This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Jiang, Chaoyang; Chen, Zhenghua; Su, Rong; Khalid Masood, Mustafa; Soh, Yeng Chai
Bayesian filtering for building occupancy estimation from carbon dioxide concentration

Published in:
Energy and Buildings

DOI:
[10.1016/j.enbuild.2019.109566](https://doi.org/10.1016/j.enbuild.2019.109566)

Published: 01/01/2020

Document Version
Peer reviewed version

Published under the following license:
CC BY-NC-ND

Please cite the original version:
Jiang, C., Chen, Z., Su, R., Khalid Masood, M., & Soh, Y. C. (2020). Bayesian filtering for building occupancy estimation from carbon dioxide concentration. *Energy and Buildings*, 206, [109566].
<https://doi.org/10.1016/j.enbuild.2019.109566>

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Journal Pre-proof

Bayesian filtering for building occupancy estimation from carbon dioxide concentration

Chaoyang Jiang, Zhenghua Chen, Rong Su, Mustafa Khalid Masood, Yeng Chai Soh

PII: S0378-7788(19)31272-1
DOI: <https://doi.org/10.1016/j.enbuild.2019.109566>
Reference: ENB 109566

To appear in: *Energy & Buildings*

Received date: 25 April 2019
Revised date: 26 July 2019
Accepted date: 27 October 2019

Please cite this article as: Chaoyang Jiang, Zhenghua Chen, Rong Su, Mustafa Khalid Masood, Yeng Chai Soh, Bayesian filtering for building occupancy estimation from carbon dioxide concentration, *Energy & Buildings* (2019), doi: <https://doi.org/10.1016/j.enbuild.2019.109566>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

HIGHLIGHTS

- The proposed framework can fuse a statistical model and an observation model for better occupancy estimation.
- The statistical model can capture the temporal dependency of the building occupancy.
- Observation model can estimate the occupancy level from carbon dioxide concentration.
- A novel ensemble learning method is proposed for observation model identification.

Journal Pre-proof

Bayesian filtering for building occupancy estimation from carbon dioxide concentration

Chaoyang Jiang, Zhenghua Chen, Rong Su, Mustafa Khalid Masood, and Yeng Chai Soh

Abstract

This paper proposes a new framework based on Bayesian filtering for building occupancy estimation from the observation of carbon dioxide concentration. The proposed framework can fuse a statistical model and an observation model for better occupancy estimation. The statistical model can capture the temporal dependency of the building occupancy, and the first-order inhomogeneous Markov model is utilized for the estimation of occupancy transition probability. The observation model can estimate the occupancy level from carbon dioxide concentration. The likelihood is obtained from the solution of the observation model. To identify the observation model, we present a novel ensemble extreme learning machine technique. Applying the Bayes filter technique, we can fuse the transition probability and the likelihood for better occupancy estimation. The proposed framework can be applied for general cases of occupancy estimation, and the solution outperforms the results of the observation model. The results of a real experiment show the effectiveness of the proposed method.

Index Terms

Building occupancy estimation, carbon dioxide concentration, Bayesian filtering, inhomogeneous Markov model, ensemble extreme learning machine.

This work was supported in part by Beijing Institute of Technology Research Fund Program for Young Scholars, in part by and in part by the Building and Construction Authority (BCA) of Singapore through the NRF GBIC Program with the project reference NRF2015ENC-GBICRD001-057.

C. Jiang is with the School of Mechanical Engineering, Beijing Institute of Technology, 5 South Zhongguancun Street, Beijing, China 100081 (e-mail: cjiang@bit.edu.cn)

Z. Chen is with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, Singapore 138632 (e-mail: chen0832@e.ntu.edu.sg)

R. Su and Y. C. Soh are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: rsu@ntu.edu.sg; eycsoh@ntu.edu.sg).

M. K. Masood is with the Secure Systems group at Aalto University, Finland. (e-mail: mkm_imn@hotmail.com).

I. INTRODUCTION

Buildings consume around 40% of the total energy in the world when providing comfortable and healthy indoor environments [1]. Heating, ventilation, air-conditioning, and lighting (HVAC&L) systems account for the largest proportion of energy consumption in buildings. The occupancy-based control of HVAC&L systems offers considerable potential for energy saving [2]–[5]. The absence and presence of occupants can suggest the active periods of HVAC&L systems [3]–[5]. During the active periods, the real-time occupancy level is a key input for the demand-driven HVAC&L operations [6]–[9]. Therefore, occupancy information is an essential component of energy-efficient HVAC&L systems in intelligent buildings.

A. Related prior works

Numerous methods have been proposed to estimate indoor occupancy. Vision-based methods [10] can provide very good results but suffer from high equipment costs and privacy concerns. Terminal-based methods require that each occupant uses a device such as a keyboard and mouse [11], smartphones with WiFi [2], [9] or radio frequency identification tags [12]. The data security in terminal-based methods is an issue and the estimation result is actually the number of devices, which may not equal the number of occupants. The pyroelectric infrared sensors can detect the movement of occupants but cannot provide the occupancy level [13]. Other sensors such as chair sensors [8] and smart meters [14] only detect the occupants performing specific activities. Consequently, it is very interesting to develop a non-intrusive and low-cost method for the estimation of occupancy levels.

Environmental sensors such as temperature, humidity and carbon-dioxide sensors are non-intrusive and available for many HVAC systems. No additional cost is required for environmental data acquisition. Hence, environmental sensor-based methods have become a popular way for occupancy estimation [15]–[22]. In these papers, various machine learning algorithms such as Naive Bayes classifier [15], hidden Markov model [16], [17], artificial neural network [17], decision tree [18], k -nearest neighbours [18], [19], linear discriminant analysis [19], [20], and extreme learning machine (ELM) [21], [22] have been applied to infer the occupancy level.

Among all environmental parameters, the carbon-dioxide concentration has the highest correlation with the number of occupants [23]. Assuming that the indoor carbon-dioxide concentration is uniform, we can directly obtain the occupancy level from the mass balance equation of the carbon-dioxide concentration [25]–[28], which is a first-order ordinary differential equation in terms of the carbon-dioxide concentration. In this method, the estimation results depend on the accuracy of calculated carbon-dioxide differentials, resulting in serious errors because of the uncertainties of carbon-dioxide measurements and

the unpredictable openings of doors and windows [23]. Based on the mass balance equation, a recent work provided a stochastic differential equation for occupancy estimation, which greatly improve the estimation accuracy [24]. Instead of the mass balance equation, the authors in [23] identified a carbon-dioxide-based occupancy estimator using a machine learning technique. In addition, carbon-dioxide sensors are often combined with non-environmental sensors such as cameras [29], computer power consumption [30], power meters and microphones [31] for occupancy estimation. More discussions about the building occupancy estimation can be found in the survey paper [32].

Many recent studies estimate the indoor occupancy level from various types of sensor observations by learning a model/function/classifier, which formulates the relationship between the occupancy levels and the corresponding sensor observations [32]. In this paper, we call the model/function/classifier as the *observation model* of occupancy levels. In addition, for many application scenarios such as offices and class rooms, the occupancy level can be estimated without any sensor observations. We can view each day's occupancy level of a certain room as a time series. Without any sensor observations, based on the statistical information of historical data, the number of indoor occupants can be predicted by using some statistical technique such as agent based models [33], Markov chain models [34], [35], and generative adversarial network [36]. We call these models the *statistical models* of occupancy levels.

B. Our contributions

The occupancy information is derived from the sensor measurements when using the *observation model*, while the *statistical model* provide the occupancy information based on a temporal model such as a time-series. The information sources of the two models are totally different, and each of them can provide different information. Hence, it has a large potential to achieve better occupancy estimation if we fuse the two models. In this paper, we propose a new framework in which we fuse the results of a statistical model and an observation model via Bayes filter for better occupancy level estimation. Bayes filter, also known as recursive Bayesian filtering, can provide real-time posterior probability density/mass function (pdf/pmf) of the state (occupancy level) based on all available information, including the initial pdf/pmf, the transition probability matrix obtained from the statistical model and the likelihood obtained from the observation model. The solution of Bayes filter is viewed as 'optimal' because the optimal estimate with respect to any criterion can be obtained from the posterior pdf/pmf [37], [38]. When using Bayes filter, the state is assumed to follow a first-order Markov process. Due to the Markov property of occupancy behavior, Markov chains were commonly used to capture the temporal dependency of indoor occupancy [16], [17], [34], [35]. Hence, Bayes filter can be naturally used to combine the statistical model and the observation model for indoor occupancy estimation.

Since the probability of each day's occupancy level is time-dependent, instead of the homogeneous Markov model used in prior works [35], [39], we leverage on the first-order inhomogeneous Markov chain to formulate the statistical model. Correspondingly, a series of time-varying (inhomogeneous) transition probability matrices are used to reveal the temporal dependency of indoor occupancy.

We propose a difference equation in terms of the occupancy level to formulate the observation model, in which the occupancy level at a particular time instant depends on the measured carbon-dioxide concentration and the estimated occupancy level in a past time horizon. The ELM algorithm has been widely used for occupancy estimation [16], [21]–[23], especially for feature selection in occupancy estimation [16], [21], [22] due to its fast learning-speed. In the ELM algorithm, however, the input-layer weights are randomly generated [40], which results in diverse solutions. To address the solution diversity of the ELM model, ensemble methods are very effective choices [41]. Hence, we select the ensemble ELM algorithm to identify the observation model. Here, 'ensemble' means that we run many different ELMs and then find the final solution based on the voting result of all ELM solutions. Considering the extremely fast learning-speed of ELM, the computational cost of the ensemble ELM is still not expensive.

From the results of the observation model, we can obtain the likelihood of the occupancy level. With the transition probability matrix and the likelihood, we can achieve the posterior pmf via Bayes filter. The optimal estimation in the sense of minimum mean-squared-error (MMSE) is the conditional mean of occupancy level, which can be easily calculated from the pmf.

We tested the proposed method in an air-conditioned office room. The experimental results show that 1) the 'ensemble' strategy is effective; 2) the Bayes filter outperforms the observation model; and 3) the Bayes filter outperforms the state-of-the-art approaches in [16]–[18], [20], [21], [23].

II. PROBLEM STATEMENT

A. Experiment description

In this work, we measured the carbon-dioxide concentration in the office space of a laboratory whose layout is shown in Figure 1. The size of the office space is 20m long, 9.3m wide and 2.6m high. The office space consists of 24 office cubicles for research students/staff and 11 open seats for undergraduate students. The five open seats in the corner are rarely occupied.

The office space is conditioned using both variable air volume and active chilled beam systems, and is mechanically ventilated using an air handling unit which delivers a constant supply of fresh air. The active chilled beam was rarely used while both the variable air volume system and the mechanical ventilation system were active everyday during 7:30AM to 10:00PM.

Figure 1: The layout of the test-bed.

We used a CL11 sensor from Rotronic, which was placed at the central area of the office space (see the red dot in Figure 1), to measure the carbon-dioxide concentration. The sensor was placed 1.1m above the floor. The resolution of the sensor is 1 parts-per-million (ppm), and the accuracy is $\pm 5\%$. The sampling time was 1 minute.

We used two Internet protocol cameras installed at two doors of the office space to record the entries and exits of occupants. The videos recorded by the two cameras were applied to obtain the number of occupants in the office space, which we used as the ground truth. Normally, people counting solutions perform automatic line or area counting from overhead cameras to obtain ground truth occupancy. However, some error accumulation is inevitable through this approach [41]. In order to get perfectly accurate ground truth data, we took a semi-automatic approach by applying highly sensitive motion detection using the iSpy software (which is freely available) to generate snapshots of the frames of the videos at every entry and exit. We then manually recorded the number of occupants corresponding to each entry/exit event. If we noted a non-zero occupancy for any midnight, we corrected the inconsistency by directly referring to the recorded video. We are thus confident that our ground truth data is error-free. Using this method, we collected the occupancy data for 30 weekdays.

B. Problem description: occupancy estimation based on carbon-dioxide observations

We aim to estimate the occupancy levels of the office space in our test-bed based on real-time carbon-dioxide measurements. Currently, it is difficult to find the exact number of indoor occupants

from environmental sensors. Part works show that if the maximum number of indoor occupants is more than 20, the accuracy tends to be around 50% [23], [28]. In a dense occupancy environment, many high cost sensors are required to achieve a high estimation accuracy with respect to the exact number. Fortunately, the alternative occupancy levels, i.e., zero, low, medium and high, are adequate for many demand-driven control systems (for HVAC&L systems) [7], [8]. For the control of HVAC systems in a large room such as large office and classroom, it is unnecessary to estimate the exact number of occupants. Overestimating/underestimating several occupants has insignificant influence on the control of HVAC systems. Hence, we introduce the concept of occupancy level.

Based on the occupancy data for 30 days in our experiment, we uniformly divide the occupancy level. For all 30 days, the maximum occupancy number in the office space was 28, which was only maintained for a few minutes. Hence, for each occupancy level we assigned around 9 occupants. In this way, increasing one occupancy level leads to an almost constant mean increment of the thermal load, which makes it easy for the demand-driven controller design of HVAC systems. The detail of the occupancy level is defined in Table I.

Table I: The definition of occupancy level

occupancy level (x)	number of occupants (o)
zero ($x = l_1$)	$o = 0$
low ($x = l_2$)	$0 < o \leq 9$
medium ($x = l_3$)	$9 < o \leq 18$
high ($x = l_4$)	$o > 18$

Theoretically, we can estimate the occupancy level per minute since the sampling time of carbon-dioxide concentration is 1 minute in our experiment. In practice, however, the requirement of the occupancy information in time domain depends on the operation frequency of the control systems. HVAC systems usually has a large time constant (around 10-20 minutes), which mainly depends on the size of the room. Hence, the operation time of HVAC control systems is also 10-20 minutes, and the requirement of occupancy information is in a corresponding time horizon. For example, the occupancy level per 15 minutes was estimated in [16] and the occupancy level per 20 minutes was estimated in [17], respectively.

In this work, occupancy estimation refers to the estimation of the occupancy level per 15 minutes. We aim to estimate the indoor occupancy level from the known information: the real-time carbon-dioxide observations and the real-time ON/OFF state of the ventilation system.

III. METHODOLOGY

A. Bayes filter

Bayes filter, also known as recursive Bayesian filtering, is considered as an optimal filter for a state estimation problem. We use a Bayes filter to recursively estimate the posterior *pmf* of the real-time occupancy level $p(x_n|\mathcal{Y}_n)$. Here, x_n is the occupancy level at the time instant t_n where $t_n = nT$, $n \in \mathbb{N}$ and T is the sampling time. $\mathcal{Y}_n$ represents a set of observations $\mathbf{y}_{0:n} := \{y_0, y_1, \dots, y_n\}$, and y_n is the observation of the occupancy level at time instant t_n , which can be obtained from the observation model.

Two assumptions are used in the Bayes filter:

- 1) the states follow a first-order Markov process, i.e., $p(x_n|\mathbf{x}_{0:n-1}) = p(x_n|x_{n-1})$;
- 2) the observation is conditionally independent of the previous observations given the current state, i.e., $p(y_n|\mathcal{Y}_{n-1}, x_n) = p(y_n|x_n)$.

The indoor occupancy level naturally follows a first-order Markov process. The observation model of an occupancy level can be formulated as

$$y_n = \hat{x}_n = x_n + \omega_n \quad (1)$$

where ω_n is the observation noise. If the *pmf* of ω_n is known, from (1) we can find that y_n is independent of $\mathcal{Y}_{n-1}$ given x_n . Later, we will detail the observation model.

From the Bayes' rule we can obtain the following conditional *pmf* of x_n :

$$\begin{aligned} p(x_n|\mathcal{Y}_n) &= \frac{p(\mathcal{Y}_n|x_n)p(x_n)}{p(\mathcal{Y}_n)} \\ &= \frac{p(y_n, \mathcal{Y}_{n-1}|x_n)p(x_n)}{p(y_n, \mathcal{Y}_{n-1})} \\ &= \frac{p(y_n|\mathcal{Y}_{n-1}, x_n)p(\mathcal{Y}_{n-1}|x_n)p(x_n)}{p(y_n|\mathcal{Y}_{n-1})p(\mathcal{Y}_{n-1})} \\ &= \frac{p(y_n|\mathcal{Y}_{n-1}, x_n)p(x_n|\mathcal{Y}_{n-1})p(\mathcal{Y}_{n-1})p(x_n)}{p(y_n|\mathcal{Y}_{n-1})p(\mathcal{Y}_{n-1})p(x_n)} \\ &= \frac{p(y_n|x_n)p(x_n|\mathcal{Y}_{n-1})}{p(y_n|\mathcal{Y}_{n-1})} \end{aligned} \quad (2)$$

Equation (2) shows that the posterior *pmf* $p(x_n|\mathcal{Y}_n)$ is determined by the three terms:

- Prior *pmf*:

$$p(x_n|\mathcal{Y}_{n-1}) = \sum_{i=1}^4 p(x_n|x_{n-1} = l_i)p(x_{n-1} = l_i|\mathcal{Y}_{n-1}) \quad (3)$$

where l_i represents the occupancy level, $[l_1, l_2, l_3, l_4] = [\text{zero}, \text{low}, \text{medium}, \text{high}]$, and $p(x_n|x_{n-1})$ represents the *transition probability* of the state, which can be obtained from the statistical model

(the first-order Markov model). In the Chapman-Kolmogorov equation (3), we use the fact that $p(x_n|x_{n-1}, \mathcal{Y}_{n-1}) = p(x_n|x_{n-1})$.

- Likelihood: $p(y_n|x_n)$ can be obtained from the observation model.
- Evidence:

$$p(y_n|\mathcal{Y}_{n-1}) = \sum_{i=1}^4 p(y_n|x_n = l_i)p(x_n = l_i|\mathcal{Y}_{n-1}). \quad (4)$$

The evidence can be directly obtained from the prior *pmf* and the likelihood. To find the prior *pmf*, the key work is the estimation of the transition probability. Therefore, our main work is to find the *transition probability* and the *likelihood*.

B. The statistical model: transition probability estimation

We view each day's occupancy level of the office space in our test-bed as a time series. The occupancy level has the Markov property. In other words, the occupancy level of the next time instant only depends on the state of the current time instant and is independent of the state of previous time instants. The Markov chain model has been shown to be effective to capture the temporal dependence in occupancy estimation [16], [17], [34], [35]. Considering the time-dependency of the transition probability in the occupancy dynamics discussed in [16], [17], we use the first-order inhomogeneous Markov chain to formulate the statistical model of the occupancy level.

Figure 2: The first-order inhomogeneous Markov chain model.

The diagram of the first-order inhomogeneous Markov chain is shown in Figure 2. The *pmf* of the occupancy level at the time instant t_n is

$$\mathbf{p}(x_n) = \mathbf{A}_{n|n-1}\mathbf{p}(x_{n-1}) \quad (5)$$

where

$$\mathbf{p}(x_n) = \begin{bmatrix} p(x_n = l_1) \\ p(x_n = l_2) \\ p(x_n = l_3) \\ p(x_n = l_4) \end{bmatrix},$$

the transition probability matrix $\mathbf{A}_{n|n-1} \in \mathbb{R}^{4 \times 4}$, $n = 1, 2, \dots, M$, and $M = 1440/T$ represents the number of data points of each day (one day consists of 1440 minutes). The i th row and j th column entry of $\mathbf{A}_{n|n-1}$ is

$$a_{n|n-1}^{ij} = p(x_n = l_i | x_{n-1} = l_j), i, j = 1, 2, 3, 4.$$

Based on the historical data, we find that the occupancy level is zero during the early morning. Hence, we set the initial probability of the occupancy level $\mathbf{p}(x_0) = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}^T$.

From the training data, we can directly calculate the transition probability:

$$\begin{aligned} a_{n|n-1}^{ij} &= p(x_n = l_i | x_{n-1} = l_j) \\ &= \frac{\sum_{k=1}^m \delta(x_n(k) - l_i) \delta(x_{n-1}(k) - l_j)}{\sum_{k=1}^m \delta(x_{n-1}(k) - l_j)} \end{aligned} \quad (6)$$

where

$$\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & \text{otherwise,} \end{cases}$$

m days occupancy data were used to learn the transition probability, and $x_n(k)$ represents the occupancy level at t_n of the k th day of the training data. In practice, however, both the numerator and the denominator in (6) may be zero. To avoid the zero probability value, we utilize the Laplacian correction technique to smooth the probability in (6), which is commonly used in Naive Bayesian classifier. The smoothed transition probability is given by

$$\begin{aligned} a_{n|n-1}^{ij} &= p(x_n = l_i | x_{n-1} = l_j) \\ &\approx \frac{1 + \sum_{k=1}^m \delta(x_n(k) - l_i) \delta(x_{n-1}(k) - l_j)}{n_L + \sum_{k=1}^m \delta(x_{n-1}(k) - l_j)} \end{aligned} \quad (7)$$

where n_L represents the number of occupancy level, i.e., 4, in our experiment. From (7), we can easily obtain the transition probability matrices $\mathbf{A}_{n|n-1}$, for all $n = 1, 2, \dots, M$.

C. The observation model: estimation of the likelihood

In this section, we firstly formulate an observation model of the number of occupants based on the carbon-dioxide measurements. We then identify the model, and calculate the likelihood based on the results of the observation model.

1) *Observation model of the occupant number:* The indoor carbon-dioxide concentration depends on the occupancy level, the venting level, and the carbon-dioxide concentration in a past time horizon [25]–[28]. The carbon-dioxide dynamic model can be described as a discrete-time state space model:

$$c_n = g(\mathbf{c}_{n-\tau:n-1}, \mathbf{v}_{n-\tau:n}, \mathbf{O}_{n-\tau:n}) \quad (8)$$

where c_n represents the carbon-dioxide concentration at t_n , $g(\cdot)$ is an unknown function, $\mathbf{c}_{n-\tau:n-1}$, $\mathbf{v}_{n-\tau:n}$, and $\mathbf{o}_{n-\tau:n}$ represent the sequence of carbon-dioxide concentration, venting level, and the number of occupants in the corresponding time-horizons, respectively. τ is the length of the time-horizon. We can identify this dynamic model and then solve a deconvolution problem to estimate the occupancy level. However, this method is prone to the errors of the identified model and the truncation errors in the deconvolution process. From (8), we know that a dynamic model in terms of the number of occupants exists. Considering equation (8), we describe the occupancy dynamic model as

$$o_n = f(\mathbf{c}_{n-\tau:n}, \mathbf{v}_{n-\tau:n}, \mathbf{o}_{n-\tau:n-1}) = f(\mathbf{z}_n) \quad (9)$$

where $f(\cdot)$ is the model to be identified. Compared with the occupancy level, the occupant number provides more information. Thus, in the dynamic model we set the occupant number as the state. With this model we can recursively estimate the real-time occupant number from which the occupancy level can be easily obtained.

2) *Ensemble ELM for occupancy dynamic model identification:* The structure of the ensemble ELM for occupancy level estimation is shown in Figure 3, which then will be explained in detail.

Figure 3: The structure of the ensemble extreme learning machine for occupancy level estimation.

Input data: As mentioned before, the input data of the dynamic model is $\mathbf{z}_n = [\mathbf{c}_{n-\tau:n}^T \quad \mathbf{v}_{n-\tau:n}^T \quad \mathbf{o}_{n-\tau:n-1}^T]^T$. In practice, the carbon-dioxide observations suffer from serious spikes due to the measurement noise, irregular occupant movements, and irregular air movements. To reduce the influence of these spikes, we smooth the carbon-dioxide data before using them as the training data. Details about the smoothing of carbon-dioxide data can be found in [23]. In our test-bed, the venting rate of fresh air is a constant. Hence, for the venting level, we only consider the ON/OFF state. If the ventilation system is ON we set $v = 1$ unless we set $v = 0$. We use the real number of occupants in the training process, while in the testing process the estimated occupant number is used.

Feature extraction: Feature extraction is an effective technique for many data-driven approaches. Some statistical features of the input data are more informative for the occupancy estimation [16], [17], [22], [23]. Deep neural networks can automatically generate the features, but it requires a large amount of labelled training data, which is expensive. Based on the discussion in [16], [17], [22], [23] and our own observations from manually tuning, we use the following feature data as the input of the ELM models:

$$\theta_n = \begin{bmatrix} s_1 \mathbf{c}_{n-\tau:n} \\ s_2 (\mathbf{c}_{n-\tau+1:n} - \mathbf{c}_{n-\tau:n-1}) \\ s_3 (\mathbf{c}_{n-\tau+\eta:n} - \mathbf{c}_{n-\tau:n-\eta}) \\ s_4 \mathbf{V}_{n-\tau:n} \\ s_5 \mathbf{O}_{n-\tau:n-1} \end{bmatrix} \quad (10)$$

where $\mathbf{c}_{n-\tau+1:n} - \mathbf{c}_{n-\tau:n-1}$ is the first-order difference of the carbon-dioxide data, $\mathbf{c}_{n-\tau+\eta:n} - \mathbf{c}_{n-\tau:n-\eta}$ is the high-order difference of the carbon-dioxide data, and $\mathbf{s} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 \end{bmatrix}$ represents the scalable parameters. Many publications [16], [17], [22], [23] have shown that the difference of carbon-dioxide data has a close relationship with the variation of occupant numbers.

The venting level only has two states: $v_n = 1$ if $t_n \in [7 : 30\text{AM}, 10 : 00\text{PM}]$ otherwise $v_n = 0$. The common statistical features such as difference and average of the venting data have no obvious influence on the occupancy estimation. In the testing process, the estimated occupant number is used as the input data, which is not the real occupant number. This error will propagate to the next time instant. We find that the statistical features of the estimated occupant number may enlarge the accumulated error. Therefore, we do not extract features from the venting state and the number of occupants.

The scalable parameters can balance the huge difference between different types of data. Well-tuned scalable parameters can improve the estimation performance. For details about the tuning of the scalable parameters, the reader may refer to [23].

Extreme learning machine (ELM): ELM is an extremely fast way to train single-hidden-layer feed-forward neural networks (SLFNs), which is shown in Figure 4. The input-to-hidden layer weights are randomly generated. We only require to find the hidden-to-output layer weights. More details about ELM can be found in [40], [43].

The output function of the ELM for SLFNs is

$$o_n = \sum_{i=1}^L \beta_i h(\mathbf{r}_i \theta_n + \varepsilon_i) = \mathbf{h}(\mathbf{R} \theta_n + \boldsymbol{\varepsilon}) \boldsymbol{\beta} \quad (11)$$

where L is the number of hidden neurons, $\mathbf{r}_i$ consists of the random weights from input layer to the i th hidden neuron and it is the i th row of the random weight matrix $\mathbf{R}$, ε_i is the i th row of $\boldsymbol{\varepsilon}$ and it is a small random bias for the input of i th hidden neuron, $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_L \end{bmatrix}^T$ represents

Figure 4: ELM for single-hidden-layer feed-forward neural network.

the hidden-to-output layer weight vector, $h(\cdot)$ is a prior known activation function, the commonly used activation function is the sigmoid function, and $\mathbf{h}(\mathbf{R}\boldsymbol{\theta}_n + \boldsymbol{\epsilon})$ is a L -dimensional row vector whose i -th element is $h(\mathbf{r}_i\boldsymbol{\theta}_n + \epsilon_i)$.

Substituting all training data into (11), we can obtain a series of linear equations with respect to different time instants from which we can easily obtain the hidden-to-output layer weight β by solving a least square problem. With β , we can use the SLFN to estimate the number of occupants.

Voting: In the ELM model, the input-to-hidden layer weights are randomly generated. Different weights lead to different solutions of the occupancy estimation. To remove the diversity of the solutions and find the best one, we implement an ensemble method. We run a group of ELMs and vote the best solution.

Specifically, we implement two votes. The first one is to vote the occupant number. The result will be recursively used for the occupancy estimation of the next time instant which is shown as the dash line in Figure 3. The voting result of the occupant number is

$$\hat{o}_n = \arg \max_i \sum_{j=0}^{n_e} \delta([\hat{o}_n^j] - i)$$

where $\hat{o}_n$ is the voting result, $o_{\max}$ represents the maximum number of the occupants in the office space, i.e., 28 in our experiment, n_e represents the number of ELMs, and $[\hat{o}_n^j]$ is an integer rounded from $\hat{o}_n^j$ which is the solution of the j th ELM.

The second one is to vote the occupancy level. The result will be used for the likelihood calculation.

The voting result of the occupancy level is

$$\hat{x}_n = \arg \max_{l_i} \sum_{j=0}^{n_e} \delta(x(\hat{o}_n^j) - l_i) \quad (12)$$

where l_i represents the i th occupancy level and $x(\hat{o}_n^j)$ represents the occupancy level of the solution of j th ELM. The definition of occupancy level is given in Table I.

3) *Likelihood calculation*: We view $\hat{x}_n$ as the observation of the occupancy level at the time instant t_n , i.e., y_n , and formulate the observation model as equation (1). Next, we calculate the likelihood $p(y_n|x_n)$.

We denote the likelihood matrix by $\mathbf{B}_n$ for all $n = 1, 2, \dots, M$. Like the method to find the transition probability matrix $\mathbf{A}_{n|n-1}$, based on the training data and the results of (12), we can obtain all entries of $\mathbf{B}_n$: For all $i, j = 1, 2, 3, 4$,

$$\begin{aligned} b_n^{ij} &= p(y_n = l_i | x_n = l_j) \\ &= p(\hat{x}_n = l_i | x_n = l_j) \\ &\approx \frac{1 + \sum_{k=1}^m \delta(\hat{x}_n(k) - l_i) \delta(x_n(k) - l_j)}{n_L + \sum_{k=1}^m \delta(x_n(k) - l_j)} \end{aligned}$$

D. Occupancy estimation via Bayes filter

With the transition probability matrix $\mathbf{A}_{n|n-1}$ and the likelihood matrix $\mathbf{B}_n$, we can directly calculate the posterior *pmf* of the occupancy level at the time instant t_n .

From (3), we can obtain the prior *pmf*:

$$p(x_n = l_i | \mathcal{Y}_{n-1}) = \mathbf{a}_{n|n-1}^i \mathbf{p}(x_{n-1} | \mathcal{Y}_{n-1}) \quad (13)$$

where $\mathbf{a}_{n|n-1}^i$ is the i th row of $\mathbf{A}_{n|n-1}$ and

$$\mathbf{p}(x_{n-1} | \mathcal{Y}_{n-1}) = \begin{bmatrix} p(x_{n-1} = l_1 | \mathcal{Y}_{n-1}) \\ p(x_{n-1} = l_2 | \mathcal{Y}_{n-1}) \\ p(x_{n-1} = l_3 | \mathcal{Y}_{n-1}) \\ p(x_{n-1} = l_4 | \mathcal{Y}_{n-1}) \end{bmatrix}.$$

Substituting (13) into (4), we can obtain the evidence:

$$\begin{aligned} p(y_n | \mathcal{Y}_{n-1}) &= \sum_{i=1}^4 p(y_n | x_n = l_i) p(x_n = l_i | \mathcal{Y}_{n-1}) \\ &= \mathbf{b}_n^{i*} \mathbf{A}_{n|n-1} \mathbf{p}(x_{n-1} | \mathcal{Y}_{n-1}) \end{aligned} \quad (14)$$

where $y_n = \hat{x}_n = l_{i^*}$ and $\mathbf{b}_n^{i*}$ is the i^* th row of $\mathbf{B}_n$.

Substituting (13) and (14) into (2), we can obtain the posterior *pmf* of the occupancy level:

$$\mathbf{p}(x_n | \mathcal{Y}_n) = \frac{(\mathbf{b}_n^{i*})^T * (\mathbf{A}_{n|n-1} \mathbf{p}(x_{n-1} | \mathcal{Y}_{n-1}))}{\mathbf{b}_n^{i*} \mathbf{A}_{n|n-1} \mathbf{p}(x_{n-1} | \mathcal{Y}_{n-1})} \quad (15)$$

where we define a new operation “ $*$ ”: for any vectors $\mathbf{a} = [a_1 \ a_2 \ \cdots \ a_n]^T$ and $\mathbf{b} = [b_1 \ b_2 \ \cdots \ b_n]^T$, $\mathbf{a} * \mathbf{b} = [a_1 b_1 \ a_2 b_2 \ \cdots \ a_n b_n]^T$.

We set the initial conditional *pmf* as $\mathbf{p}(x_0|\mathcal{Y}_0) = \mathbf{p}(x_0) = [1 \ 0 \ 0 \ 0]^T$. With the real-time observation y_n , we can recursively obtain the real-time posterior *pmf* from (15). With the posterior *pmf* $\mathbf{p}(x_n|\mathcal{Y}_n)$, we can easily find the final estimation of the occupancy level:

$$\tilde{x}_n = \mathbf{E}(x_n|\mathcal{Y}_n) = l_i, \text{ where } i = [\mathbf{I}^T \mathbf{p}(x_n|\mathcal{Y}_n)] \quad (16)$$

Here, $\mathbf{E}(\cdot)$ is the expectation operator, $\tilde{x}_n$ is the optimal estimate in the sense of MMSE. $[\mathbf{I}^T \mathbf{p}(x_n|\mathcal{Y}_n)]$ is an integer rounded from $\mathbf{I}^T \mathbf{p}(x_n|\mathcal{Y}_n)$ in which $\mathbf{l} = [1 \ 2 \ 3 \ 4]^T$ implies the occupancy level.

With the posterior *pmf* of the occupancy level, we can also find the ‘optimal’ estimate in other senses. In this work, we only consider the optimal estimate in the sense of MMSE, which is the most commonly used criterion.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Experimental setup

Following the test-bed described in Section II, we design the observation model based on the proposed ensemble ELM detailed in Section III-C2. The observation model contains nine design parameters that need to be set in advance. The tuning results of the parameters are shown in Table II. The entries of the random weight matrix $\mathbf{R}$ in (11) are independent, identically distributed (i.i.d.), and follow the uniform distribution in the interval $[-1, 1]$. Similarly, the entries of ε in (11) are i.i.d., and follow the uniform distribution in the interval $[-0.1, 0.1]$.

Table II: The parameters using in the observation model identification.

τ	η	s_1	s_2	s_3	s_4	s_5	L	n_e
30	10	0.0001	0.0025	0.02	0.004	0.008	1000	100

Notes: τ is the time horizon in (9), η is the order of the high-order difference in the feature vector (10), s_1 to s_5 are the scalable parameters in (10), L is the number of hidden neurons in (11), and n_e represents the number of ELMs used in (12).

To verify the effectiveness of the proposed ensemble method, we compare the proposed observation model with the results of the ELM in [23], which is actually a special case of the proposed ensemble method but only one ELM is used in Figure 3. We then compare both observation models (ELM and ensemble ELM) with the proposed Bayes filter. To validate the generalised performance of the proposed methods, we use leave-one-out-cross-validation (LOOCV) technique. Here, we employ five criteria to

quantify the performance of the estimation results: the estimation accuracy, the normalized root-mean-squared-error (NRMSE), the detection accuracy, the precision of the detection, and the recall of the detection, which were used in [16]. We use the first two criteria to assess the performance on occupancy level estimation. The last three criteria are used to assess the performance on occupancy detection, which is very important in lighting control systems. The precision and recall are commonly used to validate the robustness of binary classification.

Figure 5: The estimated mean occupancy level per 15 minutes in five days. For the result of each day, the other 29-day data are used to estimate the transition probability and to train the observation model.

B. Results and discussions

The HVAC systems in the test-bed make an operation every 15 minutes like the HVAC system in [44]. Hence, we apply the above mentioned methods to estimate the mean occupancy level per 15 minutes. As mentioned in Section II, the sampling time of the carbon-dioxide sensor is 1 minute; therefore, we can use the proposed observation model to estimate the occupancy level per minute. Considering the requirement of the HVAC system, we estimate the mean occupancy level every 15 minutes. For example, at 8:00AM of some day, the mean occupancy estimation is

$$\bar{o}_{8:00} = \frac{1}{15} \sum_{t=7:46}^{8:00} \hat{o}_t$$

Substituting $\bar{o}_{8:00}$ into Table I, we can obtain the estimated occupancy level per 15 minutes.

The estimated occupancy levels in five days are shown in Figure 5. Here, we use the features discussed in Section III-C2 for both one ELM and the ensemble ELM. Hence, one ELM is the same as the feature scaled ELM in [23]. In the Bayes filter, we use the observation model identified from the ensemble ELM.

Figure 5 shows that the solutions of the proposed ensemble ELM are better than those of the ELM, which was used in [23]. Specifically, in the last three days, the ensemble ELM finds the high occupancy level while the ELM does not. Because of the random weighting, the solutions of ELM models are diverse with respect to different input-to-hidden layer weights. However, the solutions of all ELM models are around the optimal solution [40]. In the proposed ensemble ELM method, we find a group of solutions from different ELM models and then select the best one by voting. Therefore, the ensemble method can provide better solutions.

It is obvious that the Bayes filter outperforms the ensemble ELM model. With the help of the prior statistical information, Bayes filter can correct part of the wrong estimation of the observation model, i.e., the ensemble ELM model. Specifically, the proposed Bayes filter improve the solution of the observation model from two perspectives:

- Utilized the prior information, i.e., the statistical model obtained from the empirical occupancy data, the Bayes filter can correct some irregular estimations based on carbon-dioxide observations. For example, in the evening of Day 4 in Figure 5, the estimated occupancy level of the ensemble ELM model jumps from low to medium, which is irregular, and the proposed Bayes filter can correct the medium to be low. In addition, almost every noon, i.e., the lunch time, the occupancy level jumps to be low. Therefore, at the noon of Day 5, the Bayes filter can correct the occupancy level medium to be low.
- Bayes filter can reduce the time-delay of the solutions obtained from the observation model around the midnight. The last-leave of each day was at around 23:30 PM, and the air-conditioning and

ventilation systems have turned off for around one and a half hours. Without the ventilation system, the reduction of carbon-dioxide concentration has a large time-delay; therefore, the solution of the observation model probably has a time-delay. The historical occupancy data provides the prior probability whether the test-bed is occupied around the midnight with which the Bayes filter can enhance the solutions of the observation model and thus reduces the time-delay.

Actually, Bayes filter can also reduce the time-delay in every morning. However, the improvement is insignificant compared with that in every midnight. The first-arrival of each day is at around 8:30 AM and the air-conditioning and ventilation system has worked around one hour. Thus, the carbon-dioxide concentration is very low when the first person arrivals. Once the carbon-dioxide concentration increases, the observation model will immediately detect the occupancy; therefore, the time-delay of the solution obtained from the observation model in every morning is much smaller than that in every midnight, and the time-delay reduction by the Bayes filter is also smaller.

Table III: The performance of the proposed Bayes filter compared with the results directly obtained from observation models.

criteria	ELM [23]	Ensemble ELM	Bayes filter
estimation accuracy (%)	82.67	83.11	85.57
NRMSE	0.1408	0.1377	0.1270
detection accuracy (%)	95.80	95.71	97.69
precision (%)	95.03	95.65	98.42
recall (%)	98.23	97.47	97.77

Notes: Both the input of the ELM and that of the ensemble ELM are the same features. The five criteria refer to the results of leave one out cross validation.

To quantitatively show the effectiveness of the proposed Bayes filter, we calculate the five criteria of the LOOCV results and shows them in Table III. The estimation accuracy and NRMSE show that the ensemble ELM outperforms ELM and that the proposed Bayes filter outperforms the ensemble ELM model. The detection accuracies also indicate that the Bayes filter greatly improve the estimation of the observation model. The precision and recall of the Bayes filter are quite high, which shows the robustness of the proposed method.

C. Comparison with the state-of-the-art

The data of this experiment has also been used in [16] in which data for 31 days were used. The active time of the ventilation system of one of the 31 days is unknown, hence in this paper, we only use the data

Table IV: The performance comparison between the proposed Bayes filter with the state-of-the-art.

criteria	Dong's [17]	Yang's [18]	Masood's [21]	Candanedo's [20]	Jiang's [23]	Chen's [16]	Bayes filter
estimation accuracy (%)	71.46	66.67	72.31	70.21	73.96	75.21	77.29
NRMSE	0.1912	0.2509	0.2322	0.2297	0.1701	0.1660	0.1588
detection accuracy (%)	93.13	90.21	92.38	88.54	92.71	94.37	97.29
precision (%)	91.83	89.65	91.09	86.52	88.99	93.45	95.92
recall (%)	94.46	90.20	93.82	91.28	100	95.29	99.65

of the other 30 days. In [16], the authors did a very good comparison among the state-of-the-art techniques. Following the same style, we compare the proposed Bayes filter with the state-of-the-art approaches. The results are shown in Table IV. Noted that our prior work [23] estimate the occupancy level only using the carbon-dioxide concentration while all other methods in Table IV using the environmental data including carbon-dioxide concentration, temperature, relative humidity and pressure.

In [16], the definition of the occupancy level is different with that in Table I. The authors in [16] considered the balance of occupancy data of each level, i.e., each occupancy level was assigned almost the same amount of training data, and set the occupancy range of low to 1-6 occupants, medium to 7-14 occupants, and high to more than 14 occupants. In addition, [16] used five-day data for the testing. Hence, for an easy comparison, all methods in Table IV use the same five-day data for the testing and the same definition of the occupancy level in [16]. It is clearly shown in Table IV that for both estimation accuracy and the detection accuracy, the proposed Bayes filter has the best performance.

Figure 6: Performance comparison between Chen's online IHMM-MLR [16] and the proposed Bayes filter.

Table IV shows that for all prior works, the inhomogeneous hidden Markov model with multinomial logistic regression (IHMM-MLR) proposed in [16] provides the best solution. We show the details of the solutions of the IHMM-MLR with the proposed Bayes filter in Figure 6. We can find from Figure 6

that Bayes filter outperforms the IHMM-MLR, especially in the areas of the marked circles. For the time slots around midnight of all the five days, the solutions of IHMM-MLR are always not better than the solutions of the proposed Bayes filter. Actually, we can view all state-of-the-art methods in Table IV as observation models and apply the proposed Bayes filter to improve the solutions of the state-of-the-art techniques.

V. CONCLUSIONS AND FUTURE WORKS

In this work, we propose a framework to apply the Bayes filter for indoor occupancy estimation from the observations of carbon dioxide concentrations. In this way, both the statistical information and the observation model of the occupancy data can be fully utilized for the occupancy estimation. The experimental results show that the proposed ensemble ELM outperforms the ELM for identification of the observation model. The proposed Bayes filter can enhance the solution of the observation model by fusing the statistical model and the observation model.

The proposed Bayes filter only requires the solution of the observation model and has no limitation for the forms of the observation model. Hence, we can apply the Bayes filter for all the current methods which can be viewed as different observation models. The performance of the Bayes filter depends on the estimation of the transition probability matrix and the solution of the observation model. The future work can focus on the improvement of the transition probability matrix and the performance of the observation model. To find better transition probability matrix, we need more historical occupancy data. If no additional occupancy data is available, the simulated data using the generative adversarial network may be used to enhance the transition probability matrix. In practice, it is expensive to obtain the real occupancy data. However, the carbon-dioxide data is easily collected. Hence, we can use semi-supervised learning techniques to achieve better observation models.

REFERENCES

- [1] J. Laustsen, *Energy efficiency requirements in building codes, energy efficiency policies for new buildings*, International Energy Agency (IEA), 2008.
- [2] W. Wang, T. Hong, N. Li, R. Q. Wang, and J. Chen. "Linking energy-cyber-physical systems with occupancy prediction and interpretation through WiFi probe-based ensemble classification." *Applied energy*, vol. 236, pp. 55-69, 2019.
- [3] J. Brooks, S. Kumar, S. Goyal, R. Subramany, and P. Barooah, "Energy-efficient control of under-actuated HVAC zones in commercial buildings," *Energy Build.*, vol. 93, pp. 160–168, 2015.
- [4] E. Soltanaghaei and K. Whitehouse, "Practical occupancy detection for programmable and smart thermostats". *Applied Energy*, vol. 220, pp.842-855, 2018.
- [5] Y. Peng, A. Rysanek, Z. Nagy, and A. Schläpfer, "Using machine learning techniques for occupancy-prediction-based cooling control in office buildings". *Applied energy*, vol. 211, pp.1343-1358, 2018.

- [6] F. Oldewurtel, D. Sturzenegger, and M. Morari, "Importance of occupancy information for building climate control," *Appl. Energy*, vol. 101, pp. 521–532, 2013.
- [7] F. Iglesias and P. Palensky, "Profile-based control for central domestic hot water distribution," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 697–705, 2014.
- [8] T. Labeodan, W. Zeiler, G. Boxem, and Y. Zhao, "Occupancy measurement in commercial office buildings for demand-driven control applications: a survey and detection system evaluation," *Energy Build.*, vol. 93, pp. 303–314, 2015.
- [9] H. Zou, Y. Zhou, H. Jiang, S.-C. Chien, L. Xie, and C. J. Spanos, "Winlight: A WIFI-based occupancy-driven lighting control system for smart building," *Energy Build.*, vol. 158, pp. 924–938, 2018.
- [10] Y. Cho, S. O. Lim, and H. S. Yang, "Collaborative occupancy reasoning in visual sensor network for scalable smart video surveillance," *IEEE Trans. Consum. Electron.*, vol. 56, no. 3, 2010.
- [11] Y. Zhao, W. Zeiler, G. Boxem, and T. Labeodan, "Virtual occupancy sensors for real-time occupancy information in buildings," *Build. Environ.*, vol. 93, pp. 9–20, 2015.
- [12] N. Li, G. Calis, and B. Becerik-Gerber, "Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations," *Autom. constr.*, vol. 24, pp. 89–99, 2012.
- [13] P. Liu, S.-K. Nguang, and A. Partridge, "Occupancy inference using pyroelectric infrared sensors through hidden Markov models," *IEEE Sensors J.*, vol. 16, no. 4, pp. 1062–1068, 2016.
- [14] W. Kleiminger, C. Beckel, T. Staake, and S. Santini, "Occupancy detection from electricity consumption data," in *Proc. 5th ACM Workshop on Embedded Syst. Energy-Efficient Build.*, 2013, pp. 1–8.
- [15] L. Zimmermann, R. Weigel, and G. Fischer, "Fusion of non-intrusive environmental sensors for occupancy detection in smart homes," *IEEE Internet Things J.*, 2017.
- [16] Z. Chen, Q. Zhu, M. Masood, and Y. C. Soh, "Environmental sensors based occupancy estimation in buildings via IHMM-MLR," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2184–2193, 2017.
- [17] B. Dong, B. Andrews, K. P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, and D. Benitez, "An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network," *Energy Build.*, vol. 42, no. 7, pp. 1038–1046, 2010.
- [18] Z. Yang, N. Li, B. Becerik-Gerber, and M. Orosz, "A systematic approach to occupancy modeling in ambient sensor-rich buildings," *Simulation*, vol. 90, no. 8, pp. 960–977, 2014.
- [19] A. Szczurek, M. Maciejewska, and T. Pietrucha, "Occupancy determination based on time series of CO₂ concentration, temperature and relative humidity," *Energy Build.*, vol. 147, pp. 142–154, 2017.
- [20] L. M. Candanedo and V. Feldheim, "Accurate occupancy detection of an office room from light, temperature, humidity and CO₂ measurements using statistical learning models," *Energy Build.*, vol. 112, pp. 28–39, 2016.
- [21] M. K. Masood, Y. C. Soh, and V. W.-C. Chang, "Real-time occupancy estimation using environmental parameters," in *Int. Joint Conf. Neural Netw.*, 2015, pp. 1–8.
- [22] M. Masood, Y. C. Soh, and C. Jiang, "Occupancy estimation from environmental parameters using wrapper and hybrid feature selection," *Appl. Soft Comput.*, vol. 60, pp. 482–494, 2017.
- [23] C. Jiang, M. K. Masood, Y. C. Soh, and H. Li, "Indoor occupancy estimation from carbon dioxide concentration," *Energy Build.*, vol. 131, pp. 132–141, 2016.
- [24] S. Wolf, D. Cali, J. Krogstie, and H. Madsen, "Carbon dioxide-based occupancy estimation using stochastic differential equations," *Applied energy*, vol. 236, pp. 32–41, 2019.
- [25] S. Kar and P. K. Varshney, "Accurate estimation of gaseous strength using transient data," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 4, pp. 1197–1205, 2011.

- [26] D. Calì, P. Matthes, K. Huchtemann, R. Streblow, and D. Müller, "CO₂ based occupancy detection algorithm: Experimental analysis and validation for office and residential buildings," *Build. Environ.*, vol. 86, pp. 39–49, 2015.
- [27] G. Ansanay-Alex, "Estimating occupancy using indoor carbon dioxide concentrations only in an office building: a method and qualitative assessment," in *Proc. 11th REHVA World Congr. Energy Efficient Smart Healthy Build.*, 2013, pp. 1–8.
- [28] M. Zuraimi, A. Pantazaras, K. Chaturvedi, J. Yang, K. Tham, and S. Lee, "Predicting occupancy counts using physical and statistical CO₂-based modeling methodologies," *Build. Environ.*, vol. 123, pp. 517–528, 2017.
- [29] F. Wang, Q. Feng, Z. Chen, Q. Zhao, Z. Cheng, J. Zou, Y. Zhang, J. Mai, Y. Li, and H. Reeve, "Predictive control of indoor environment using occupant number detected by video data and CO₂ concentration," *Energy Build.*, vol. 145, pp. 155–162, 2017.
- [30] J. A. D'Áz and M. J. Jimenez, "Experimental assessment of room occupancy patterns in an office building. Comparison of different approaches based on CO₂ concentrations and computer power consumption," *Applied energy*, vol. 199, pp. 121–141, 2017.
- [31] M. Amayri, A. Arora, S. Ploix, S. Bandhyopadhyay, Q.-D. Ngo, and V. R. Badarla, "Estimating occupancy in heterogeneous sensor environment," *Energy Build.*, vol. 129, pp. 46–58, 2016.
- [32] Z. Chen, C. Jiang, and L. Xie, "Building Occupancy Estimation and Detection: A Review," *Energy Build.*, vol. 169, pp. 260–270, 2018.
- [33] C. Liao, Y. Lin, and P. Barooah, "Agent-based and graphical modelling of building occupancy," *J. Build. Perf. Simul.*, vol. 5, no. 1, pp. 5–25, 2012.
- [34] Z. Chen, J. Xu, and Y. C. Soh, "Modeling regular occupancy in commercial buildings using stochastic models," *Energy Build.*, vol. 103, pp. 216–223, 2015.
- [35] J. Page, D. Robinson, N. Morel, and J.-L. Scartezini, "A generalised stochastic model for the simulation of occupant presence," *Energy Build.*, vol. 40, no. 2, pp. 83–98, 2008.
- [36] Z. Chen and C. Jiang, "Building occupancy modeling using generative adversarial network," *Energy Build.*, vol. 174, pp. 372–379, 2018.
- [37] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. signal process.*, vol. 50, no. 2, pp. 174–188, 2002.
- [38] Z. Chen, *Bayesian filtering: from kalman filters to particle filters, and beyond*, Technical Report, McMaster University, Hamilton, 2003.
- [39] C. Wang, D. Yan, and Y. Jiang, "A novel approach for building occupancy simulation," *Build. Simul.*, vol. 4, no. 2, pp. 149–167, 2011.
- [40] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.
- [41] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 41–53, 2016.
- [42] F. C. Sangoboye and M. B. Kjærgaard, "Plcount: A probabilistic fusion algorithm for accurately estimating occupancy from 3D camera counts," in *Proc. 3rd Int. Conf. Syst. Energy-Efficient Built Environ.*, 2016, pp. 147–156.
- [43] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst. Man Cybern. Part B-Cybern.*, vol. 42, no. 2, pp. 513–529, 2012.
- [44] E. Png, S. Srinivasan, K. Bekiroglu, C. Jiang, R. Su, and K. Poolla, "An internet of things upgrade for smart and scalable heating, ventilation and air-conditioning control in commercial buildings," *Applied Energy*, vol. 239, pp. 408–424, 2019.

Conflict of interest statement

Manuscript title: Bayesian filtering for building occupancy estimation from carbon dioxide concentration.

Authors:

Chaoyang Jiang, Beijing Institute of Technology, China. Email: cjiang@bit.edu.cn

Zhenghua Chen, Agency for Science, Technology and Research (A*STAR), Singapore, Email:

chen0832@e.ntu.edu.sg.

Rong Su, Nanyang Technological University, Singapore. Email: rsu@ntu.edu.sg.

Yeng Chai Soh, Nanyang Technological University, Singapore. Email: eycsch@ntu.edu.sg

Mustafa Khalid Masood, Aalto University, Finland. Email: mkm_inn@hotmail.com

The authors have no conflicts of interest to declare.

The authors confirm that this manuscript has not been published elsewhere and is not under consideration by another journal. The authors have approved the manuscript and agree with submission to Energy and Buildings.

The authors declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in the above manuscript.