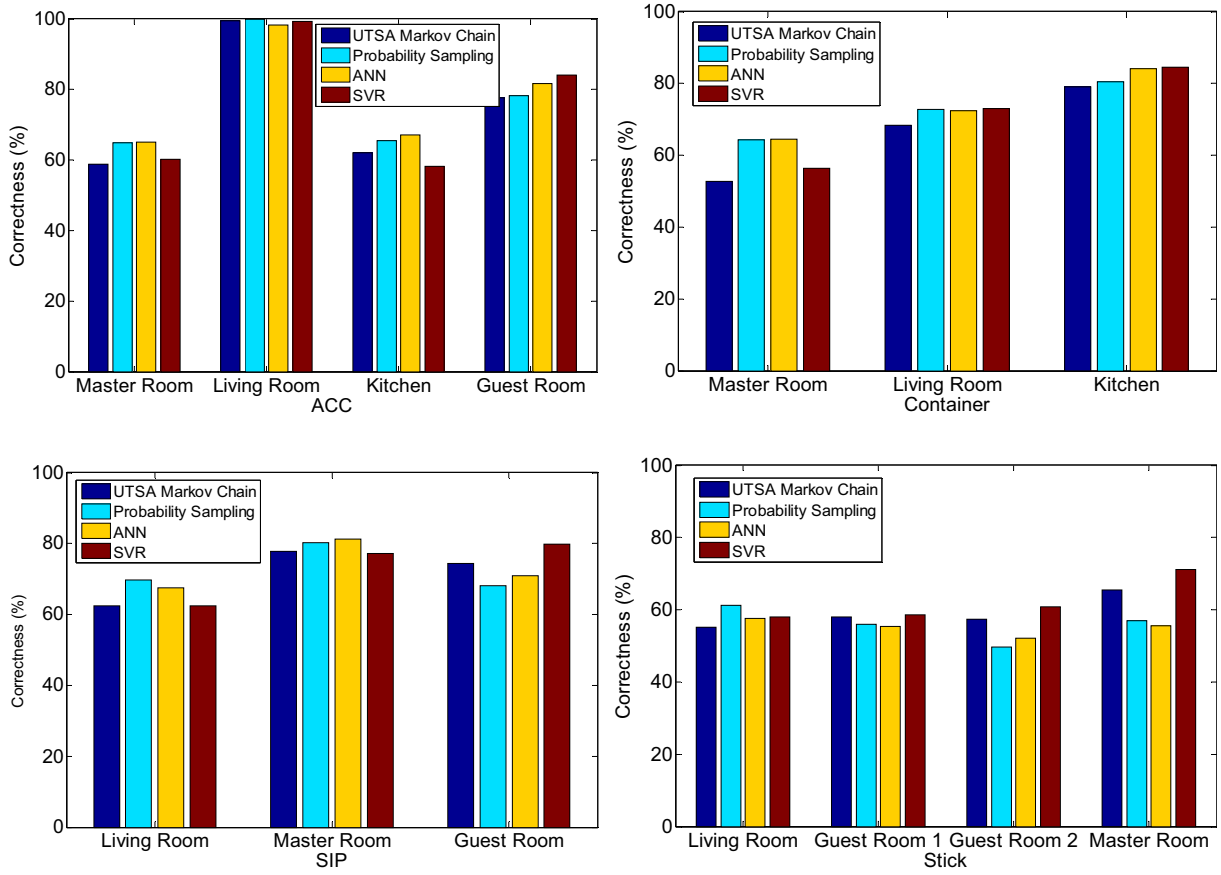

Fig. 12. Presence predictions for the residential houses (24-hour ahead in 1-hour resolution).

to adapt to the diversity of the residential occupancies at the room level (Figs. 8–12). From the literature review in Section 2, it is noticed that the main advantage for most of the building occupancy studies to adopt the probability sampling model is not because of better accuracy but the easier integration with other tools for total building performance simulation. The sampling model in the office studies are normally developed for estimations of the first arrival time, the last departure time and the intermediate departures. It is unlikely to adopt the same strategy to the residential buildings because there is no regular arrival time, departure time or intermediate absence (e.g. the office breaks and meetings) in most single-family houses.

Another issue is the expected accuracy for occupancy behavior prediction used for different applications. The performance evaluation of the occupancy models for building energy simulation is different than ones for building controls. The model accuracies from predictions rather than estimations are at best individually claimed and verified for finite samples [13,35,36]. By far, there are several potential performance matrixes to measure the predictive power of the occupancy models for office occupants: the first arrival error, last departure error, the occupancy state one-to-one matching error, the number of transitions error, the duration of the intermediate presence and the duration of the intermediate absence [13]. However, it is unlikely to adopt those criteria for the residential samples. Therefore, only one of the errors mentioned above, the occupancy state one-to-one matching error, is used in this study. One recent study reported that the 80th percentile of the matching errors for a one year period of a single worker's office [13]. For three tested methods, the errors are spanning from 0.45 to 0.48, equal to 45%–48%. Comparing to the same prediction horizon (24-hour ahead in 15-min resolution) in this study, the results from the residential tests actually have a decent higher accuracy. This can be explainable by the success of the modification of the methods. Further studies can be investigated in a longer and general data set.

The last important aspect needs to be considered is the temporal difference of the forecast window for the occupancy presence in various applications. In other research domains, the accuracy of the models' predictions could be improved by changing the window of the forecast [26]. A more recent study to predict the occupancy level of the office workers draw a similar conclusion [36]. However, in this study, no significant changes of prediction accuracies are observed for most samples when the prediction horizon increases from 15-min to 24-hour ahead. For the smart buildings, the temporal changes of the occupancy models actually have less influence on the smart controller like Nest [37]. Those advanced interfaces not only record occupancy presence and the human building interactions from sensors, but also analyze the preference of occupants. This advanced control strategy diminishes the stochasticity of users' overrides and increase the predictive power of the occupancy models. Although an even higher resolution of the occupancy monitoring, such as one minute interval, could be used to improve model performance. The control algorithms will instead have a more frequent track to the occupancy model. Such frequent responses from occupancy-based controller can highly violate the operations of the systems. Unless the occupants are extremely insensitive to the comfort changes, the predictive performances and control difficulties should be equally addressed in a relaxed forecast window, namely 15-min, or even hourly scale.

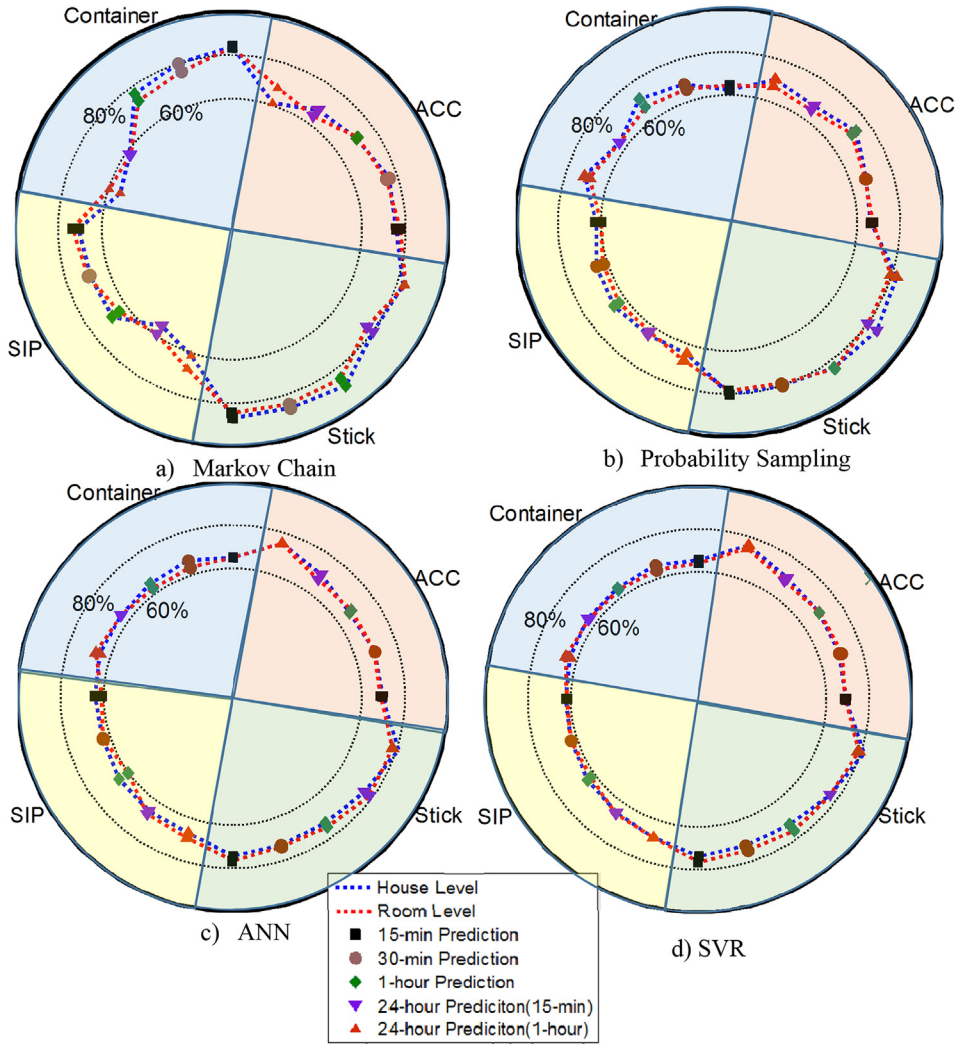

Fig. 13. Comparison between models based on the modeling level.

6. Conclusion

This paper aims to develop and demonstrate an innovate approach for residential occupancy presence forecasting. By predicting future occupancy presence of different time scale (15-min to 24-hour ahead), the proposed Markov model demonstrates its predictive power specifically for the purpose of building control applications. The results are validated through measured data from the field tests of the residential houses and compared to other commonly used methods and models for occupancy presence predictions such as the probability sampling, Artificial Neural Network and Support Vector Regression. The final results show that the proposed Markov model outperforms the other methods in terms of an average 5% correctness. Maximum difference of 11% in one time step ahead forecast (15-min ahead) is observed for the occupancy prediction of samples with large variances. In 24-hour ahead prediction, not much differences could be found among the models. Implementing such kind of occupancy model will be a solution for characterizing the large dynamics existing in residential occupancy patterns and help buildings to optimally control the energy devices. This study observes a relatively lower performance in 24-hour ahead prediction cases compared to the other prediction windows (e.g. 15-min to 1-hour ahead). It is challenging to improve the forecast accuracy in this case even with the changes of temporal resolution (sampling rate) from 15-min to 1-hour. However, the results show competitive performances compared to recent studies [13].

The limitations of this study includes: 1) Potential high computational cost. The proposed method integrates a change point analysis looping all the data in the moving window. The optimization could become slower if the data pattern becomes more stochastic. The situation may become worse if longer period of training data is used. However, if the prediction horizon window of the predictive control design is around 15-min, the model developed in this study could have a great potential in implementing online through increasing forgetting factor during training process. 2) Limited data to investigate seasonal factors. Due to privacy issues, data collection becomes extremely difficult. Often, we do not have a continuous data set across a whole year. Hence the seasonal or other time-related factors cannot be identified. 3) The generality of the developed model. Through this study, we cannot conclude the generality of the developed model again due to limited data. The purpose of this study is to propose and test a new model with limited data and vent its prediction capability. In the future, we will test our models when more data is available.

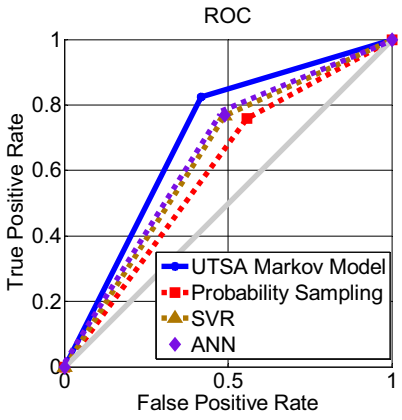

Fig. 14. ROC of 15-minute ahead prediction for ACC at house level.

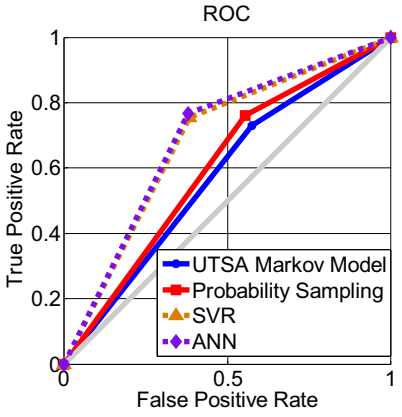

Fig. 15. ROC of 24-hour ahead prediction for ACC at house level.

Appendix

The result in the appendix shows the model performance based on receiver operating characteristic (ROC) curve based on data from one house (ACC). The ROC scores are plotted base on the true positive rate against the false positive rage. The ROC score is calculated as follows:

$$ROC = \frac{\sum TP}{\sum FP} \tag{12}$$

where *TP* represents the true positive and *FP* represents the false positive. Since the study predicts binary data, single ROC score point is plotted. Lines are added indicating the deviation of the score pints for each methods from the diagonal (which is the 50% line). The further the RCO scores line deviate from this diagonal line to the upper left space, the better the predictions are. Fig. 14 illustrates the prediction made at house level is 15-minute ahead with 15-minute sampling resolution. Fig. 15 illustrates the prediction made at house level is 24-hour ahead with 15-minute sampling resolution. Both results consistently show that the proposed Markov model has better performance in short term prediction, while machine learning method has better performance in the 24-hour ahead prediction.

References

[1] Energy Efficiency Strategic Plan, The government of California, Accessed at, http://www.energy.ca.gov/ab758/documents/CAEnergyEfficiencyStrategicPlan_Jan2011.pdf, 2011.

[2] B. Dong, Z. Li, G. Mcfadden, An investigation on energy-related occupancy behavior for low-income residential buildings, Sci. Technol. Built Environ. 21 (6) (2015) 892–901.

[3] H.B. Gunay, W. O'Brien, I. Beausoleil-Morrison, A critical review of observation studies, modeling, and simulation of adaptive occupant behaviors in offices, Build. Environ. 70 (2013) 31–47.

[4] A.I. Dounis, C. Caraiscos, Advanced control systems engineering for energy and comfort management in a building environment—a review, Renew. Sustain. Energy Rev. 13 (6) (2009) 1246–1261.

[5] M.A. ul Haq, M.Y. Hassan, H. Abdullah, H.A. Rahman, M.P. Abdullah, F. Hussin, D.M. Said, A review on lighting control technologies in commercial buildings, their performance and affecting factors, Renew. Sustain. Energy Rev. 33 (2014) 268–279.

[6] T. Hong, H. Sun, Y. Chen, S.C. Taylor-Lange, D. Yan, An occupant behavior modeling tool for co-simulation, Energy Build. 117 (2016) 272–281.

[7] X. Feng, D. Yan, C. Wang, On the simulation repetition and temporal discretization of stochastic occupant behaviour models in building performance simulation, J. Build. Perform. Simul. (2016) 1–13.

[8] C. Wang, D. Yan, H. Sun, Y. Jiang, A generalized probabilistic formula relating occupant behavior to environmental conditions, Build. Environ. 95 (2016) 53–62.

[9] S. D'Oca, T. Hong, Occupancy schedules learning process through a data mining framework, Energy Build. 88 (2015) 395–408.

[10] X. Feng, D. Yan, T. Hong, Simulation of occupancy in buildings, Energy Build. 87 (2015) 348–359.

[11] T. Hong, D. Yan, S. D'Oca, C.F. Chen, Ten questions concerning occupant behavior in buildings: the big picture, Build. Environ. 114 (2017) 518–530.

[12] D. Yan, W. O'Brien, T. Hong, X. Feng, H.B. Gunay, F. Tahmasebi, A. Mahdavi, Occupant behavior modeling for building performance simulation: current state and future challenges, Energy Build. 107 (2015) 264–278.

[13] A. Mahdavi, F. Tahmasebi, Predicting people's presence in buildings: an empirically based model performance analysis, Energy Build. 86 (2015) 349–355.

[14] J. Tanimoto, A. Hagishima, H. Sagara, A methodology for peak energy requirement considering actual variation of occupants' behavior schedules, Build. Environ. 43 (4) (2008) 610–619.

[15] D. Wang, C.C. Federspiel, F. Rubinstein, Modeling occupancy in single person offices, Energy Build. 37 (2) (2005) 121–126.

[16] I. Richardson, M. Thomson, D. Infield, A high-resolution domestic building occupancy model for energy demand simulations, Energy Build. 40 (8) (2008) 1560–1566.

[17] J. Page, D. Robinson, N. Morel, J.L. Scartezzini, A generalised stochastic model for the simulation of occupant presence, Energy Build. 40 (2) (2008) 83–98.

[18] C. Wang, D. Yan, Y. Jiang, A novel approach for building occupancy simulation (Vol. 4, No. 2, pp. 149–167), in: Building Simulation, Tsinghua University Press, co-published with Springer-Verlag GmbH, 2011, June.

[19] P.D. Andersen, A. Iversen, H. Madsen, C. Rode, Dynamic modeling of presence of occupants using inhomogeneous Markov chains, Energy Build. 69 (2014) 213–223.

[20] V.L. Erickson, A.E. Cerpa, November. Occupancy based demand response HVAC control strategy, in: Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-efficiency in Building, ACM, 2010, pp. 7–12.

[21] U. Wilke, Probabilistic Bottom-up Modelling of Occupancy and Activities to Predict Electricity Demand in Residential Buildings (Doctoral dissertation, École Polytechnique Fédérale de Lausanne), 2013.

[22] C. Manna, D. Fay, K.N. Brown, N. Wilson, November. Learning occupancy in single person offices with mixtures of multi-lag Markov chains, in: Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on, IEEE, 2013, pp. 151–158.

[23] S. Liu, M. Yamada, N. Collier, M. Sugiyama, Change-point detection in time-series data by relative density-ratio estimation, Neural Netw. 43 (2013) 72–83.

[24] V. Tabak, B. de Vries, Methods for the prediction of intermediate activities by office occupants, Build. Environ. 45 (6) (2010) 1366–1372.

[25] D. Aerts, J. Minnen, I. Glorieux, I. Wouters, F. Descamps, A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison, Build. Environ. 75 (2014) 67–78.

[26] B. Dong, Z. Li, S.M. Rahman, R. Vega, A hybrid model approach for forecasting future residential electricity consumption, Energy Build. 117 (2016) 341–351.

[27] Z. Li, S.M. Rahman, R. Vega, B. Dong, A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting, Energies 9 (1) (2016) 55.

[28] S.M. Rahman, Rolando Vega PhD, P.E, Machine learning approach applied in electricity load forecasting: within residential houses context, ASHRAE Trans. 121 (2015) 1V.

[29] K. Levenberg, A method for the solution of certain non-linear problems in least squares, Q. Appl. Math. 2 (2) (1944) 164–168.

[30] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intelligent Syst. Technol. (TIST) 2 (3) (2011) 27.

[31] J. Torriti, Demand side management for the european supergrid: occupancy variances of european single-person households, Energy Policy 44 (2012) 199–206.

[32] J. Widén, A. Molin, K. Ellegård, Models of domestic occupancy, activities and energy use based on time-use data: deterministic and stochastic approaches

with application to various building-related simulations, J. Build. Perform. Simul. 5 (1) (2012) 27–44.
[33] M.A. López-Rodríguez, I. Santiago, D. Trillo-Montero, J. Torriti, A. Moreno-Munoz, Analysis and modeling of active occupancy of the residential sector in Spain: an indicator of residential electricity consumption, Energy Policy 62 (2013) 742–751.
[34] U. Wilke, F. Haldi, J.L. Scartezzini, D. Robinson, A bottom-up stochastic model to predict building occupants' time-dependent activities, Build. Environ. 60 (2013) 254–264.
[35] B. Dong, K.P. Lam, A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting, Vol. 7, No. 1, pp. 89–106, in: Building Simulation, Springer Berlin Heidelberg, 2014, February.
[36] Z. Chen, Y.C. Soh, Comparing occupancy models and data mining approaches for regular occupancy prediction in commercial buildings, J. Build. Perform. Simul. (2016) 1–9.
[37] A. Meier, How People Actually Use Thermostats. ACEEE Summer Study on Energy Efficiency in Buildings, American Council for an Energy Efficient Economy, Pacific Grove, Calif, 2012.