Energy and Buildings 158 (2018) 268–281

Contents lists available at ScienceDirect

Energy and Buildings

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

Original Studies

Short term predictions of occupancy in commercial buildings—Performance analysis for stochastic models and machine learning approaches

Zhaoxuan Li, Bing Dong*

Department of Mechanical Engineering, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, United States

ARTICLE INFO

Article history:
Received 21 June 2017
Received in revised form 17 August 2017
Accepted 19 September 2017
Available online 21 September 2017

Keywords:
Occupancy prediction
Moving window
Machine learning
Field data

ABSTRACT

Real-time occupancy predictions are essential components for the smart buildings in the imminent future. The occupancy information, such as the presence states and the occupants' number, allows a robust control of the indoor environment to enhance the building energy performances. With many current studies focusing on the commercial building occupancy, most researchers modeled either the occupancy presence or the occupants' number without evaluating the model potentials on both of them. This study focuses on 1) providing a unique data set containing the occupancy for the offices located in the U.S with difference pattern varieties, 2) proposing two methods, then comparing them with four existing methods, and 3) both presence of occupancy and occupancy number are predicted and tested using the approaches proposed in this study. In detail, the paper develops a new moving-window inhomogeneous Markov model based on change point analysis. A hierarchical probability sampling model is modified based on existed models. They are additional compared to well-known models from previous researchers. The study further explores and evaluates the predictive power of the models by various temporal scenarios, including 15-min ahead, 30-min ahead, 1-h ahead, and 24-h ahead forecasts. The final results show that the proposed Markov model outperforms the other methods with a max 22% difference in terms of presence forecasts for 15-min, 30 min and 1-h ahead. The proposed Markov model also outperforms other models in occupancy number prediction for all forecast windows with 0.34 RMSE and 0.23 MAE error respectively. However, there is not much performance difference between models for 24-h ahead predictions of occupancy presence forecast.

1. Introduction

Modern cities consist of sustainable and resilient infrastructures, where building is a major constituent. Buildings' energy consumption contributes to more than 70% of the electricity usages in the cities [1], profoundly impacting the electrical grid operations. It is necessary to let the building operations not only provide comfort but also minimize energy cost simultaneously. The integrated way to control the building systems, demonstrated as the smart building in California [2], is a complex problem especially when occupants' comforts are usually the first priority. Humans spend more than 90% of their time in buildings [3]. The human and building interaction, such as lighting operations, consumes around 27% to 43% of the total amount commercial building energy [4]. Another example, the control of the air conditioning via thermostat, uses around 40% of the total energy source of the residential buildings [5,6]. Occupancy-based smart controls deployed on those system could potentially save large amount of energy consumptions [7].

Comparing with residential buildings, human and building interactions in office environment are either scheduled or limited while the occupancy information is easier to be collected [8]. Therefore, it is possible to use occupancy information in the commercial building controls, especially when building automations systems become cable to couple with the occupancy sensors [9,10]. Occupancy detection nowadays is able to record the detailed occupants' number and movements in the office spaces [10,11], and can be used to satisfy occupants' illumination comfort by using occupancy-based automations of the window, blinds and lighting systems [8,10,11]. However, the slow response of the thermal system for large commercial building requires the occupancy information to be known ahead of time rather than instantaneously detection. Typical or average occupancy schedules can be used

* Corresponding author.
E-mail addresses: Zhaoxuan.li@utsa.edu (Z. Li), bing.dong@utsa.edu (B. Dong).

https://doi.org/10.1016/j.enbuild.2017.09.052
0378-7788/© 2017 Elsevier B.V. All rights reserved.

as indicators to achieve precooling or heating for thermal conditioning system at building level [12]. However, larger saving and individual comforts can be achieved through occupancy information learning and prediction, which is exactly what Nest thermostat is doing for residential houses [13]. Many building Model Predictive Control (MPC) studies further demonstrated the need to have flexible occupancy models of different temporal resolutions (e.g. intra-hour and hourly) and prediction horizons (e.g. hour-ahead to day-ahead) [7,14–16]. In conclusion, reaching this vision for smart building controls requires accurate occupancy forecasts within one day ahead scale to capture the randomness of occupied periods and changes of occupancy patterns. The occupancy forecast model for smart building controls should be capable to predict for different temporal resolutions (e.g. 15-min to 24-h 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 usually studied separately by previous researchers.

In this paper, the authors provide a review of current studies and methods on modeling and predicting the occupancy. One new Markov model and a modified probabilistic model to predict both the presence status and the occupancy numbers are proposed. Insights on the current popular models and their approaches are presented in Section 2. Section 3 introduces the new developed Markov chain model, the modified probability sampling model, and a brief description of the models used to be compared with. Section 4 provides the results of each model on occupancy prediction. Temporal and spatial differences among models' predictive performance are further assessed. Section 5 discusses importance and concludes this study.

2. Current state-of-the art

Two types of occupancy information can be used for the smart buildings especially by the predictive controls of the air conditioning system. The first one, the binary states (the presence and absence) of individual occupant at space level, addresses individual thermal comfort and maximum energy saving potentials. The second one, the number of occupants at space or building level addresses the ventilation and occupants' group comforts. There are three major approaches to model those two types of the occupancy: 1) the sampling method based on the random sampling, 2) the Markov method based on the Markov chain, and 3) the statistical learning methods including data mining, agent-based model, and machine learning.

The random sampling, in essence, generates the distributions of the key information (e.g. first arrival times) through training data, and then uses different sampling methods to reproduce the occupancy [17–24]. One popular building occupancy simulator is the light-switch algorithm developed by Reinhart [17,18]. It modeled the arrival and departure time per day using cumulative distribution functions (CDFs). Three intermediate phases, morning, lunch, and afternoon are fitted to Probability Distribution Functions (PDFs) with starting times and duration lengths of absences. To simulate the occupancy, the first arrival time and last departure time are firstly identified using random sampling through the fitted CDFs and PDFs. Then, intermediate absence start times can be determined by comparing the pseudorandom numbers against the intermediate absence PDFs. The durations of the intermediate events can be obtained by the random sampling again from the duration PDFs. Later studies developing similar models concentrated on the improvement of the model's simulation accuracy and generality by introducing different kinds of distribution types, building samples, and sampling approaches. Wang et al. simulated occupancy in a single-person office with the occupancy divided to occupancy and vacancy events. The intervals of the events are fitted by maximum-likelihood algorithm using exponential distributions [19]. Sun et al. modelled the overtime occupancy and Kolmogorov-Smirnov tests were used to calibrate the exponential distributions of overtime durations and the binomial distributions of the number of overtime occupants respectively [20]. Tabak and Vries utilized a different statistical curve, the S-curve, to fit more detailed intermediate events in the office environment including vacancy times due to restrooms, visitors, printing, smoking, and sports [21]. Silva and Ghisi implemented the Latin hypercube sampling on the occupants schedules to design the occupancy model for the EnergyPlus simulation [22]. Chang and Hong [23] demonstrated the generality of the sampling models by fitting the daily absences, absence duration, and the starting time of each absences of 200 single-occupant office data collected on-site from a three floor commercial building. Besides occupancy simulation, researchers recently started to use the sampling method to predict the occupancy presence for building predictive controls. Mahdavi and his group [24] used the original Reinhart's model compared to two other types of models, a Markov model developed by Page [25] and a non-probabilistic schedule. They predicted one day ahead occupancy using the university offices collected within a year. The results did not show significant differences between the models and a non-probabilistic schedule, demonstrating the need to revise the occupancy models original developed for building simulation. Nonetheless, the random sampling is one possible model to be integrated with advanced building control, such as the Model Predictive Control (MPC) of air conditioning system.

The first order Markov chain is another popular stochastic occupancy modeling approach that has been applied in building simulations [25–33]. One classic model is the generalized inhomogeneous Markov chain model developed by Page [25]. This model is used to model the binary office occupancy with occasional periods of long absence. The transitional probabilities are inverse sampled from the trained transitional PDFs (e.g. presence to presence). The other types of the transitions (e.g. presence to absence) are further calibrated using fitted profile of mobility parameters. The mobility parameters are empirically defined by the authors to describe the relationships among the different transitions. On the other hand, Wang [26] proposed an approach based on homogenous Markov chain to simulate the occupants' number of office building. The model generates the location for each occupant through a homogenous Markov matrix and further aggregated as the zonal level occupants' number while moving around events such as going to the office, getting off work and lunch break have separate transitional matrixes. However, most occupancy simulations used inhomogeneous Markov chains for residential building due to the more stochastic behaviors and occupancy events of the residential occupants [27–29]. Furthermore, Anderson [30] designed a sophisticated dynamic estimation approach "apple to apple" comparing the simulation accuracy of the inhomogeneous and homogenous Markov chains. The training of the Markov chains with binary occupancy states are transformation to a generalized linear regression problem using a logistic form of the transitional probabilities. The logarithm likelihood of the joint form of all transitional probabilities at each time is regressed with only time information of two polynomial orders. Statistical modelling using spline function and exponential smoothing are utilized to further improve the model accuracy. Besides first-order Markov chains discussed above, a higher-order Markov model is used to connect several occupancy events (e.g. the durations) together [31,32]. However, it is overqualified to an online predictive control problem of building systems, which only interests in binary occupancy states and limited occupants' number.

Unlike random sampling, Markov occupancy models have been widely used in building control studies, especially the predictive controls of air conditioning systems. However, the first type of predictive Markov model appeared in control studies is the hidden Markov chain. Hidden Markov chain introduces additional states (the hidden states) to achieve real-time prediction of the occupancy. The hidden states are the observations of the buildings' environment measurements. They are assumed to have probabilistic correlations to the occupancy states [33]. Various control studies have demonstrate the significant savings achieved by using this occupancy prediction tool [33,34]. Meanwhile cost and time-delay problems, to measure the hidden states, prohibit the further application in real time control. Namely, the CO_2 sensors are expensive to install and have time lagging issues. Recent research focused on utilizing the first order Markov chain to achieve fast and cost effective online occupancy predictions. As mentioned earlier, Mahdavi and his groups used both Reinhart's sampling model and Page's Markov model to predict one day ahead occupancy for a single office [24]. Similar studies have been done by Erikson and his group but several modifications were made on the Markov model to predict the occupants' number [35–39]. With too many ocs ccupants' numbers ahain states, the prediction step may have no transitional probability for a specific occupant's number (a sink state) when using finite training window in building MPC. A closest distance method is used to discover the closest sink state in training and guarantees a representative transitional probability for an incoming occupants' number [36]. A blended strategy can also be used by training the data in lower time resolution and linear combine all possible transitional matrixes to avoid sink states [37]. A more thorough but computational expensive way is to have a minute level transitional matrix in a moving window that contains all observable chain states (the occupants' number) for hourly ahead predictions [38]. Dobbs and Hencey implemented a different approach, the Bayesian Inference, and demonstrated a self-adaptive training algorithm for binary occupancy states [39]. By treating the binary occupancy as a binomial distribution, transitional probability only needs to be calculated as the expected value of a certain transitional distribution at that specific time. Using fractional occupancy (binary data averaged out by a fixed dwell time), the transitional distributions as posterior distributions are estimated from a uniform distribution using Bayesian inference. Clearly noticed from those studies that the occupancy models in building MPC require different time resolution for predictions spanning from intra-hour to intra-day scale. All models are also specially designed to predict one aspect of the commercial building occupancy, either binary occupancy states (the presence and the absence) or multiple occupancy levels (the occupants' number in certain space). Therefore, it would be interesting to explore the possibility to use one general predictive occupancy model covering both occupancy binary states and multiple occupancy levels at different temporal resolutions.

The statistical learning methods, including data mining, agent-based model, and machine learning, emerged recently in building occupancy simulations. The first method, data-mining, models the occupancy events through clustering analysis based on a large amount of data, such as hierarchical agglomerative clustering [40], nearest neighbor clustering [41], and decision tree classification [42]. The method required high quality occupancy data which is not applicable to all building cases. The second learning model, the agent-based model, is to use the agent learning ability to simulate occupancy events based on predefined event rules and decision-making mechanisms [43]. One typical study done by Liao and Barooah [44] developed a Monte-Carlo algorithm based on mixed rules of the agents (the occupants) to simulate the occupancy events. For individual agents, the probability of that agent occupying certain space node at the simulation time step is integrated with an acceleration rule and a damping rule. The final profile is generated by associating with other agents in a reduced-order graphical model. Cook and his group [45] implemented a similar agent-based model in a smart home environment and different occupancy events were simulated from real collected smart home data. However, the difficult to scale up the model for general building MPC studies is obvious due to the model formulation largely relies on the specific design rules and tedious calibrations. The last method, the machine learning approach, has actually been used in building occupancy predictions. Yu applied a generic programming by using time information (day, hour, minute), and occupancy durations (the length of time the occupant spent in the state prior to the previous state, the length of time the occupant has been in the office since the first arrival of the day) to predict the occupancy for five different offices [46]. Dong et al. applied more robust machine learning tools, such as Artificial Neural Network and Support Vector Regression, to predict the occupants' number in building MPC studies [34,47,48]. Chen and Soh [49] compared the machine learning methods to Markov model, Multivariate Gaussian Sampling, and Autoregressive model to predict the occupants' number in a research lab. Superior of the Support Vector Regression is demonstrated with lower root mean square error while Artificial Neural Network has lower mean error.

In conclusion, occupancy models in building simulations could be used for real occupancy predictions with certain changes. Namely, Page's Markov model and Reinhart's sampling approach could be trained in fixed moving window strategy to predict binary occupancy states [24]. However, major modifications should be carried out. Namely, Page's Markov model is designed to simulate long term data. Hence a week profile of mobility parameters including weekends is utilized by a long term period training (e.g. 5 years data in the study). Firstly, it is very difficult for every building to have such high quality long term data. Secondly, with fixed training window in building predictive control, the data normally would not be expected to be fully used more than one month. How this effects the model parameter calibrations is unknown. Similar issues existed for Reinhart' model. Additionally, previous research did not investigate much on the models' potentials to forecast both the occupancy presence and the occupants' number. A general occupancy predictor could be very helpful to fulfill multiple control tasks in a smart building. It is also necessary to evaluate the model's generality at different temporal resolution and prediction horizons, as mentioned earlier in the previous discussions on the current building MPC studies.

3. Methodology

Given the current state of the art, the authors have developed two systematic approach of short-term predictions of the occupancy profile for commercial buildings. One is a new integrated Markov model including a change-point analysis with a moving window training. The other is a modified random sampling approach that able to predict both occupancy presence and number. Here, "short-term" means within 24-h ahead prediction. They are compared with two simulation models (Page's Markov model and Reinhart's Light-switch sampling model [17,25]) and two machine learning methods (Artificial Neural Network and Support Vector Regression) for predictions of occupancy presence and number respectively. Occupancy presence data from four different offices are collected every 5-min time step. The data log for motion sensors entails a sequence of times stamped from the presence (value of 1) to the absence (values of 0). Occupancy number data from one laboratory is recorded by a monitoring system with cameras. The image process technique processes hundreds of videos to ensure the correct documents of the occupants' number. A low pass filter is utilized to both data to generate the time series occupancy in 15-min, 30-min and 1-h interval for different prediction window. The occupancy model, developed upon these data, is designed to deliver the necessary inputs for the model predictive controls of the built

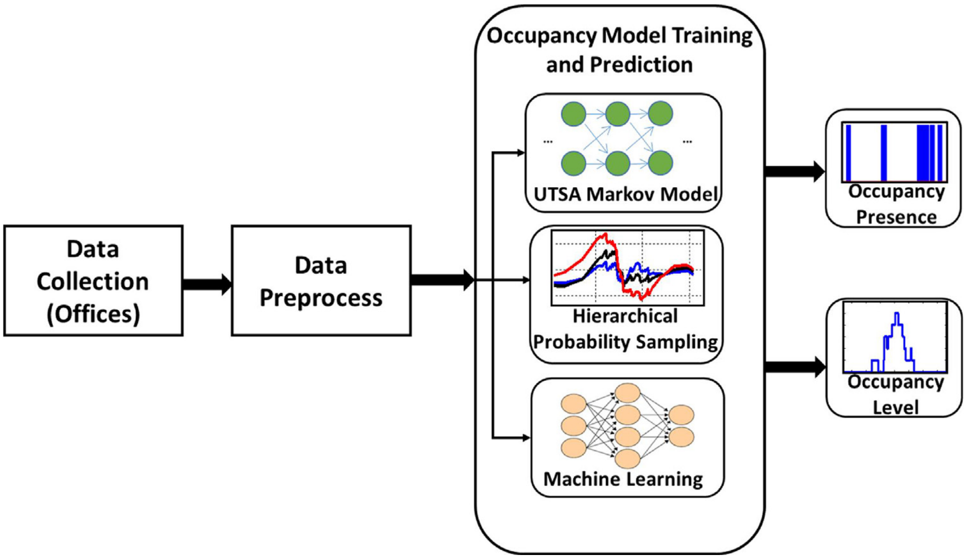

Data Collection (Offices)
Data Preprocess
Occupancy Model Training and Prediction
UTSA Markov Model
Hierarchical Probability Sampling
Machine Learning
Occupancy Presence
Occupancy Level

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

environment. In the following sections, the developed new Markov process model to predict occupancy is presented and discussed in Section 3.1. Section 3.2 demonstrates an integrated hierarchical probability model based on the random sampling approach that capable to forecasts both occupancy presence and occupancy number. Section 3.3 presents the brief model settings of the two simulation models and two machine learning approaches (Fig. 1).

3.1. An innovative inhomogeneous markov chain model

Let a Markov chain X at time step k be a time sequence x_1, x_2, ..., x_k and the occupancy states be $S = \{s_1, s_2, ..., s_n\}$ where $n \le k$. The chance of x_k(containing occupancy state s_i) transit to x_{k+1} (containing occupancy 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 $p(x_{k+1} = s_j | x_1, x_2, \ldots, x_k) = p(x_{k+1} = s_j | x_k)$.

Given n_{ij} pairs of the transitional states observed as $\{s_i, s_j\}$ of all pairs of the transitional states $\{s_i, s_l\}$ that belong to the training data, the transition probability is estimated as:

$$\hat{p}_{ij} = \frac{n_{ij} + \alpha}{\sum_{l=1}^{k} (n_{il} + \alpha)} \tag{2}$$

Where α is a smooth factor. A smooth factor could increase the likelihood of the states' transitions and reduce the "sink" states, discussed in the Section 2. It can also enhance the dramatic occupancy changes at the morning ramp-up and decrease at the evening ramp-down. Noticed that offices will have very low or zero transition probabilities during long vacancy and occupancy periods with constant occupancy number. A smooth factor could force the unlikely occupancy to happen. To address the issue, authors define the smooth factor as time serious step function with small values of 0, 0.05 and 0.1 only decided by empirical. Namely, a factor of 0 for conservative estimated long working periods in the morning and afternoon. A higher value of 0.1 for ramp-up or ramp-down periods. Similar rules are applied to the transitional probabilities of the occupancy number.

The authors use a new approach to develop the inhomogeneous Markov chain by integrating a change-point analysis with an optimal moving window training strategy [50]. Modifications are made for office occupancy data containing both binary occupancy and occupancy number. Assuming a building MPC is rolling at 15-min resolution, there will be total 96 set of transition probabilities need to be updated for a day ahead MPC optimization. For each set of the time inhomogeneous transitional probabilities, it is estimated within an optimal window before each of the occupancy predicted time step. The period of the optimal window is decided by the changing point of the occupancy rate based on a daily profile of the historical occupancy, as shown in the hypothetical MPC of Fig. 2. Training data contains historical occupancy in this training window ("training window" in Fig. 2) from recent working days. For the hypothetical example to predict the chain state x_{k+1}, a change point is detect after the chain state x_m. Hence, the optimal window length spans from $m+1$ to k time steps. The training data set is the historical occupancy states observed in this time window of the recent working days as $\{(x_{m+1}, x_{m+2}, ..., x_k);..;(x_{m+1-96\times d}, x_{m+2-96\times d}, ..., x_{k-96\times d})\}$ where d is the historical day index. Here, authors selected limited training size of 10 working days for one-step ahead forecast and 20 working days for day-ahead forecast.

As shown in Fig. 2, the training data set of the model relies on the occupancy change-point analysis. Since the authors design the model for a rolling MPC, the daily profile is constantly updated. Thus a visual identifications of change points are not adaptive. Let $D = \{d_1, ..., d_{24\times z}\}$ represents all historical occupancy before prediction day. Here, historical data contains all the working days z that are updated for MPC. A discrete profile of the occupancy rate in daily scale is generated by:

$$P(i) = \frac{\sum_{j=1}^{z} (\lambda^{z-j} \cdot d_{(j-1)\times 24+i})}{z \times \max(D)} \tag{3}$$

Where d is the chain state which could be binary occupancy contains only 0 and 1 or multiple occupancy states contains occupants' number, i is the time step of the daily profile where $1 \le i \le 96$ if the occupancy data is in 15-min scale, max(D) gives the maximum number observed of the total occupancy data, and λ is an exponential forgetting factor, which is below 1. The forgetting effect reduce the influence from the too old occupancy information.

The change-point detection algorithm uses relative density-ration estimation with the Pearson divergence scoring the possible

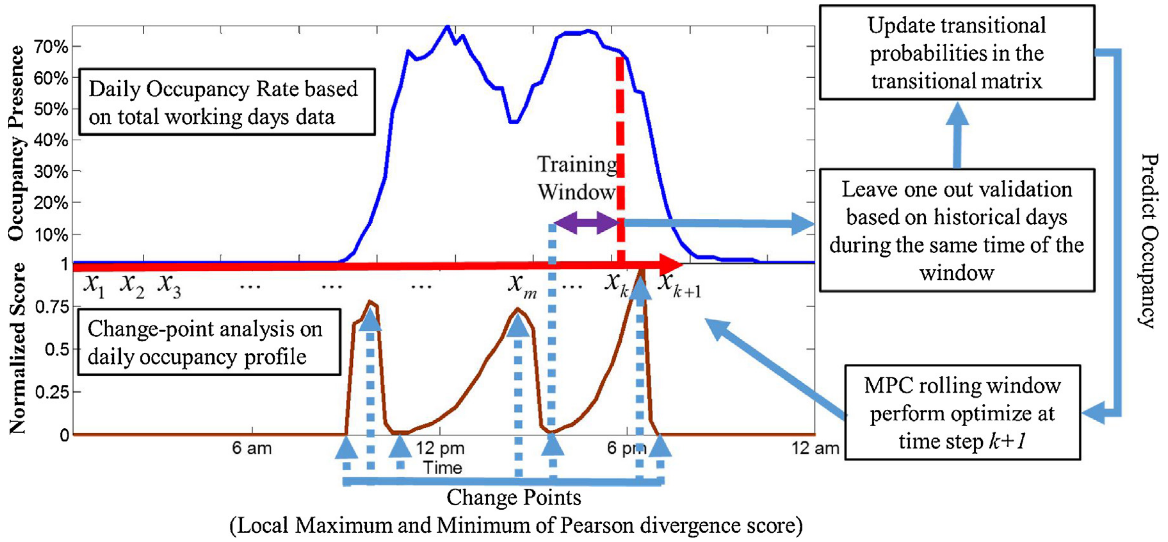

Occupancy Presence
70%
60%
50%
40%
30%
20%
10%
Daily Occupancy Rate based on total working days data
Training Window
Normalized Score
1
0.75
0.5
0.25
0
x_1 x_2 x_3 … … … x_m … x_k x_{k+1}
Change-point analysis on daily occupancy profile
6 am
12 pm
Time
6 pm
12 am
Change Points
(Local Maximum and Minimum of Pearson divergence score)
Update transitional probabilities in the transitional matrix
Leave one out validation based on historical days during the same time of the window
MPC rolling window perform optimize at time step k+1
Predict Occupancy

Fig. 2. Markov model training and predicting in a rolling MPC.

change points of the daily profile $P(i)$. For a data set m sampled from the daily profile D, the divergence score is defined as follows [51]:

$$\int P_\beta(m)\left[\frac{P(D)}{P_\beta(m)}-1\right]^2 d(m)+\int P(D)\left[\frac{P_\beta(m)}{P(D)}-1\right]^2 d(D) \qquad (4)$$

Where $P_\beta(m)=\beta P(D)+(1-\beta)P(m)$, P is the probability density function of the corresponding data set, and the factor β is a weight factor which is 0.5 to put equal weight of the distributions. For the example of Fig. 2, the day file as the sample D contains 96 occupancy rates in 15-mintue resolution. The data set m is sampled using a sliding window size of 12 (3 h data). The sliding continues forwardly until end of the day and then resample backwardly again. A symmetric score is calculated by the summation of forward sliding and backward sliding scores using Eq. (4). MATLAB toolbox developed by Liu is used in this study [51].

As shown in Fig. 2 again, the predictive power of the model depends on a modified leave-one-out validation for multiple training rounds. For each round of training validation, $d-1$ (e.g. 9 for one-step ahead forecast) days of historical data in the optimal window randomly selected to calculate the transitional probabilities based on Eq. (2). Validations will be performed in the same optimal window of all d historical days. The process compares a threshold value with transitional probabilities to generate occupancy. For example, the binary transition probability from presence states to absence states give a vacancy if below the threshold of 0.5 (mean value of a uniform distribution). Similar strategy used for occupancy number prediction. Number of the rounds of validations is the same as the number of training days and total d (e.g. 10 for one-step ahead forecast) contingency tables can be produced. The correctness C of each contingency table, calculated by comparing the generated occupancy and the training occupancy data, is:

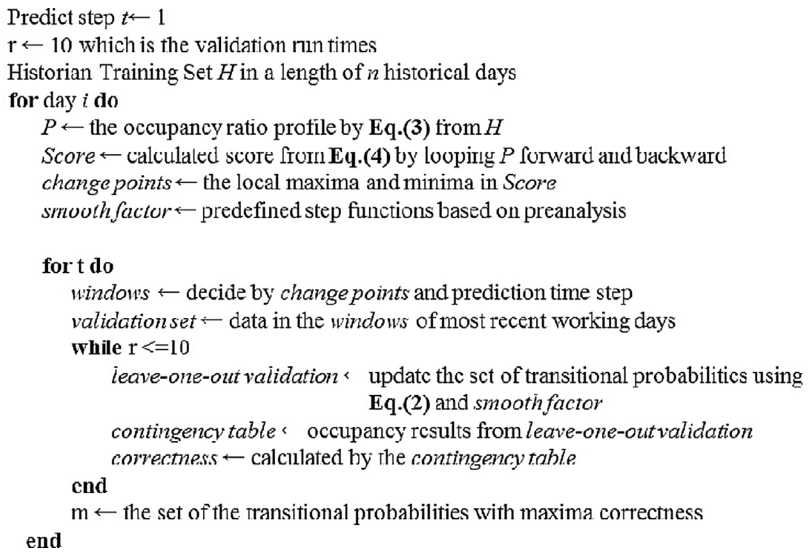
$$C=\frac{\sum TP}{N} \qquad (5)$$

Where TP represents the number of correct validated occupancy and N is the total number of the occupancy at the validation process. The set of the transitional probabilities produce the maximum correctness will be the final transitional probabilities for that trained time step (the time step k for the hypothetical example). By moving the optimal window and change the window length according to the change points, each time step will have a cross-validation validated time inhomogeneous probabilities for the Markov chain. The pseudo algorithm of the complete training process in Fig. 2 is illustrated as following (Fig. 3).

3.2. Hierarchical probability sampling

The authors develop a hierarchical approach based on a classic model by Reinhart's light-switch and modify it to predict both the presence states and occupancy levels. The hierarchical model has two levels. The first level predicts presence and the second level predicts the occupancy number. For the first level, the light-switch model utilized the probability density functions (PDFs) on the intermediate breaks of the morning, the lunch and the afternoon periods, along with the fitted PDFs of the absence lengths for each type of the intermediate absences. Details of the methods have discussed in the review part of Section 2. However, the light-switch used empirical rules to visually determine the different periods (e.g. morning) which is not adaptive enough in a rolling window MPC. Hence, the change-point analysis illustrated in previous Section 3.1 is used again to auto determining the morning, noon, and afternoon phases based on the local change-point scores, as illustrated in Fig. 4. Here, the "local" is statistically defined extremes in a given range that is opposite to the "global" maxima or minima. From Fig. 4, there are apparently 3 peaks, which are three "local" maxima and one highest peak, which is the "global' maxima. Beginning and ending of the morning and afternoon periods only decided by the first and last local maximum score. The noon period is determined by the middle local maximum score. The middle score is found by ranking the local maximum values by appearing time. To determine the noon period, the distance between the closest local minima and the middle maxima is found and the noon period is the double values of this distance as the middle maxima is the center time point. Noticed if no middle maxima is found, there is no obvious noon period (pattern in Fig. 6(b) and (d)) A single long period for intermediate events will be used instead.

The first arrival, last departure, intermediate absence starting time, and the intermediate absence durations are fitted by twelve common probability distributions. They are t location-scale, Loglogistic, Logistic, Inverse Gaussian, Brinbaum-Saunders, Lognormal, Rician, Normal, Weibull, Extreme Value, Rayleigh, and Exponential. Noticed that a direct fitting to get the training data's empirical distribution is possible. However, it is a bootstrap sample limited by the training period. Hence, authors assume the training data belongs to a 'true' distribution that a statistical fitting need to be



```
Predict step $t$ ← 1
r ← 10 which is the validation run times
Historian Training Set $H$ in a length of $n$ historical days
for day $i$ do
    $P$ ← the occupancy ratio profile by Eq.(3) from $H$
    $Score$ ← calculated score from Eq.(4) by looping $P$ forward and backward
    $change\ points$ ← the local maxima and minima in $Score$
    $smooth\ factor$ ← predefined step functions based on preanalysis

    for t do
        $windows$ ← decide by $change\ points$ and prediction time step
        $validation\ set$ ← data in the $windows$ of most recent working days
        while r <=10
            $leave$-$one$-$out\ validation$ ‹  update the set of transitional probabilities using
                                               Eq.(2) and $smooth\ factor$
            $contingency\ table$ ‹  occupancy results from $leave$-$one$-$out\ validation$
            $correctness$ ← calculated by the $contingency\ table$
        end
        m ← the set of the transitional probabilities with maxima correctness
    end
```

Fig. 3. Algorithm of the Markov model training for predictive purpose.

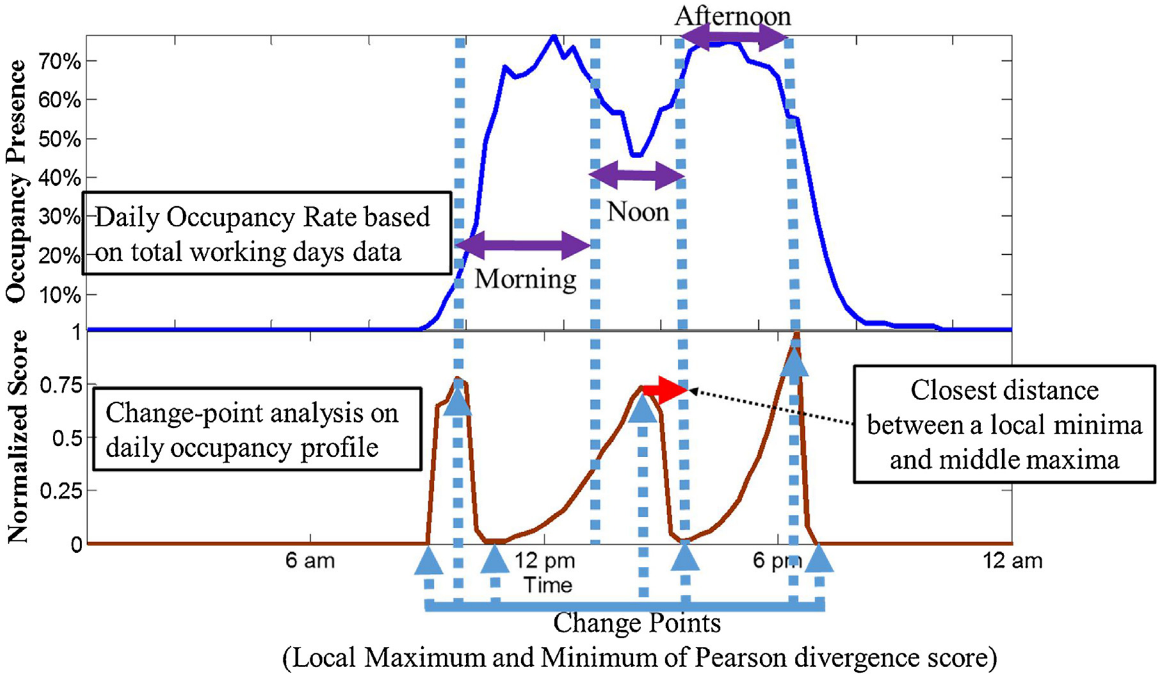


Fig. 4. Detection of morning, noon and afternoon periods.

performed. All the distributions are fitted using maximum likelihood and ranked according to the Bayesian Information Criteria (BIC) defined as follows:

$$BIC = -2\mathrm{Ln}\hat{L} + k\mathrm{Ln}(n) \tag{6}$$

Where $\hat{L}$ is the maximized value of the likelihood function of the fitted distribution, k is the number of free parameters to be estimated based on fitting distribution type and n is the number of fitting data. The ones with the highest ranked Bayesian Information Criteria are used to represent the cumulative distribution functions (CDFs) of the first arrival time and the last departure time, the PDFs of the intermediate absence beginning times, and the CDFs of the intermediate vacancy durations for different periods (e.g. morning, noon, etc).

For the second level of the model, the number of occupants is fitted by an empirical distribution of the occupancy level at each time step. The authors avoid to use continuous distributions like Gaussian distribution in previous studies [35,49] owing to the occupancy number is discrete integer. Let the observed set of occupancy levels is $L = \{l_1, l_2, ..., l_n\}$ at time step k for all training n days. Then the empirical CDF is defined as:

$$F_n(l) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le l) \tag{7}$$

Where I is the indicator function calculating the observed frequency when occupancy number l happens. The function gives value 1 if the occupancy level x_i at the time step i is smaller than l. Otherwise, it gives value 0. The pseudo algorithm of the training is illustrated as following:

3.3. Model Comparison

This section will briefly discuss the previous researchers' model used for comparison in this study. For binary occupancy, authors select two well know models, Page's and Reinhart's occupancy simulation models, and two machine learning approaches. The machine learning approaches are further used to compare for the predictions of the occupancy number. To have "apple-to-apple" comparison, how to do predictions are also discussed for all the models including the two models proposed by authors.

3.3.1. Machine learning models

Two machine learning techniques are tested here for occupancy predictions: Artificial Neural Network (ANN) and Support Vector Regression (SVR). For ANN, feed forward neural network (FFNN) is used with three layer structure. FFNN is modeled with 20 training neurons at the training layer and multiple input neurons at the input layer (time lagged values of historical occupancy states). The input layer uses the hyperbolic tangent sigmoid functions for the neurons and the linear transfer function for the training layer. Weights of training neurons are learned from Levenberg-Marquardt back-propagation algorithm. 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. All those model settings are adjusted in Neural Network Toolbox of MATLAB [52]. For support vector regression (SVR), LibSVM is used in this study through MATLAB interface [53]. A radial basis function of 3 ° with weighting factor 0.1 is used. The authors tune the model parameters using 10 fold cross-validation based on mean square error. The grid search cross validation is used with parameters searched in a range of 10^3 to 10^{-3}. More detailed methodologies, model formulations, and model settings can refer to authors previous studies [54,55].

The machine learning methods train a black-box model with a training label set and the validation label set. The validation set is $\{x_{k-n},x_{k-n+1},..., x_k\}$ which is n time steps back historical occupancy. To train for one time step ahead forecast (the 15-min to 1 h ahead), the training label uses a Markov order 4 label, as follows:

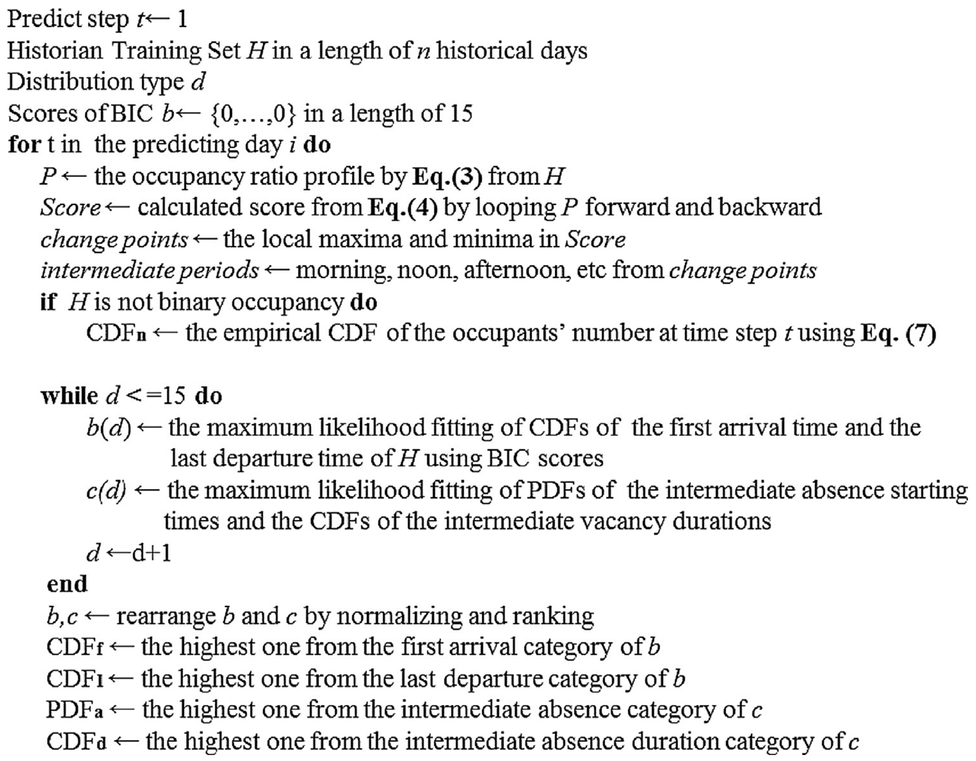
$$H_1 : \left\{ \begin{matrix} x_{k-n-1} & x_{k-n} & \cdots & x_{k-1} \\ x_{k-n-2} & x_{k-n-1} & \cdots & x_{k-2} \\ x_{k-n-3} & x_{k-n-2} & \cdots & x_{k-3} \\ x_{k-n-4} & x_{k-n-3} & \cdots & x_{k-4} \end{matrix} \right\} \quad (8)$$

For day ahead forecast, the training label uses older historical information:

$$H_2 : \left\{ \begin{matrix} x_{k-n-d} & x_{k-n+1-d} & \cdots & x_{k-d} \\ x_{k-n-2d} & x_{k-n+1-2d} & \cdots & x_{k-2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k-n-5d} & x_{k-n+1-5d} & \cdots & x_{k-5d} \end{matrix} \right\} \quad (9)$$

Where d is the time length of one day (e.g. 96 for 15 min resolution). By comparison to one time step ahead training, day ahead training needs the historical occupancy at the same time from yesterday, the day before yesterday, ., and the day before one week. These matrix orders are selected by minimizing the coefficient of determination of the validation label. The authors produce two predictive labels as the inputs to predict the occupancy state x_{k+1} based on the trained models. For one step ahead and one day ahead prediction, the predict labels are $\{x_k, x_{k-1}, x_{k-2}, x_{k-3}\}^T$ and $\{x_{k+1-d}, x_{k+1-2d},$ ..., $x_{k+1-5d}\}^T$. To predict using machine learning is straightforward:

1) define the prediction time step and prediction horizon (one step ahead or one day ahead);
2) determined training and validation labels using Eqs. (8) and (9);
3) set model parameters discussed in beginning of Section 3.3.1 and further tune the model structure using the training and validation labels;



```
Predict step $t$ ← 1
Historian Training Set $H$ in a length of $n$ historical days
Distribution type $d$
Scores of BIC $b$ ← {0,…,0} in a length of 15
for t in the predicting day $i$ do
    $P$ ← the occupancy ratio profile by Eq.(3) from $H$
    $Score$ ← calculated score from Eq.(4) by looping $P$ forward and backward
    $change\ points$ ← the local maxima and minima in $Score$
    $intermediate\ periods$ ← morning, noon, afternoon, etc from $change\ points$
    if $H$ is not binary occupancy do
        CDFn ← the empirical CDF of the occupants' number at time step $t$ using Eq. (7)

    while $d$ <=15 do
        $b(d)$ ← the maximum likelihood fitting of CDFs of the first arrival time and the
                 last departure time of $H$ using BIC scores
        $c(d)$ ← the maximum likelihood fitting of PDFs of the intermediate absence starting
                 times and the CDFs of the intermediate vacancy durations
        $d$ ←d+1
    end
    $b,c$ ← rearrange $b$ and $c$ by normalizing and ranking
    CDFf ← the highest one from the first arrival category of $b$
    CDFl ← the highest one from the last departure category of $b$
    PDFa ← the highest one from the intermediate absence category of $c$
    CDFd ← the highest one from the intermediate absence duration category of $c$
```

Fig. 5. Algorithm of the training of the hierarchical probability sampling.

4) predict the occupancy using prediction labels on the trained machine learning models.

3.3.2. Occupancy Models from Page and Reinhart

As mentioned in Section 2, Reinhart's model have distributions for four key occupancy information: the first arrival, the last departure, and intermediate departure, and the intermediate absence duration for the intermediate phase (morning, lunch and afternoon) [17,18]. The approach is corresponding the presence predictions of the hierarchical probability sampling approach developed in this study. To be fair, both approaches predict occupancy using trained distributions (CDFs and PDFs refer to Fig. 5) as follows:

1) Draw two random number r_1 and r_2 from a uniform distribution;
2) Inverse sample the first arrival time t_f based on $CDF_f(t_f) = r_1$;
3) Inverse sample the last departure time t_l based on $CDF_l(t_l) = r_2$;
4) Loop between t_f and t_l to draw two random number r_3 and r_4 from a uniform distribution for each prediction step;
5) if $r_3 \leq PDF_a(t)$, an absence event happened, and inverse sampling the absence period based on the duration type (e.g. morning) where $CDF_d(T) = r_4$;
6) Jump T time steps ahead in looping if an absence event happened or else continue step 4) until the last departure time is reaches;
7) Predict the occupancy number N during looping only if occupied at the time step and inverse sample the number by $CDF_n(N) = r_5$ where r_5 is also draw from a uniform distribution.

The other occupancy model developed by Page derives the transitional probabilities based on aggregated presence probabilities of one weeks in the original work [25]. It is not possible to generate accurate weekly presence profile unless long period data (e.g. one year) is used. Hence, authors here only use aggregated one day profile of the presence probabilities. A mobility parameter defined by user as μ is used to calculate both the transitional probability $T_{01}(t)$ and $T_{11}(t)$ at individual time step t (absence state 0 to presence state 1 and presence state 1 to presence state 1 respectively):

$$\begin{aligned} T_{01}(t) &= \frac{\mu-1}{\mu+1}P(t) + P(t+1) \\ T_{11}(t) &= \frac{P(t)-1}{P(t)}\left(\frac{\mu-1}{\mu+1}P(t) + P(t+1)\right) + \frac{P(t+1)}{P(t)} \end{aligned} \tag{10}$$

Where $P()$ is the aggregated one day profile of the presence probabilities and more details on how to derive Eq. (10) can refer to Page's original work [25]. The thresholds of parameter of μ are estimated with "low", "medium" and "high" mobiles based on Page's definitions [25] and constant numbers 0.3, 0.6 and 0.9 are used for different periods of a day based on empirical rules. With the predestinated mobility parameter and daily presence probabilities, each time step will have a time inhomogeneous set of transitional probabilities generated.

To predict the occupancy, both Page's model and this study's Markov model use the same procedures. For one step ahead prediction, predictions are based on only observed states. Namely, if current observed state is presence, only T_{11} of this time step is used. If a random number draw from a uniform distribution is smaller than T_{11}, a prediction of presence state is predicted. For day ahead prediction, the prediction continues based on predicted occupancy. For example, the first step will predict based on initial observation. However, the following steps can only predict base on predicted occupancy state and treat them as the "true observations" until predictions are made for the whole day. Similar strategy is adopted for occupancy number predictions.

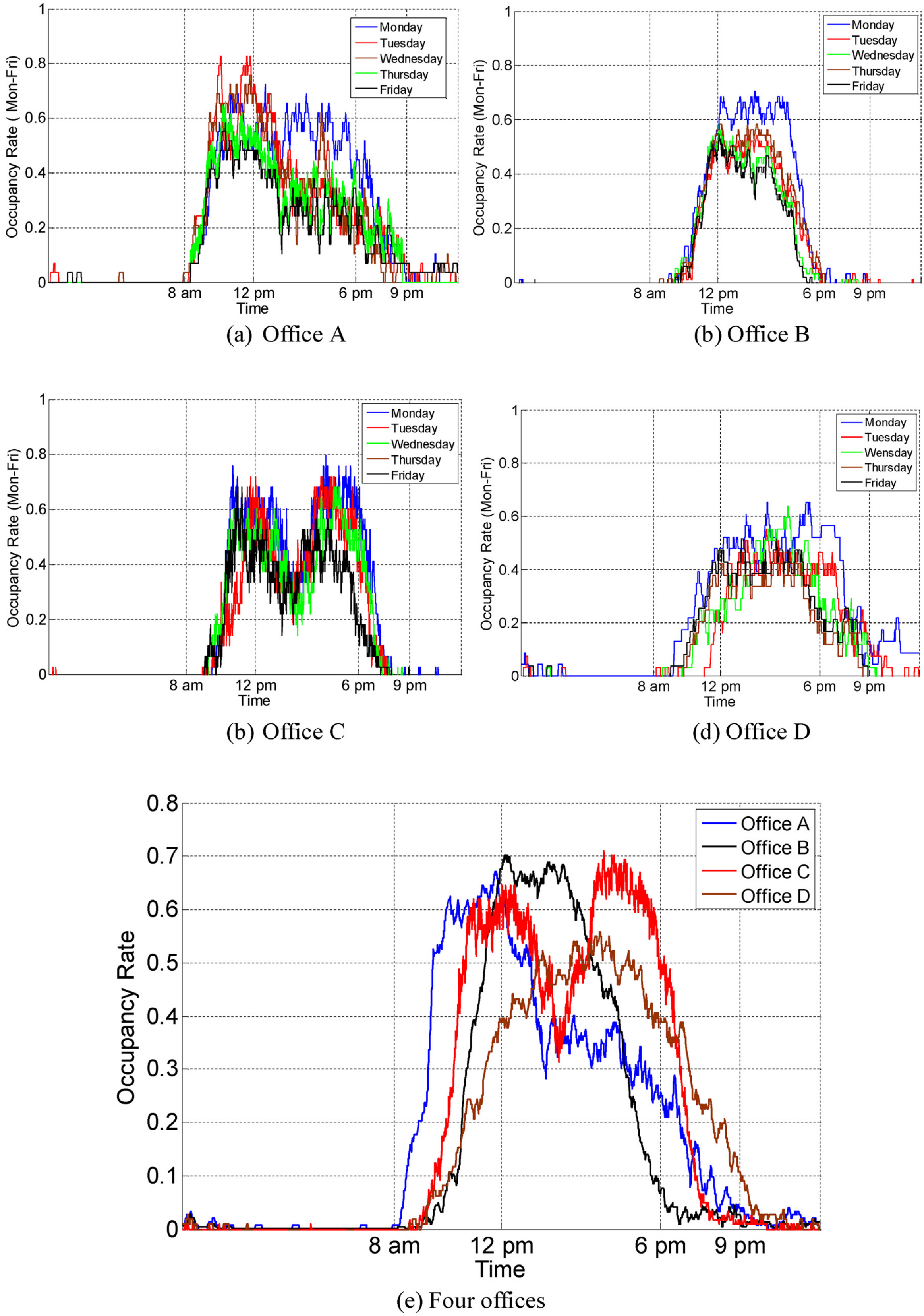
4. Results

4.1. Prediction of occupant presence

In this section, four university offices located on the second floor of a university building are used to do prediction test. The test is performed from Oct 1st 2015 to April 1st 2016. The presence probabilities of the samples during this period are presented in Fig. 6. These offices belong to a research institute that provides services for the visiting scholars and graduate students. Frequent group meetings outside of the individual offices are expected, which explain the varied 50% to 80% max presence rate. Office A is occupied by a junior assistantship. From Fig. 6(a), consistent patterns can be observed except Monday, where a higher presence rate is discovered during the afternoon. For working days besides Monday, there is clearly high working load in the morning and other tasks need absences from the office in the afternoon. Office B is occupied by a full-time advisor, which illustrates a single-mountain pattern in Fig. 6(b). Again, the occupancy at Monday is expected to be the highest. Office C is occupied by a senior administrator who demonstrates a twin-summit pattern: two highest presence peaks during the morning and the afternoon. A typical lunch break is observed as the deep valley in Fig. 6(c). Office D is occupied by two staffs. One is a student assistantship and the other is a financial secretary. This explains the long presence periods, sometimes lasting until midnight in Fig. 6(d).

The authors used a fixed sliding window to predict the occupancy by looping forward the whole test period. A two week window size (10 working days) is used for one-step ahead forecast and one month is used to predict day-ahead forecast. Different training algorithms of the models in Section 3 are applied to the historical data of the sliding window. Predictive performance is evaluated by the exact one-to-one match between the ground-truth and the occupancy predictions. Among the total l predictions, if there are m predicted presence while the observations of the rooms are occupied and n predicted absence while the observations of the rooms are not occupied, the overall accuracy is thus calculated as a percentage, $100 \times (m+n)/l$. Additionally, the evaluation periods are limited for the working time. Office A's hours are from 8 am to 7 pm, Office B from 9 am to 6 pm, Office C from 9 am to 7 pm, and lastly Office D from 10 am to 9pm. All these periods are matching with the working periods shown in Fig. 6(e).

The prediction results are presented in Table 1. In general, the probability sampling approaches, either Hierarchical Probability sampling or Reinhart's model, are less accurate during the prediction especially for one step ahead forecast (15-min, 30-min and 1-h ahead forecast). Difference up to 22% accuracy can be observed between the new Markov model and the sampling approaches. The authors further compare with Page's Markov model and machine learning approaches while the differences among them are not significant. It is evident that Markov models and machine learning approaches all have separate inputs' configurations for one step ahead and day ahead forecast. Namely the machine learning input in Eq. (8) (the last column of the training label) use occupancy information of four time steps before the prediction step to train the model. This kind of inputs' configurations catches the key information that may help predict the irregular short absence or presence in the more stochastic pattern as shown in the occupancy patterns of Fig. 6. Meanwhile, the probability sampling approaches are developed based on a "day by day" simulation mechanism without such separated input configurations. Additionally, the light-switch uses the empirical rules to determine the periods of the morning, lunch, and afternoon during training [17]. However, a slight higher prediction accuracy is observed for the hierarchical sampling with a maximum 5% difference. The improvement of the hierarchical probability sampling is due to the optimizations of the morning,



(a) Office A

(b) Office B

(b) Office C

(d) Office D

(e) Four offices

Fig. 6. Occupancy rates of the offices samples.

lunch, afternoon periods. Nonetheless, from the highlighted red numbers, the proposed Markov model outperforms other models at most prediction horizons for 15-min, 30-min, and 1-h cases.

Day ahead predictions are provided based on two different time resolutions (15-min and 1-h) for all methods. The need for different time resolutions of predictive control design are discussed

Table 1
Comparison between models in terms of the correctness.

Methods		UTSA Markov (%)	Page's Model (%)	Hierarchical Model (%)	Reinhart's Model (%)	ANN (%)	SVR (%)
Office A	15-min	78.16	66.69	66.36	63.71	75.35	75.76
	30-min	72.28	68.78	69.23	65.91	73.81	71.54
	1-h	68.00	63.16	65.25	62.53	69.77	70.56
	24-h(15)	64.93	63.58	66.59	63.61	69.41	69.69
	24-h(1 h)	67.07	64.82	65.01	62.43	69.58	69.52
Office B	15-min	87.59	86.23	73.02	70.52	81.96	81.92
	30-min	82.41	81.58	76.27	71.34	79.76	79.80
	1-h	81.68	81.58	80.02	78.38	74.29	74.51
	24-h(15)	67.40	72.47	73.62	69.69	73.36	74.28
	24-h(1 h)	63.15	74.31	78.75	78.53	75.29	75.20
Office C	15-min	84.50	87.57	74.11	74.13	84.09	81.41
	30-min	81.89	78.57	70.50	70.25	71.89	78.05
	1-h	78.65	71.14	71.43	71.77	76.34	74.80
	24-h(15)	64.41	69.40	73.68	74.92	73.80	75.33
	24-h(1 h)	69.65	68.13	69.41	69.90	75.89	73.65
Office D	15-min	85.22	78.71	67.11	63.13	76.41	76.57
	30-min	79.34	71.03	64.37	60.25	75.18	75.05
	1-h	75.86	71.94	69.62	66.12	72.59	72.09
	24-h(15)	69.41	70.74	66.37	62.33	70.20	72.22
	24-h(1 h)	71.51	71.32	70.93	67.28	70.62	72.37

early in Section 2. It is clearly that ANN and SVR models are capturing the patterns in a 24-h scale shown by the red numbers of the one day ahead predictions. The Markov models does not outperform machine learning approaches while probability sampling approaches are competitive for some samples. For example, the hierarchical mode outperforms ANN and SVR for 24-h ahead forecast in 1-h resolution and most of the prediction errors of the day-ahead cases are not significantly different than the machine learning approaches. In summary, for extremely short term forecast from 15-min to 1-h ahead, the Markov model is recommended while the machine learning approaches are suggested for 24-h ahead forecasts.

The authors further evaluate the first arrival, the last departure, the duration of intermediate absences, and the starting times of intermediate departures. One thing should be noticed that these evaluations are not available to be evaluated based on the one step forecasts. For example, the first arrival time may be only useful for the preheating and precooling of the conditioning systems. If perdition is too short at the time scale, the thermal comfort will be largely influenced and mostly be violated due to the slow response of the building thermal envelope. Hence, only day ahead predictions are evaluated and results for 15-min resolutions are provided in Table 2. The error of the first arrival time is calculated by subtracting the ground truth from the predicted values. Same applies to last departure. The error of the intermediate departure time, however, is calculated by an absolute value between the prediction and the ground truth. The occupancy duration is estimated by counting the number of occupied intervals.

Comparing the models' predictions in Table 2, there is no single model consistently outperforming other models but machine learning approaches show marginal advantages. The probability sampling models (hierarchical model and Reinhart's model) tend to have better predictions of the first arrival and the last departure for the occupancy shape of the sharp summits (Office A and Office C). The errors stays in the maximum range around −0.95 h to +0.79 h with smallest departure error of 0.44 h for Office A and 0.35 h for Office C. Machine learning approaches could achieve a similar performance with better performances on the first arrival time predictions (0.44 h for Office A and 0.18 for Office C). Machine learning approaches also tend to have competitive performances on the predictions of occupancy durations and the intermediate departures for the occupancy shape of the two sharp summit, especially Office C (indicated by red numbers). Office B has a high presence rate with a gradually increasing and decreasing ramps indicating a shorter consistent presence period (12p.m. to 3p.m.). The Markov models (Page's model and UTSA model) show better predictive performances with low errors of occupancy duration (0.96 h) and the inter departure time (0.51 h). Office D has a very smooth presence rate but lower in the presence (around 0.5) and more random at the first arrival and the last departure (long ramps). Such scenario creates difficulties for all models. It is obvious that the absolute errors are larger than 2 h for most predictions regardless of the model used, except machine learning approach, such as ANN.

4.2. Prediction of occupant level

The data used to test the model's ability to forecast the number of occupants was collected from a student laboratory in the same building tested for occupancy presence predictions. The test period is from February 1st through June 30th, 2016. All the absence days without any occupancy were eliminated. The average occupancy number and the variances are presented in Fig. 7 using 1.5 interquartile range (99% confidence interval) of box-and-whisker plots. The blue box in Fig. 7 for each time marks the full range of variation from the 1st quartile to the 3rd quartile of the data (total of 50% of the data covered from the first quartile to the third quartile). Additional 1.5 quartiles beyond the first and third quartile as the black lines shown are expanded to cover the 99% of the data at that time. Daily analyses reveal the difference of occupancy number based on the working days, as shown from Fig. 7(a)–(e). These plots illustrated the occupancy of all the days from Monday to Friday. Monday, Wednesday and Thursday have very similar occupancy patterns (blue boxes in (a), (c) and (d)), where there are large variances of occupants' number during the daytime for Wednesday, Thursday and Friday (black lines in (c), (d), and (e)). Overall, gradual ramp-up and ramp-down are observed on most days between 7 am to 6 pm, except Tuesday. On Tuesday, more frequent meeting outside laboratory is scheduled which explains the low occupancy level (maximum 4 people) and less varied pattern (fewer black lines besides blue boxes). By averaging all occupancy numbers for all days, a gradual ramp-up is observed between 7 am to 11 am while a ramp-down is identified from 2 pm to 6 pm in Fig. 7(f). This indicates a very irregular first arrival time and last departure time. Large variances are also observed during the whole occupied period. The average variance is more than 2 people at each time's occupancy around the median (blue boxes of Fig. 7(f)). This is

Table 2
Comparison between models in terms of the statistical means.

		UTSA Markov (hr)	Page's Model (hr)	Probabilistic Model (hr)	Reinhart's Model (hr)	ANN (hr)	SVR (hr)
Office A	First Arrival Time	1.22	−0.89	−0.53	−0.95	0.44	−0.67
	Last Departure Time	0.73	0.71	0.62	0.70	0.76	0.85
	Occupancy Duration	0.36	4.38	2.19	2.33	−0.47	0.39
	Inter Departure Time	2.7	2.50	1.05	1.87	1.49	1.33
Office B	First Arrival Time	−2.69	−1.81	−2.39	−3.35	−1.63	−1.82
	Last Departure Time	2.15	2.46	2.39	2.74	2.170	2.44
	Occupancy Duration	0.96	1.82	2.87	2.63	1.49	1.43
	Inter Departure Time	0.97	0.51	1.08	1.88	1.17	0.58
Office C	First Arrival Time	1.14	−0.92	−0.35	−0.20	0.18	−0.53
	Last Departure Time	1.47	0.41	0.35	0.79	0.84	1.25
	Occupancy Duration	0.40	1.70	1.88	1.11	−0.09	0.91
	Inter Departure Time	2.37	2.20	2.94	4.13	2.03	2.30
Office D	First Arrival Time	−1.15	−2.16	−3.78	−3.54	−1.46	−1.56
	Last Departure Time	1.63	2.08	3.78	3.82	1.31	1.80
	Occupancy Duration	2.42	2.01	2.91	2.98	1.63	2.09
	Inter Departure Time	2.84	2.18	3.65	2.54	1.67	2.43

Table 3
Comparison between models.

Methods		UTSA Markov (hr)	Hierarchical Model (hr)	ANN (hr)	SVR (hr)
RMSE	15-min	0.510	1.135	1.283	1.241
	30-min	0.736	1.303	1.394	1.281
	1-h	1.046	1.280	1.463	1.388
	24-h(15 min)	1.052	1.146	1.387	1.088
	24-h(1 h)	1.069	1.255	1.588	1.127
MAE	15-min	0.204	0.675	0.730	0.709
	30-min	0.327	0.711	0.807	0.744
	1-h	0.523	0.666	0.951	0.839
	24-h(15 min)	0.524	0.625	0.849	0.679
	24-h(1 h)	0.567	0.629	1.061	0.705

caused by the different working schedules from individual student. The maximum occupancy contains 6 students (black lines up to 6 occupancy level of Fig. 7(f)).

We applied two widely used evaluation error criteria, the root mean square error (RMSE) and the mean absolute error (MAE), to quantify the performances of the predictions:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \tag{11}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{12}$$

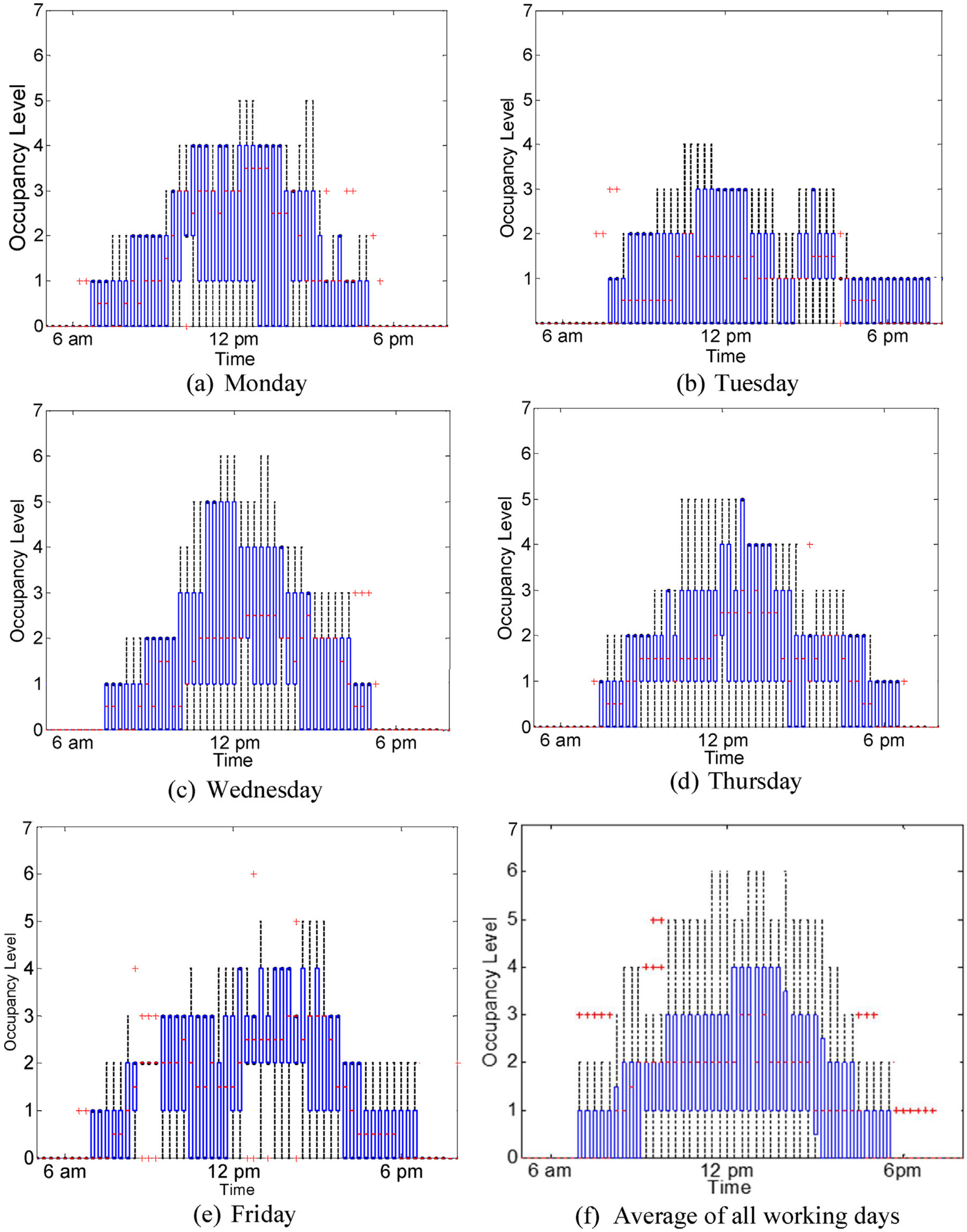
Where y_i and $\hat{y}_i$ are the actual and predicted occupancy number at each time step, and N is the total length of the data. The first error RSME amplifies and severely punishes large errors using the square form while the second error MAE provides a view on how close the forecasts and the measurements are in absolute scale. Predictions under the two error criteria of all forecast horizons are shown in the following table.

As shown in Table 3, the occupancy number can be more accurately predicted (0.204–0.523 by MAE error) in the extremely short-term forecast (e.g. 15-min till 1-h) based on the proposed Markov model. Poor performances are observed for other models while Hierarchical Probability Sampling Model has a similar performance compared to the proposed Markov model in 1-h ahead case (0.666 MAE and 1.280 RMSE). In terms of RMSE, the proposed Markov model has large errors when prediction horizons are extended (e.g. from 1-h ahead to 24-h ahead). It shows the difficulty for the proposed model to predict longer period of the dynamic occupancy pattern in this study. The occupancy of the samples has an average variance of 2 people at each time step, shown as the 1st quartile to the 3rd quartile of the data deviated from the median (blue boxes of Fig. 7(a)). It is also noticed that the Hierarchical Probability Sampling and SVR have a very similar predictive performances at 24-h ahead forecasting cases although the proposed Markov model still has marginal improvements on both RMSE and MAE. In general, the prediction for 24-h ahead case in 15-min resolution is easier than the cases in 1-h resolution for all the methods (e.g. all the errors of 24-h ahead cases in 15-min resolutions being smaller than the cases in 1-h resolutions). Meanwhile, it is noted that the errors in this study for longer prediction window (24-h ahead) are comparable to a recent research study, which reported the error ranges from 0.53 to 1.27 in terms of RMSE, and from 0.21 to 0.80 in terms of MAE [49].

5. Results and Discussions

In general, the real occupancy patterns in building environments may differ significantly from each other as shown in Figs. 6 and 7. Meanwhile, most studies to model occupancy in commercial buildings usually focus on one part of the information for limited samples, such as presence only or occupants' number for single office space. By realizing the issues, IEA-EBC Annex 66 is promoting a broader and systematic definition and simulation of occupancy behavior in buildings [56,57]. Recent studies are expanding the limitations of the test-beds (e.g. multiple office rooms) or the prediction horizon (e.g. day-ahead forecast) [35–40]. Although many still argued that a single model is unlikely to be general enough to cover all solutions and not necessary for building simulations [57], the advanced predictive controls of the building systems discussed in Section 2 demonstrated the need to develop occupancy models able to forecast in different time scales and occupancy types. For real-time control applications, an integrated occupancy model, such as the proposed hierarchical sampling model and the new developed Markov model could be adapt enough to provide universal solution from extremely short term forecasts (e.g. 15-min ahead) to day-ahead predictions (e.g. 24-h ahead) of different occupancy information including the presence and the number.

One of the key contributions of this study is to investigate whether the proposed Markov Model and Hierarchical Probability Sampling Model are adaptive enough to handle the temporal changes at various prediction horizons (15-min ahead to 24-h ahead) and different types of the occupancy (the presence and number of occupancy). The proposed Markov model is specifically designed to handle not only binary occupancy states (presence and



(a) Monday (b) Tuesday

(c) Wednesday (d) Thursday

(e) Friday (f) Average of all working days

Fig. 7. Occupancy number of the laboratory.

absence) but also multiple occupancy level (the number of occupants). The moving window optimization for different prediction horizons can find the change of the presence patterns based on change-point analysis and the cross validations insure the optimal estimations of the transition probabilities of the model based on historical information. If only the binary forecast is needed, the proposed Markov model states can only use the binary presence information. Otherwise, the Markov chain states will fully utilize the occupants' number information. The study further expands the traditional probability sampling model to create a hierarchical framework that could produce occupancy number predictions based on the presence forecast. By comparison to other state-of-the-art models, results in Section 3 suggest that the proposed models are competitive for the applications of the predictive control.

Another contribution is the investigation on the popular occupancy models' predictive performances on different patterns of occupancy presence. The authors evaluated different samples including twin summits shape and single summit shape with high and low occupancy rate at different periods of time. The predictive

Table 4
Comparison between models and studies.

Study	Method	First Arrival (hr)	Last Departure (hr)	Occupancy State Matching (%)	Occupancy Duration (hr)
Mahdavi [11]	Reinhart	1.2	2.4	52	2.3
	Page	1.4	2.4	52	2.2
	Aggregated Optimized Schedule	1.0	2.4	55	1.6
UTSA Offices[a]	UTSA Model	1.5	1.5	76	1.0
	Hierarchical Model	1.8	1.7	72	2.5
	Page	2.9	2.7	69	2.9
	Reinhart	2.0	2.0	66	2.1

[a] Absolute values of the average errors of the all four offices.

performances of the models are evaluated from different performance matrixes: the prediction correctness, the first arrival error, the last departure error, the occupancy state one-to-one matching error, and the occupancy duration of the intermediate presence [24]. Detailed analysis in Section 3.1 suggests the possible impacts from the occupancy diversities on the models' performances. Different models have their own advantages to predict certain types of occupancy presence patterns. Meanwhile, the average errors of all four offices are compared to a recent study conducted in the office environment [24], as shown in Table 4. It is noticed again that there are significant differences between the models' performance due to different samples of the occupancy presence profiles even for the same models (e.g. Page and Reinhart's models in Mahdavi's study and this study). For the same prediction settings (24-h ahead in 15-min resolution) of this study and a recent reported study, the results of the proposed Markov model and aggregated optimized schedule have slight higher predictive powers.

6. Conclusion

This paper aims to develop and demonstrate an innovative Markov approach and an integrated sampling model to forecast occupancy of office buildings. By predicting future occupancy presence and occupancy number at different time scales (15-min to 24-h ahead), the proposed Markov model and the integrated sampling model demonstrate their predictive power specifically for the purpose of control application. The results are validated through long term measured data from the field tests of the offices and compared to other commonly used models for occupancy predictions, such as the Page's model, Reinhart's model, Artificial Neural Network and Support Vector Regression. The final results show that the proposed models outperform the other methods in terms of an average 7% correctness with 22% maximum difference for one time step ahead forecast of the occupancy presence. Meanwhile, maximum 0.34 RMSE and 0.23 MAE differences for the occupancy number predications are observed at all time steps. In day ahead prediction, not much difference could be concluded among the models. Artificial Neural Network and Support Vector Regression tend to have slightly better performance for presence predictions for day-ahead cases and the proposed Markov model could outperform them to predict occupancy number. In conclusion, implementing such adaptive occupancy models will be a solution for integrated predictive controls that handle multiple building optimizations.

This study also observes a significantly lower performance for 24-h ahead prediction scenario compared to the other prediction window (e.g. 15-min to 1-h ahead) for both occupancy presence and number predictions. It is extremely necessary for advanced building control to have more accurate forecasts on longer window of the occupancy presence such as the day ahead (24-h) forecast when renewable building systems are involved. However, it is challenging to improve the forecast accuracy even with the changes of temporal resolution (sampling rate) between 15-min and 1-h resolution. But the results show competitive performances compared to recent studies [24,51]. Another limitation related to the proposed model is the lack of cross sectional comparison when testing the forecast ability of the models for both the presence and the occupancy level. The authors already deployed more than ten occupancy detection systems of the test building. Owing to the privacy and installation issues, four of the offices were able to have the occupancy presence information collected while only one lab successfully record long term pattern of the occupancy number. The total sample size is abundant compared to previous research, but more samples from different occupants should be included to further validate the generalization of the proposed models.

Further investigation on improvements of the day ahead predictions could be conducted by the more advanced statistical inference or analysis to detect the irregular dynamics at daily scale in the occupancy data. By evaluating the uncertainty of the occupancy patterns, a classification process of the data before the predictions could also help to realize the difficulty to forecast specific types of occupancy. The last but not the least, software packages for the models should be developed on an open source programming language (e.g. Python) which is much preferred by the real-time control tests.

Acknowledgements

This paper is based on the research supported by the IEA-EBC Annex 66 (Definition and Simulation of Occupant Behavior in Buildings) under Subtask A (Occupant movement and presence models in buildings), and National Science Foundation (NSF) under *EAGER: Collaborative Research: Empowering Smart Energy Communities: Connecting Buildings, People, and Power Grids,* with award number: CBET-1637249. Any results and material reported in this work are authors research and do not necessarily reflect the views of the IEA-EBC and NSF.

References

[1] U.S. Department of Energy, Energy Efficiency Trends in Residential and Commercial Buildings, 2010 (Available at:) https://www1.eere.energy.gov/buildings/publications/pdfs/corporate/building_trends_2010.pdf.

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

[3] N.E. Klepeis, W.C. Nelson, W.R. Ott, J.P. Robinson, A.M. Tsang, P. Switzer, J.V. Behar, S.C. Hern, W.H. Engelmann, The National Human Activity Pattern Survey (NHAPS): a Resource for Assessing Exposure to Environmental Pollutants, 2001 (Available at) https://indoor.lbl.gov/sites/all/files/lbnl-47713.pdf.

[4] National Grid, Managing Energy Costs in Office Buildings, 2002 (Available at:) https://www9.nationalgridus.com/non_html/shared_energyeff_office.pdf.

[5] Residential energy end-use splits by fuel type (quadrillion BTU): http://buildingsdatabook.eren.doe.gov/TableView.aspx?table=1.1.4.

[6] U.S. Census Bureau. https://www.census.gov/construction/chars/highlights.html.

[7] A. Mirakhorli, B. Dong, Occupancy behavior based model predictive control for building indoor climate-A critical review, Energy Build. 129 (2016) 499–513.

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

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

[10] M.A. 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.

[11] X. Guo, D.K. Tiller, G.P. Henze, C.E. Waters, The performance of occupancy-based lighting control systems: a review, Light. Res. Technol. 42 (4) (2010) 415–431.

[12] X. Peng, P. Have, M.A. Piette, J.E. Braun, Peak demand reduction from pre-cooling with zone temperature reset in an office building, in: ACEEE Summer Study on Energy Efficiency in Buildings, Pacific Grove, CA, U.S.A., 2004.

[13] R. Yang, M.W. Newman, Living with an intelligent thermostat: advanced control for heating and cooling systems, in: HomeSys Workshop: Ubicomp, Seattle, U.S.A., 2012.

[14] J. Shi, N.P. Yu, W.X. Yao, Energy efficient building HVAC control algorithm with real-time occupancy prediction, Energy Procedia 111 (2017) 267–276.

[15] R. Halvgaard, N.K. Poulsen, H. Madsen, Economic model predictive control for building climate control in a smart grid, in: Proceedings of the 2012 IEEE PES Innovative Smart Grid Technologies, Washington D.C., U.S.A, 2012.

[16] J. Dobbs, B. Nencey, Predictive HVAC control using a Markov occupancy model, in: Proceedings of the American Control Conference, Portland, Oregon, U.S.A, 2014.

[17] C.F. Reinhart, Daylight availability and manual lighting control in office buildings −simulation studies and analysis of measurements, in: Ph.D. Thesis, Technical University of Karlsruhe, Faculty of Architecture, 2001.

[18] C.F. Reinhart, Lightswitch-2002: a model for manual and automated control of electric lighting and blinds, Sol. Energy 77 (2004) 15–28.

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

[20] K.Y. Sun, D. Yan, T.Z. Hong, S.Y. Guo, Stochastic modeling of overtime occupancy and its application in building energy simulation and calibration, Build. Environ. 79 (2014) 1–12.

[21] V. Tabak, B. De Vrie, Methods for the prediction of intermediate activities by office occupants, Build. Environ. 45 (2010) 1366–1372.

[22] A.S. Silva, E. Ghisi, Uncertainty analysis of user behaviour and physical parameters in residential building performance simulation, Energy Build. 76 (2014) 381–391.

[23] W.K. Chang, T.Z. Hong, Statistical analysis and modeling of occupancy patterns in open-plan offices using measured lighting-switch data, Build. Simul. 6 (1) (2013) 23–32.

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

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

[26] C. Wang, D. Yan, Y. Jiang, A novel approach for building occupancy simulation, Build. Simul. 4 (2) (2011) 149–167.

[27] J. Widen, A. Molin, K. Ellegard, 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.

[28] M.A. Lopez-Rodriguez, I. Santiago, D. Trillo-Montero, J. Torriti, A. Moreno-Munoz, Analysis and modeling of active home occupancy of the residential sector in Spain: an indicator of residential electricity consumption, Energy Policy 62 (2013) 742–751.

[29] U. Wilke, F. Haldi, Scartezzini, D. Robinson, A bottom-up stochastic model to predict building occupants' time-dependent activities, Build. Environ. 60 (2013) 254–264.

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

[31] U. Wilke, Probabilistic Bottom-up Modelling of Occupancy and Activities to Predict Electricity Demand in Residential Buildings Dissertation, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, 2013.

[32] K.P. Lam, M. Hoynck, R. Zhang, B. Andrews, Y.S. Chiou, B. Dong, D. Benitez, Information-theoretic environment features selection for occupancy detection in open office spaces, in: Building Simulation 2009, Glasgow, Scotland, 2009.

[33] G. Flett, N. Kelly, An occupant-differentiated, higher-order Markov Chain method for prediction of domestic occupancy, Energy Build. 125 (2016) 219–230.

[34] B. Ai, Z.Y. Fan, R.X. Gao, Occupancy estimation for smart buildings by an auto-regressive hidden Markov model, in: American Control Conference, Portland, Oregon, U.S.A, 2014.

[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, Build. Simul. 7 (1) (2013) 89–106.

[36] V.L. Erikson, Y.Q. Lin, A. Kamthe, R. Brahme, A. Surana, A.E. Cerpa, M.D. Sohn, S. Narayanan, Energy efficient building environment control strategies using real-time occupancy measurements, in: Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, Berkeley, California, U.S.A, 2009.

[37] V.L. Erikson, A.E. Cerpa, Occupancy based demand response HVAC control strategy, in: Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-efficiency in Building, Zurich, Switzerland, 2010.

[38] V.L. Erickson, M.A. Carreira-Perpinan, A.E. Cerpa, OBSERVE: occupancy-Based system for efficient reduction of HVAC energy, in: 10th International Conference on Information Processing in Sensor Network, Chicago, Illinois, U.S.A., 2011.

[39] V.L. Erikson, M.A. Carreira-Perpinan, Occupancy modeling and prediction for building energy management, ACM Trans. Sens. Netw. 10 (3) (2014) (article42).

[40] J.R. Dobbys, B.M. Hencey, Model predictive HVAC control with online occupancy model, Energy Build. 82 (2014) 675–684.

[41] 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.

[42] M. Baptista, A.J. Fang, H. Prendinger, R. Prada, Y. Yamaguchi, Accurate household occupant behavior modeling based on data mining techniques, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Quebec, Canada, 2014.

[43] S. D'Oca, T.Z. Hong, Occupancy schedules learning process through a data mining framework, Energy Build. 88 (1) (2015) 395–408.

[44] C.J. Andrews, D. Yi, U. Krogmann, J.A. Senick, R.E. Wener, Designing building for real occupants: an agent-based approach, IEEE Trans. Syst. Man Cybernetics-Part A: Syst. Hum. 41 (6) (2011) 1077–1091.

[45] E. Azar, C.C. Menassa, Agent-based modeling of occupants and their impact on energy use in commercial buildings, J. Comput. Civil Eng. 26 (4) (2012).

[46] T. Yu, Modeling occupancy behavior for energy efficiency and occupants comfort management in intelligent buildings, in: Ninth International Conference on Machine Learning and Applications, Washington D.C., U.S.A., 2010.

[47] R. Zhang, K.P. Lam, Y.S. Chiou, B. Dong, Information-theoretic environment features selection for occupancy detection in open office spaces, Build. Simul. 5 (2) (2012) 179–188.

[48] B. Dong, K.P. Lam, Building energy and comfort management through occupant behavior pattern detection based on a large-scale environmental sensor network, J. Build. Perform. Simul. 4 (4) (2010) 359–369.

[49] Z.H. Che, Y.C. Soh, Comparing occupancy models and data mining approaches for regular occupancy prediction in commercial buildings, J. Build. Perform. Simul. (2016) 1–9.

[50] Z.X. Li, B. Dong, A new modeling approach for short-term predictions of occupancy presence in residential buildings, Build. Environ. 121 (2017) 277–290.

[51] 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.

[52] MathWorks, Neural Network Toolbox, 2017 (Available at:) https://www.mathworks.com/products/neural-network.html.

[53] C.C. Chang, C.J. Lin, LIBSVM-A Library for Support Vector Machines, 2016 (Available at:) https://www.csie.ntu.edu.tw/∼cjlin/libsvm/.

[54] Z. Li, S.M. Mahbobur Rahman, R. Vega, B. Dong, A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting, Energies (2016) (Special Issue: Forecasting Methods and Measurements of Forecasting Errors for Renewable Energy Sources).

[55] B. Dong, Z.X. Li, S.M. Rahman, R. Vega, A hybrid model approach for forecasting future residential electricity consumption, Energy Build. 117 (1) (2016) 341–351.

[56] Energy in Buildings and Communities Programme. IEA-EBC Annex 66: Definition and Simulation of Occupant Behavior in Buildings. Available at: https://www.annex66.org/.

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