Building and Environment 123 (2017) 517–528

Contents lists available at ScienceDirect

Building and Environment

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

Predicting occupancy counts using physical and statistical Co_2-based modeling methodologies

M.S. Zuraimi*, A. Pantazaras, K.A. Chaturvedi, J.J. Yang, K.W. Tham, S.E. Lee

Building Energy Efficiency (BEE) Hub, Department of Building, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 23 June 2017
Received in revised form
17 July 2017
Accepted 17 July 2017
Available online 20 July 2017

Keywords:
Carbon dioxide
Occupant counting
Large occupant numbers
Physical model
Statistical model

ABSTRACT

Energy consumption and indoor environment quality (IEQ) of buildings have been linked to human occupants. Predicting the number of occupants in a space is essential for the effective management of various building operation functions as well as improve energy efficiency. This study is the first to compare the performance of physical and statistical models in predicting occupant counts in a high volume lecture theatre (Occ = 200) using CO_2 sensors. CO_2 measurements and actual occupant numbers were obtained for 4 months to provide robust data comparison of the methodologies. It was found that that the dynamic physical models and Support Vector Machines (SVM) and Artificial Neural Networks (ANN) models utilizing a combination of average and first order differential CO_2 concentrations performed the best in terms of predicting occupancy counts with the ANN and SVM models showing higher predictive performance. RMSE values for the corresponding models were 12.8, 12.6 and 12.1 respectively and correlation coefficients were all greater than 0.95. The relatively good agreement between dynamic physical model predictions and ground truth shows that the dynamic mass balanced model is adequate for predicting occupancy counts provided that the air exchange rates measured are accurate. Model average accuracies across all tolerance was between 70 and 76% demonstrating good performance for a large number of occupants. A discussion on the merits and limitations of each model types was presented to provide guidance on the adoption of various models.

1. Introduction

Rising energy costs and consumption in recent years, especially in buildings, have led researchers to consider new methods and approaches for reducing energy use. The building sector accounts for about 40% of total energy consumption and over 30% of the greenhouse gas (GHG) emissions [41]. The increasing trend towards building consumption is expected to persist in the coming years [40]. As a result, proposed solutions to reduce building energy consumption and GHG emissions has been provided in the form of implementation of low-and zero-carbon buildings, smarter and more efficient building energy systems as well as behavioral change among building occupants [40]. Human occupants is a key element in the relationship of buildings energy consumption and indoor environment quality (IEQ). Occupant presence release latent and sensible heat that changes thermal conditions warranting increased in air-conditioning. They also release pollutants such as carbon dioxide (CO_2) which is also an indicator for ventilation adequacy. Through human occupant behavior such as window operation, lighting, blind and fan operation, thermostat changes and use of plug-in appliances, indoor loads will be affected and subsequently impact building energy consumption [40]. Research on human occupants behavior modeling and energy consumption have been performed in various building systems such as ventilation, air-conditioning and lighting. Findings have shown that up to 30% cooling energy savings can be achieved with the aid of occupancy sensors to control lighting [27]. In another study, a 30% cooling energy reduction was achieved by using occupancy control in an open office by utilizing thermostat setting of the unoccupied zone higher than that of the occupied zones [21]. In summary, occupants and their behavior affect building IEQ and energy consumption significantly and is a leading source of uncertainty to predict building energy use [40].

Gathering real time data on IEQ has been identified as one of the core strategies to reduce the uncertainty of predicting energy use [40,41]; as well as providing good control systems to changes in IEQ [17]. Standardized occupancy schedules used as characterization

* Corresponding author.
E-mail address: bdgzm@nus.edu.sg (M.S. Zuraimi).

http://dx.doi.org/10.1016/j.buildenv.2017.07.027
0360-1323/

tools are not able to provide a good level of resolution because they provide only generic and average patterns. In recent years, a myriad of small and inexpensive IEQ sensor technologies have been made available on the market with varying degrees of sophistication and functionality. Many studies have reported the use of CO_2 sensors to detect and count occupant numbers [13,18,33,25,43,44]. Indeed, occupancy counting through CO_2 sensors is a promising approach as indoor CO_2 concentrations are indicative of the presence of humans. Also, CO_2 sensors are cheap, small, non-intrusive and non-terminal-based. Since they are also conventionally available with standard HVAC systems, no additional investment to existing infrastructures are required.

To model occupancy counts using CO_2 sensors include the use of physical modeling as well as statistical modeling techniques. For physical models, research on dynamic occupancy counting has been performed for a single zone utilizing a mass balanced model [8,6,12,19,35,43,44]. The main factors to achieve a well performing dynamic physical model are the assumptions of a well-mixed indoor space, correct use of CO_2 generation rates and accurate measurements of CO_2 concentrations at select measurement points as well as ventilation rate [1,2]. Other studies have used the steady state version of the mass balanced model only to observe significant delay compared with the true occupancy counts despite the algorithm being stable [12,44]. For statistical models, various CO_2 information is used for occupancy prediction based on different standard or advanced machine learning techniques. For example [13], explored standard Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Hidden Markov Models (HMM) to estimate occupancy numbers in a small office using several CO_2 parameters that includes various order differences, outdoor difference and moving averages [33]. developed an occupancy prediction model at the future state based on CO_2 parameters such as indoor concentration, its rate of change and moving averages as well as indoor-outdoor concentration difference and ratios, using HMM algorithm to detect occupancy in a controlled test-room. Jiang et al. [18] developed an occupancy estimator using Feature Scaled Extreme Learning Machine (FS-ELM) algorithm based on CO_2-measurements for an office room occupied with up to 35 persons. They reported an accuracy value up to 94% with a tolerance of four occupants.

Despite the widespread use of CO_2 sensor to estimate occupant count, methodologies using physical and statistical models have mostly been published in isolation. Studies using CO_2 based statistical models have dealt with low number of occupants and has not been shown to be applicable for rooms with more than 35 occupants. In institutional buildings where there are large numbers of occupants and energy performance are more sensitive to occupancy patterns [15,42], there is a need to evaluate the performance of various models. In addition, many studies have a provided a cursory evaluation of the performance without using appropriate criteria or standard [22]. This paper presents the findings of a study to predict and compare occupancy counts using CO_2 -based physical and statistical approaches in a large lecture theatre capable of accommodating a large number of people.

2. Materials and methods

2.1. Test environment and equipment

The test environment used in this study is a lecture theatre (LT) of volume 876 m^3 (Fig. 1). The LT is served by a constant air volume (CAV) with a VSD (Variable Speed Drive) providing air flow modulation air conditioning system (AC). The conditioned air is delivered via nine square supply diffusers mounted on the suspended ceiling while the room's air is extracted to the ceiling plenum via six square extract grilles. Every day the air conditioning system is turned on at 7:00 a.m. and turned off at 10:00 p.m. The amount of outdoor air brought into the LT is regulated by an indoor temperature sensor. The LT is able to accommodate up to 200 people.

A single pan-tilt-zoom (PTZ) camera (Model-VIVOTEK SD8364E, New Taipei City, Taiwan) monitors a wide area of the LT and captures high-resolution zone images aided by the optical zoom. The images taken from the camera were used to physically count the actual number of occupants in the LT as the standard against which all occupancy estimation methods would be tested. The PTZ camera is placed on the front wall next to the lecturer podium and is able to capture occupant information by dividing the space into 6 zones, with the algorithm described in Ref. [9]. High resolution zone images are captured in the theatre at 5 min intervals and the aggregated results are tallied as total captured occupants of the 6 zones, hereby described as ground truth.

Three CO_2 sensors (Blugraph), based on non-dispersive infrared principle have been placed at different locations in the LT; one at the front of the room, one on the wall at the height of the middle row and one at the final row. The placement of the sensors at three different positions was done to mitigate possible uneven mixing effects due to different student sitting distribution and/or air movement within the room. One CO_2 sensor was positioned outside the LT to monitor the ambient levels. The CO_2 sensors have an accuracy of 30 parts per million (ppm) $\pm$ 3% of the reading. The sensors are connected via the wireless university network to a server and transmit their measurements at 1 min intervals. These measurements are updated and saved in the server so they can be viewed either as historical or current data and can be downloaded for further processing. The CO_2 measurements were grouped in 5 min averages, both to compare with the other occupancy counting methods and to decrease the effect of momentary spikes in CO_2 caused by occupants breathing directly on the sensors.

Using the measured CO_2 as a tracer gas, air exchange rates (AER) after each lecture session in the LT were analyzed using a tracer gas decay technique [2]. AER was computed using the TGD software [30]; LESO-PB, EPFL, Lausanne, Switzerland) based on a first order decay. This software has been validated in the very same lecture theatre [30]. To compare the AER utilized with CO_2 as a tracer gas, a reference dynamic model using multiple SF_6 decay tests at different fan speed settings in the LT was developed. The multiple tests were performed to calculate the AER at every setting that were to occur during the four months. Firstly, pure SF_6 tracer gas was released into the LT and its concentration decay at six distributed points within the LT was measured using a photoacoustic spectrophotometer (Innova 1312 Photoacoustic Multi-gas Monitor). Simultaneously, 5 min intervals fan energy consumption data (in kWh) was obtained from the AC system. Then, total supply air into the LT was measured using a 6710 Kanomax TAB master Air Capture Hood. By correlating the measured supply air with the fan energy consumption data, a relationship was derived between the two by regression analysis. From the AER measurements, it was determined that the supplied fresh air is a relatively stable percentage of total supply air (between 14 and 16%) at different fan speeds. Thus, through the relationship between total supply air and fan energy consumption, a very good approximation of the ventilation rate can be calculated, which was then used to retrieve the past AER based on historical fan energy data. CO_2 measurements and ground truth numbers were obtained for 4 months. The result was a set of 6189 data points that were used in this study. The sensor and camera setup in the LT is indicated in Fig. 1.

2.2. Physical models

Physical models are based on a fully mixed dynamic mass

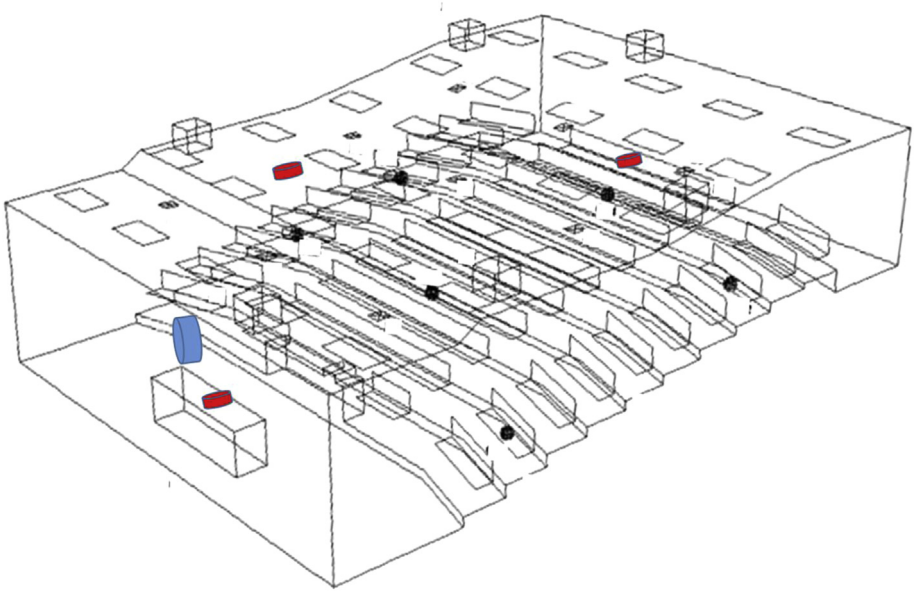

Fig. 1. Location of PTZ camera (blue cylinder) and CO_2 sensors (red cylinders) in the lecture theatre. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

balance model given as:

$$V.d[CO_2]_{in}/dt = \lambda v[CO_2]_{out} + R - \lambda v[CO_2]_{in} \quad (1)$$

where $[CO_2]_{in}$ is the indoor CO_2 concentration ($ppm/10^6$) in the LT; $[CO_2]_{out}$ is the outdoor CO_2 concentration ($ppm/10^6$); λ_v is the AER (h^{-1}) of the LT; R is the rate CO_2 is being generated within the LT ($Lmin^{-1}$) and V is the volume of the LT (m^3). CO_2 concentrations throughout the LT are assumed equal and will be confirmed using measurements at multiple locations.

Solving equation (1), the following relation to obtain R is obtained:

$$R = (\lambda_v V)\{[CO_2]_{in}(t) - [CO_2]_{in}\exp(-\lambda_v t) - [CO_2]_{out}(1 - \exp(-\lambda_v t))\}/(1 - \exp(-\lambda_v vt)) \quad (2)$$

where $[CO_2]_{in}(t)$ is the indoor CO_2 concentration ($ppm/10^6$) at time (t).

The dynamic mass balance model can be simplified by using the steady state assumption [1,2] to give the following relation:

$$R = \lambda_v V([CO_2]_{in} - [CO2]_{out}) = Q_v([CO_2]\ in - [CO_2]_{out}) \quad (3)$$

Where Q_v is the outdoor air flow rate ($Lmin^{-1}$) of the LT.

The occupancy counts in the LT Occ_p is calculated as R/G_P, where G_P = CO_2 generation rate per person ($Lmin^{-1}$) which is assumed to be constant. G_P will vary depending on a number of factors such as age, weight, gender, activity level, and the proportion of consumed macronutrients [1,2,12,26,28,34]. For this study, G_p is obtained from a field environmental chamber experiment using 16 subjects from Singapore under the same environmental conditions of the LT [39]. The value used for this study was 0.21 L min^{-1}.

Researchers have reported that mass balanced model utilizing steady state approach may cause a time delay compared to the ground truth [12,44]. The basic mass balance approach using equation (3) was then modified to include lagged CO_2 concentration to consider the time delay. This accounted for the time it takes to achieve near steady-state conditions in the LT for a CO_2 measurement at a specific instant in time. The intent of including this model is to provide a simplified steady state occupancy count without time delay problems for building operator to use.

2.3. Statistical models

For statistical models, we consider three standard machine learning techniques: 1) Artificial Neural Networks (ANN); 2) Prediction Error Minimization (PEM); and 3) Support Vector Machines (SVM) for predicting occupancy counts. Although many machine learning techniques are available, these three have been chosen because of their widespread use and ease of application.

We use the CO_2 information in various parameter forms to assess their predictions on occupancy data. The CO_2 parameters include average CO_2 concentration (AVG), its 15 min moving average (MA) and average CO_2 concentration in addition to the first order difference (AVG + DIF). For analyses, weekends and public holidays have been removed as during these periods the occupancy dynamics are different and the LT under study remained unoccupied. For consistency and comparability of results across the different models, we have adopted a 60% of the data for training and 40% for validation. In total, the number of data samples used for validation and comparison was 2477.

2.3.1. Artificial neural network (ANN)

Artificial neural network is based on mimicking the function of the human brain [23]. Currently, the applications of ANN are well acknowledged in a number of disciplines, particularly for non-linear modeling. The basic element in ANN is the artificial neuron that are aligned in layers and connected to neurons in other layers through links. These links are known as the synaptic weights and one of the objectives of the training process is to derive these weights. The activation of a neuron is determined by the summation of the weighted inputs and can be mathematically denoted as in Eq. (4):

$$O = f\left(\sum w_{ij}x_j\right) \quad (4)$$

Here, O is the output of the neuron, x_j is the input to that neuron, w_{ij} is the weight of the connection of the input to the neuron and *f* is the transfer function. The transfer function used in neural networks is the Sigmoidal function which has the following form (5):

$$S(t) = \frac{1}{1 + e^{-t}} \quad (5)$$

The output of a neuron is passed to the neuron in the next layer through these weights. The output of this neuron in the next layer can be represented in a very simple form as in Eq. (6), where, O_l is the output of the neuron in the next layer and w_{jl} is the weight connecting the previous neuron to the neuron in this next layer:

$$O_l = f\left(\sum\left(w_{j1} f\left(\sum w_{ij} x_i\right)\right)\right) \quad (6)$$

The learning process in an ANN involves minimizing the squared error between the predicted and the measured outputs. Here, error is minimized by the gradient descent method which involves computing the partial derivative of the error with respect to each weight in the network. We adopted the back propagation algorithm discussed in detailed here [32]:

$$E = \sum \frac{1}{2}(O_p - O_m)^2 \quad (7)$$

As a typical feedforward ANN, the network used in this study consists of an input layer, a hidden layer and an output layer (Fig. 2). The number of inputs corresponds to the number of neurons in the input layer. In the same way, the output corresponds to the neuron in the output layer. We fixed the number of neurons in the hidden layer to 10 as it generates accurate results without consuming much time.

2.3.2. Prediction error minimization (PEM)

The prediction error method as utilized by MATLAB is, in essence, a combination of a subspace identification method and a standard prediction error minimization method. The system to be identified is assumed in state space form:

$$a(t+1) = Aa(t) + Bu(t) + v_1(t) \quad (8)$$

$$b(t) = Da(t) + Eu(t) + v_2(t) \quad (9)$$

where a the state vector, u the input vector, b the output vector and A, B, D and E are matrices with parameters that need to be estimated. $v_1(t)$ and $v_2(t)$ are independent white noise terms. Initial estimation of the system order (i.e. the order of the matrices) and the matrix parameters is done through the N4SID algorithm, which is a singular value decomposition of a weighted matrix. A full description of this method can be found in Ref. [37]. The initial estimations are then updated through an iterative minimization of a cost function, which is usually the sum of the squared prediction error at each point.

The model parameters are estimated by minimizing the error between the model output and the measured data. This error, called the loss function, is a function of prediction errors $e_s(t)$. Generally, this function is a weighted sum of squares of the errors. For a model with ny-outputs, the loss function $V(\theta)$ has the following form:

$$V(\theta) = \frac{1}{N}\sum_{t=1}^{N} e_s^T(t,\theta) W(\theta) e_s(t,\theta) \quad (10)$$

Where N is the number of data samples and $e(t,\theta)$ is ny-by-1 error vector at a given time t, parameterized by the parameter vector θ. $W(\theta)$ is a weighting matrix. The model parameters are estimated by minimizing $V(\theta)$ with respect to θ.

2.3.3. Support vector machine (SVM)

Support Vector Machines have been widely applied in classification, forecasting and regression of random data sets based on Vapnik-Chervonenkis theory [38]. They have been popular due to their capability in modeling non-linear time series. SVMs are based on the structural risk minimization (SRM) inductive principle which seeks to minimize an upper bound of the generalization error consisting of the sum of the training error and a confidence level. This is different from the commonly used empirical risk minimization (ERM) principle which only minimizes the training error. Based on such an induction principle, SVMs have been reported to achieve higher generalization performance than traditional neural networks that implement the ERM principle in solving many data mining problems.

The principle of SVMs is to map non-linear functions in a low dimensional space to a higher dimensional space through the use of a kernel function. All necessary computations can then be performed directly in the high dimensional feature space, and a linear function is trained in such a space without having to compute the map. Some popular kernel functions as described by Ref. [10] include linear, polynomial, radial basis function (RBF) and sigmoid. We used an SVM with a radial basis function (RBF) kernel

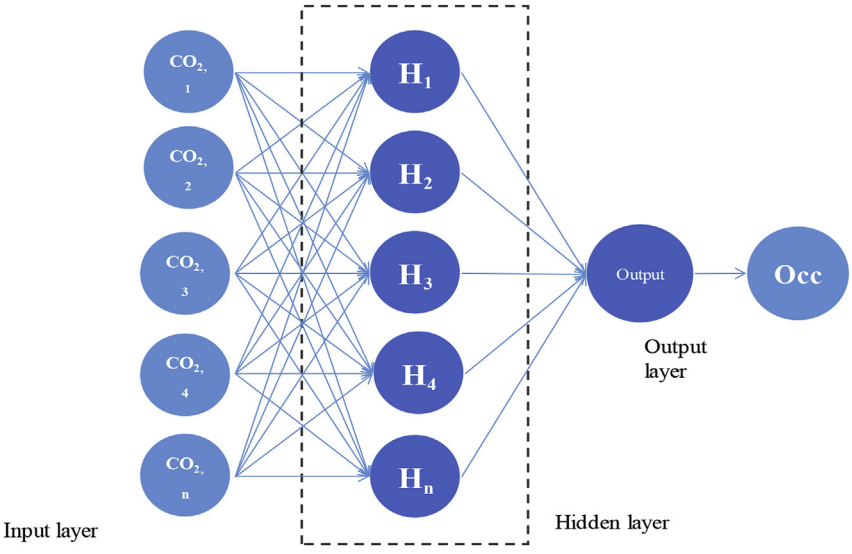

Fig. 2. Structure for feed forward multilayer perceptron ANN.

application on the data set with a penalty C ranging from 2^{-10} to 2^{10} (Equation (11)).

$$K(x_i, x_j) = \exp\left(-\gamma \parallel x_i - x_j\parallel^2\right), \quad \gamma > 0 \tag{11}$$

where γ is the kernel parameter.

RBF kernel parameter "ϒ" and penalty parameter of error term, "C" are two parameters that can affect the performance of the SVM model. In order to ensure that the best values of (C, ϒ) are chosen, we conducted cross-validation and grid search. In this method, the training set is divided into ν subset of equal sizes and each subset is tested using the classifier trained on the remaining ν - 1 sets. Here, we used a ν = 10. Various values for (C, ϒ) are chosen and the cross validation method is tried on each of them. We chose the (C, ϒ) values for the best cross-validation accuracy. The LibSVM toolkit developed by Ref. [10] is used to train and test the data sets. For instance, for the SVM model using AVG + DIF parameters, the results for the cross-validated grid search are as shown below (Fig. 3). The darker areas indicate better performing values for (C, ϒ) and in this case the best performance of the SVM model is with C = 4 and ϒ = 0.

The ASTM guidelines recommend that the r to be greater than 0.9, the intercept be within 25% of the average and the slope be between 0.75 and 1.25.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} Occ_a(i) - Occ_p\ (i)}{n}} \tag{12}$$

$$r = \frac{\sum\left[\left(Occ_a - \overline{Occ_a}\right)\cdot\left(Occ_p - \overline{Occ_p}\right)\right]}{\sqrt{\left[\sum\left(Occ_p - \overline{Occ_p}\right)\right]\cdot\left[\sum\left(Occ_a - \overline{Occ_a}\right)\right]}} \tag{13}$$

Next, we compute the normalized bias (F_b) to determine possible bias in the models (Eqn (14)). F_b ranges from −2 to 2 with a value of 0 indicating perfect agreement. ASTM recommends that the absolute value of F_b for an adequate model be less than or equal to 0.25.

$$F_b = \frac{2\cdot\left(\overline{Occ_p} - \overline{Occ_a}\right)}{\left(\overline{Occ_p} + \overline{Occ_a}\right)} \tag{14}$$

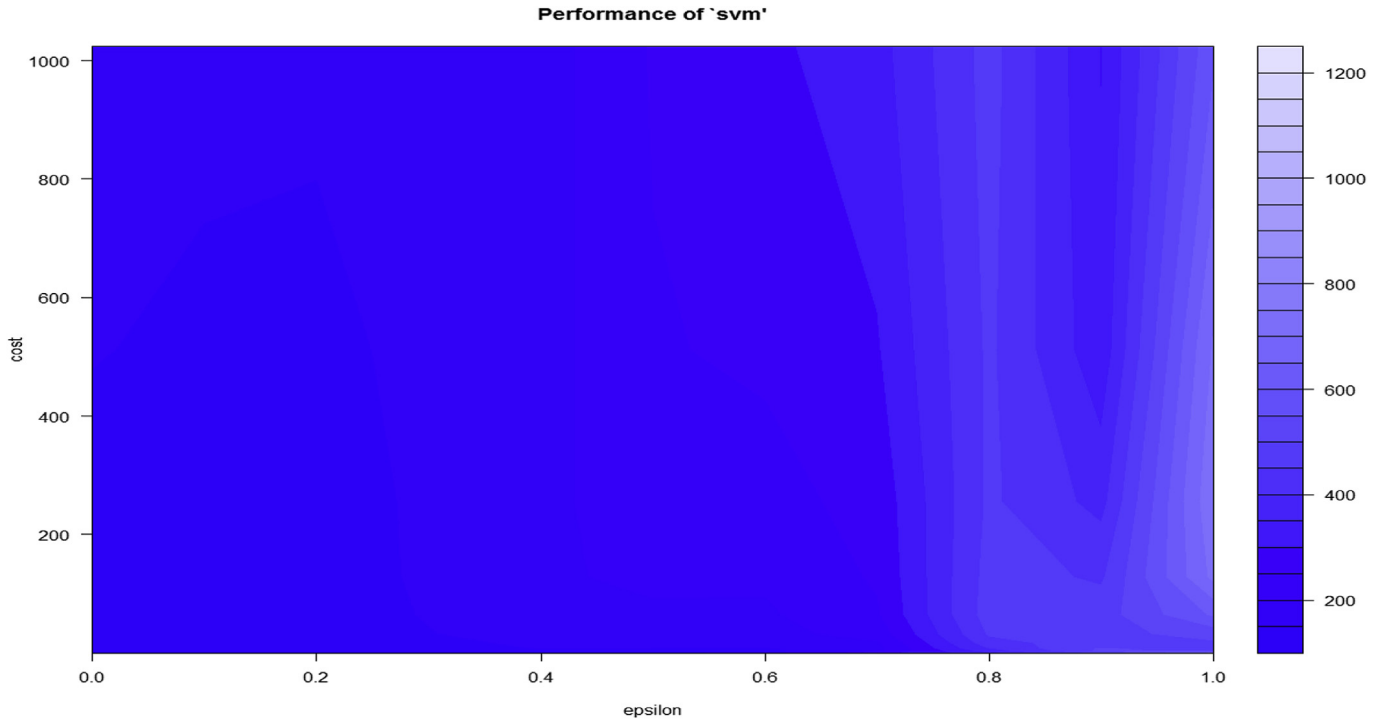

Fig. 3. Cross validation and grid search results for SVM training data.

2.4. Model performance

The performance of the prediction models is determined by evaluating the difference between the predicted and the actual occupancy counts using several metrics. The root mean square error (RMSE) is first calculated by measuring the difference between the predicted (Occ_p) and actual values (Occ_a) involving N number of data points (Eqn (12)). We use ASTM guideline [3] to calculate the correlation coefficient, r between Occ_p and Occ_a as given by Eqn (13). The least squares best fit regression line, provides the slope and intercept information on how well the model fits the data. The ideal line has a slope of 1, an intercept of 0, and a regression r^2 of 1.

In addition to the ASTM guidelines, we also compute the accuracies from the models by determining the prevalence of correctly predicting occupancy counts. We followed the procedure by Ref. [18] to report the prevalence of accuracy (Acc) that the predicted occupancy counts are similar to the ground truth taking into account various tolerances, α (Equations (15) and (16)).

$$Acc\ (Occ, \alpha) = \frac{\sum X(|Occ_a - Occ_p|, \alpha)}{N} \tag{15}$$

where

$$X(|Occ_a - Occ_p|, \alpha) = \begin{cases} 1, & if\ |Occ_a - Occ_p| \le \alpha \\ 0, & otherwise \end{cases} \quad (16)$$

Finally, we compute the models' false positives and false negatives. False positive (F_p) represents the prevalence that the LT is predicted to be occupied while the ground truth state is unoccupied. This could potentially lead to unnecessary AC energy use to maintain the LT's indoor environment. False negative (F_n) represents the prevalence that the LT is predicted to be unoccupied while the ground truth state is occupied. This could potentially lead to a reduction of AC system function which can cause occupant discomfort.

3. Results and discussions

3.1. Qualitative comparison

The model predictions for occupancy counts were visually inspected first by comparing them to the ground truth. Figs. 4 and 5 illustrate a weekly profile for predicted occupancy counts using various models plotted against ground truth occupancy counts. The ground truth occupancy count is represented by the orange line while the blue line represents the predicted occupancy count. Fig. 4A and D represent profiles of physical modeling while Fig. 5A and F represent profiles of statistical modeling. Fig. 4A represents the profiles using dynamic physical modeling via AER information obtained using SF_6 as a tracer gas and fan energy consumption. Fig. 4B represents the profiles using dynamic physical modeling via AER information obtained using CO_2 as a tracer gas. Fig. 4C represents the profiles using steady state physical modeling via AER information obtained using CO_2 as a tracer gas. Fig. 4D is similar to C except the data incorporate lagged CO_2 concentration to consider corresponding time delay. The steady state condition was estimated to be achieved within 45 min.

Fig. 4A shows that the model slightly overpredicted occupancy counts for some lecture sessions and was close to the ground truth in others. Within each lecture session, predicted occupancy counts traces the ground truth relatively well with no time lag when ground truth occupancy is changing. However, there is noticeable noise in the predicted occupancy counts and in non-occupied periods where one would expect zero values. Using CO_2 as a tracer gas at the end of each session, the models tend to underpredict occupancy counts (Fig. 4B). This is because the measured AERs at the end of each session were lower than during the session where temperature sensors would detect higher temperature caused by the presence of humans. As a result, there is a time lag especially when the ground truth occupancy is increasing causing a delay to reach steady state values. Compared to Fig. 4B, the predicted counts in Fig. 4A are more unstable considering the oscillations associated with dynamic AER values and also from the AC control algorithm affected by the sources of heat loads which include among others, outdoor air temperature. As can be seen in Fig. 4C and D, the

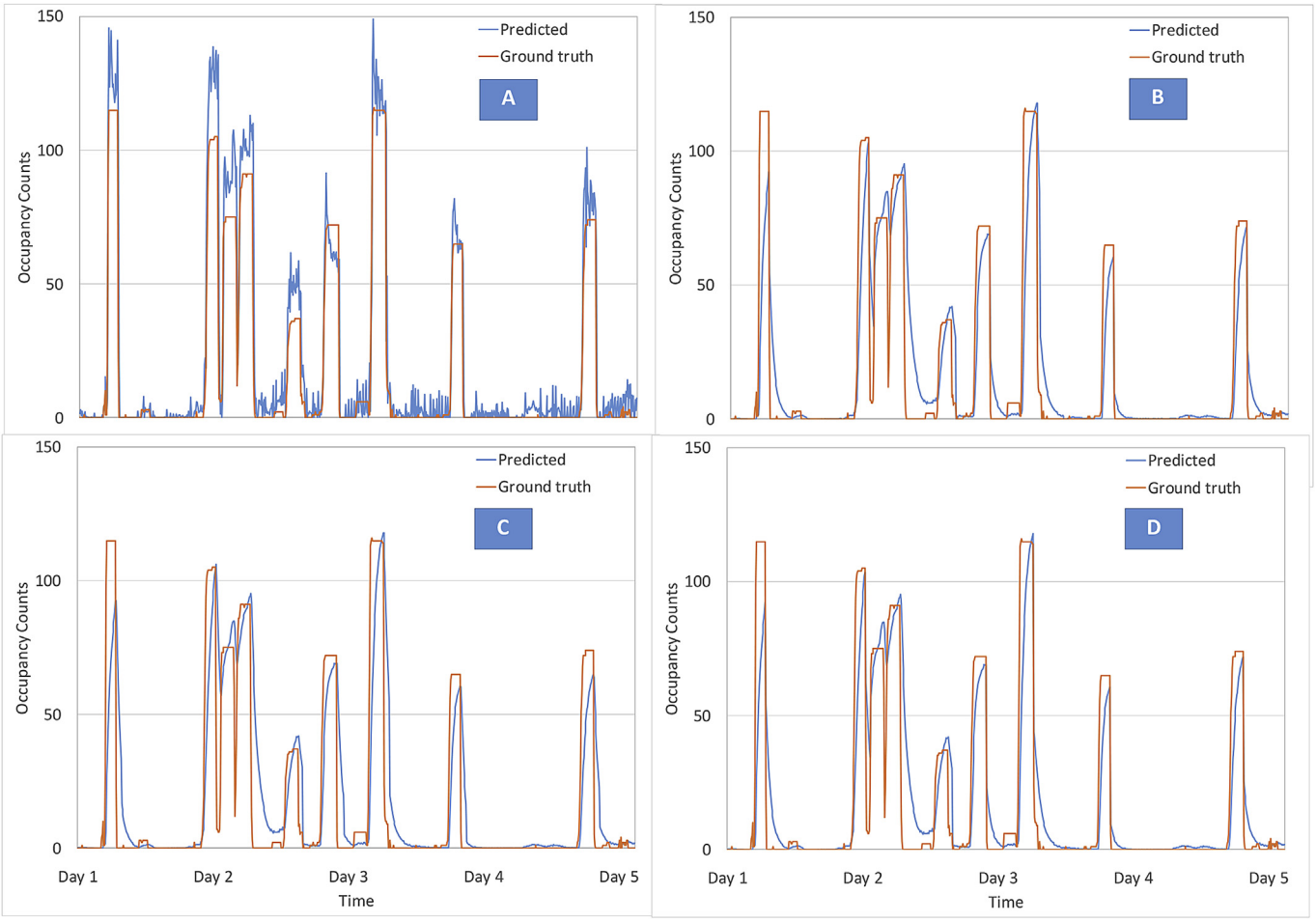

Fig. 4. Profiles of predicted against ground truth occupancy counts for various physical models (A: dynamic SF_6; B: dynamic CO_2; C: Steady state CO_2; D: Steady state CO_2 lagged) in a typical week.

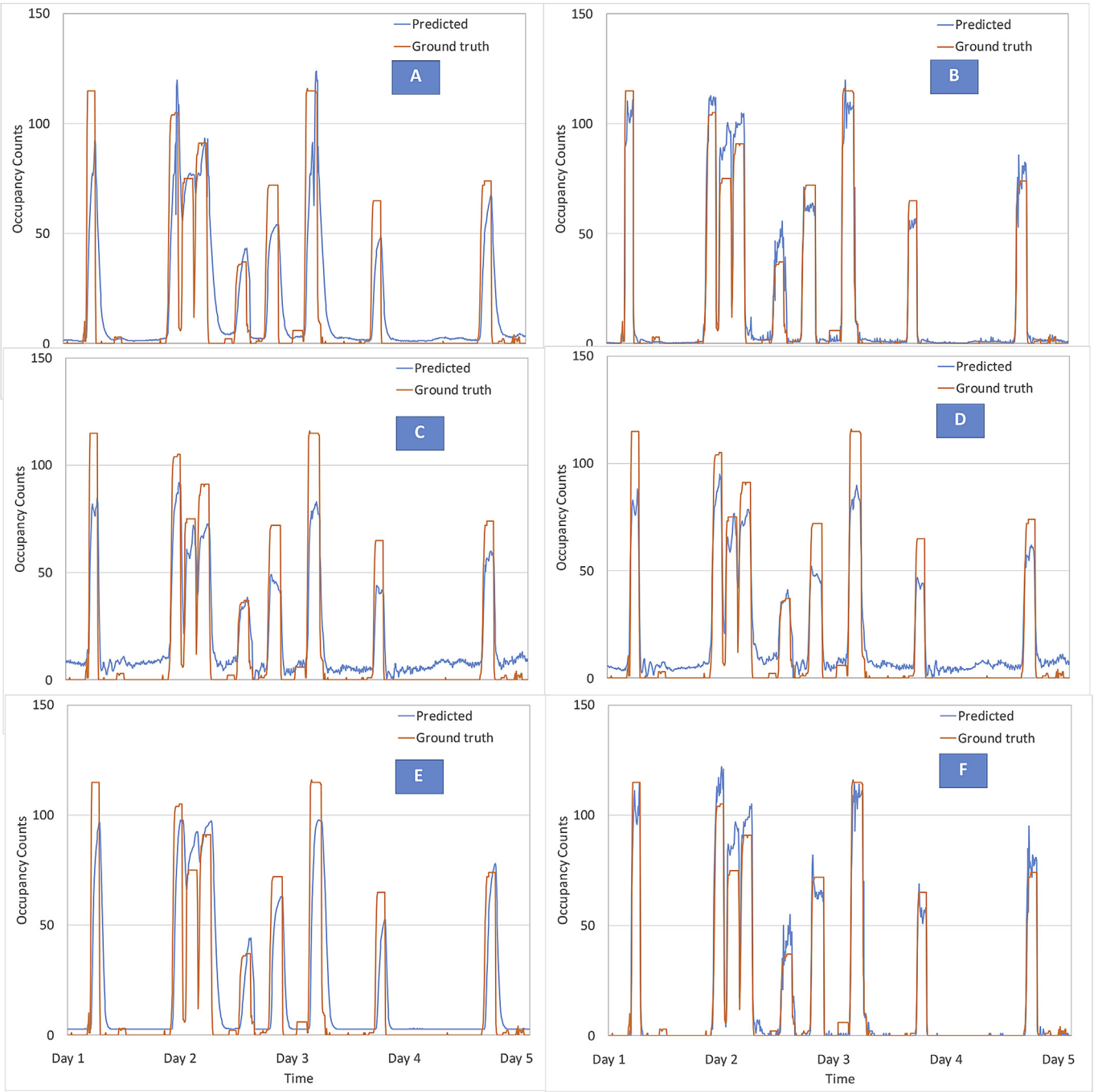

Fig. 5. Profiles of predicted against ground truth occupancy counts for various statistical models (A: ANN - CO_2 AVG; B: ANN - CO_2 AVG + DIF; C: PEM - CO_2 AVG; D: PEM - CO_2 AVG + DIF; E: SVM - CO_2 AVG; F: SVM - CO_2 AVG + DIF) in a typical week.

occupancy estimated by the steady-state model method appear to be as stable as dynamic CO_2 model. However, in Fig. 4C, significant delay compared with the true occupancy profiles can be observed. This delay was larger than that estimated by the dynamic detection model. It is to be noted that the two steady-state models are not able to consider the effects of temperature variations affecting LT AER. Any AER changes caused by the occupancy dynamics will not be able to be realized by the steady models. The use of lagged indoor CO_2 data helped to alleviate this problem to a certain degree (Fig. 4D). Similar to Fig. 4B, the "pointed" peaks at the maximum predicted occupancy counts within each LT sessions in Fig. 4C and D showed that steady-state levels were not reached.

Fig. 5A and E illustrate the models predicted by the ANN, PEM and SVM statistical models. The statistical models using MA data were similar with the model using AVG data, thus only the parameters utilizing AVG and AVG + DIF are presented. From Fig. 5A, C and 5E (models using AVG data), ANN model displayed predicted occupancy counts reaching closest to the ground truth. There appears to be a time lag with the ANN and SVM predicted occupancy counts when compared to the ground truth. During non-occupancy periods, all the statistical models using AVG data overpredict occupancy counts, exhibiting non-zero baseline values. The predicted

occupancy counts improved notably for all models when AVG data is used in conjunction with their first order difference (Fig. 5B, D and 5F). As can be seen in these figures, the predicted occupancy counts traces the ground truth relatively well with no time lag. The predicted occupancy values during non-occupancy period were notably improved especially for the ANN and SVM models.

3.2. Quantitative comparison

The quantitative measures of model fit are given in Table 1. In general, the slope for many models were lower than unity. Only the reference dynamic mass balance model displayed a slope close to 1 and dynamic CO_2, steady state CO_2 lagged, ANN and SVM CO_2 AVG + DIF models meeting the ASTM requirement of greater than 0.75. Correlation coefficients for various models range from 0.80 to 0.97. With the exception of the ANN model utilizing CO_2 AVG + DIF parameters, most statistical models exhibit high intercept values (above the 25% of the average predicted occupancy counts). The lowest RMSE value obtained was derived from the ANN and SVM models utilizing CO_2 AVG + DIF parameters. Only the reference mass balanced model as well as the ANN and SVM models utilizing CO_2 AVG + DIF parameters models accomplished in meeting all the ASTM criteria for the agreement between model predictions and ground truth. The dynamic CO_2 mass balanced model almost meet all ASTM requirements except for a slightly lower 0.89 correlation coefficient. The RMSEs for the models meeting ASTM requirements range between 12 and 13 [18]. reported RMSE values 2 to 6 to predict occupancy count in a test lab that can accommodate up to 20 people using non-conventional machine learning techniques. The simplified steady state physical models does not manage to adequately predict occupancy counts based on ASTM criteria.

Results points to the inadequacies of the statistical models relying purely on CO_2 concentrations (Table 1). Model predictions did not perform better for models where MA parameters are used. However, utilizing CO_2 AVG + DIF parameters greatly improved the model performance measures. It is a general rule of thumb that the more relevant inputs are assigned to each output of a model, the better the fitting gets. In our study, clearly the added information provided by the first order difference in CO_2 at each step is important. This is easy to explain since incorporating the stepwise first order difference in CO_2 concentration is directly linked to occupant numbers as it indicates the magnitude and direction of change. On the other hand, utilizing average or moving average concentrations by themselves contain information in a much more indirect and diffused manner. Furthermore, the PEM, SVM and ANN models assume that each data point is independent and identically distributed. In this case, average and moving average data have strong temporal correlations as illustrated by the autocorrelation function (Fig. 6). We noted significant ($p < 0.001$) correlation values across all lags [7].

Because of the non-linear and varying nature of the factors that affect LT CO_2 concentrations, it is not surprising that statistical model approach better predicts occupancy when compared against the physical models. The latter models are more structured using AER derived from measured data which could not account for the variations in a systematic way.

The results indicate that all the models are more likely to make a false detection when the LT is vacant. False positives range between 18 and 100% with values averaging 58% for the physical models and 83% for the statistical models. For the physical models [14], reported similar estimation errors for a room of up to 3 occupants. To overcome this, the authors have recommended that the control performance of the physical based algorithm to deliberately overestimate the number of people. For the statistical models (with exception of the SVM model utilizing CO_2 AVG + DIF parameters), high F_p values are attributed to elevated and non-stable input values caused during two types of vacant periods. Firstly, this occur during transitional periods at the end of classes where students leave the LT en masse in a very short period, while the CO2 sensors are still picking up high values. Secondly, this occur during vacant periods where CO_2 concentrations displayed inconsistent baseline values (mean CO_2: 527 ppm; minimum CO_2: 429 ppm; maximum CO_2: 1539 ppm). High vacant CO_2 concentrations in the LT typically happened on days where preceding night's CO_2 concentrations remained through the morning. As a result, the statistical models "registered" the vacant periods with these residual CO_2 concentrations causing a positive baseline shift in the predictive output. Interestingly, the SVM model utilizing CO_2 AVG + DIF parameters performed the best in terms of false positives. The SVM is able to acquire relatively low false positives plausibly due to the tenfold cross validation on the training set to avoid overfitting.

The good agreement between dynamic physical model predictions and ground truth while at the same time meeting the ASTM requirement, demonstrates its adequacy provided that the AER measured is accurate. Part of the slightly lower r value between predicted and measured values for dynamic CO_2 model is caused by measurement errors caused by the use of lower AER measured at the end of the lecture sessions. Indeed, as can be seen in Fig. 7, the AER using CO_2 decay at the end of the lecture sessions are slightly

Table 1
Summary of performance to predict occupancy counts for various model.

Modeling Techniques		RMSE	slope	intercept	r	F_b	F_p (%)	F_n (%)
ASTM recommended values			0.75–1.25	within 25% of average	>0.9	Abs <0.25		
Physical Models								
Mass Balanced Models	Dynamic $SF_6^{\ddagger}$	12.8	1.09*	3.29*	0.97*	0.22*	55.5	6.3
	Dynamic CO_2	19.5	0.75*	4.78*	0.89	−0.04*	58.2	5.9
	Steady State CO_2	23.8	0.73	4.19*	0.84	0.03*	58.2	6.9
	Steady State CO_2 Lagged	26.2	0.76*	5.17*	0.87	−0.01*	58.5	5.8
Statistical Models								
ANN	CO_2 AVG	26.3	0.59	9.22	0.79	−0.04*	100	0.0
	CO_2 MA	27.4	0.52	10.07	0.76	0.05*	100	0.0
	CO_2 AVG + DIF	12.6	0.90*	2.86*	0.95*	0.10*	67.9	1.6
PEM	CO_2AVE	19.6	0.59	6.65	0.93*	−0.13*	87.8	2.4
	CO_2 MA	18.3	0.62	6.08	0.95*	−0.04*	75.6	7.2
	CO_2 AVG + DIF	15.8	0.67	7.47	0.96*	0.08*	98.2	0.8
SVM	CO_2 AVG	24.9	0.65	8.60	0.80	0.02*	100	0.0
	CO_2 MA	23.0	0.63	10.42	0.84	0.17*	100	0.0
	CO_2 AVG + DIF	12.1	0.93*	1.66*	0.96*	0.01*	18.2	17.1

*Meeting ASTM recommended values, ‡Reference physical model.

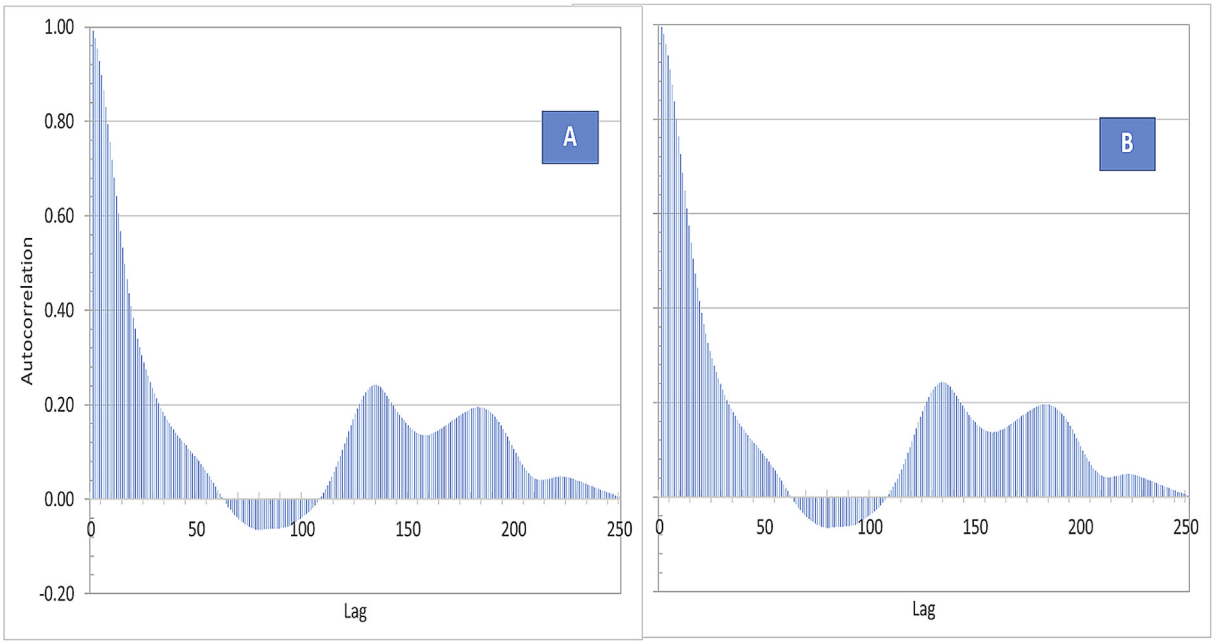

Fig. 6. Plot of autocorrelation versus lag using average CO_2 (A) and moving average (B) data.

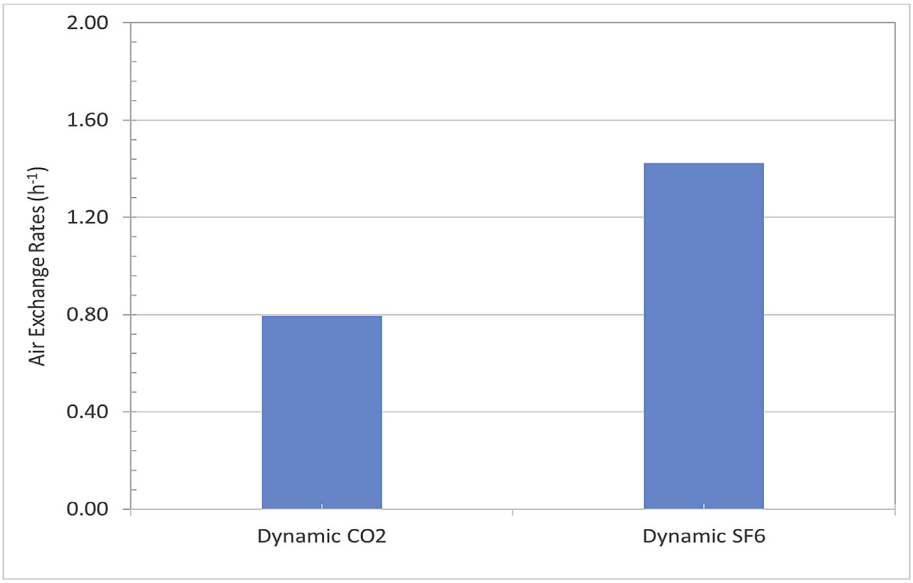

Fig. 7. Air exchange rates for the two dynamic mass balance models.

lower compared to those determined using SF_6 decay method combined with real time fan energy data. To overcome this problem, other ways of dynamically computing AER using CO_2 measurements have to be considered. For example [12], had measured AER in hospital rooms measuring CO_2 concentrations at the supply, return and exhaust air streams while ASTM incorporates a technique to measure percentage outdoor air intake in the supply air intake by measuring CO_2 concentrations at the supply, return and outdoor air streams [2].

Fig. 8 illustrates the accuracy based on α-tolerances for all physical and statistical models. Calculating accuracies with varying tolerances allows for their use in different applications [13,18]. Again, the best performing models are those from ANN and SVM models utilizing CO_2 AVG + DIF parameters. Average accuracies across all tolerance were 70 and 76% were respectively. For physical models, average accuracies across all tolerance ranged from 60 to 62%. With the exception of the SVM model utilizing CO_2 AVG + DIF parameters, all models displayed 0-tolerance accuracy no more than 50%. The accuracy improves significantly with broader occupancy. While these numbers appear to be non-ideal in terms of attaining 100% accuracy, accuracy of predictive models has a strong relation with the actual level of occupancy [5,18]. Unfortunately, this infer that the models are not sufficiently accurate to detect LT occupant presence (i.e. occupancy 'on/off) thus potentially excluding their application to control for the on/off controls of an air-conditioning, lighting or other system. In smaller indoor spaces, presence can be more accurately predicted (60–80%) using conventional machine learning tools [8,11,13]. Dong et al. [13] reported

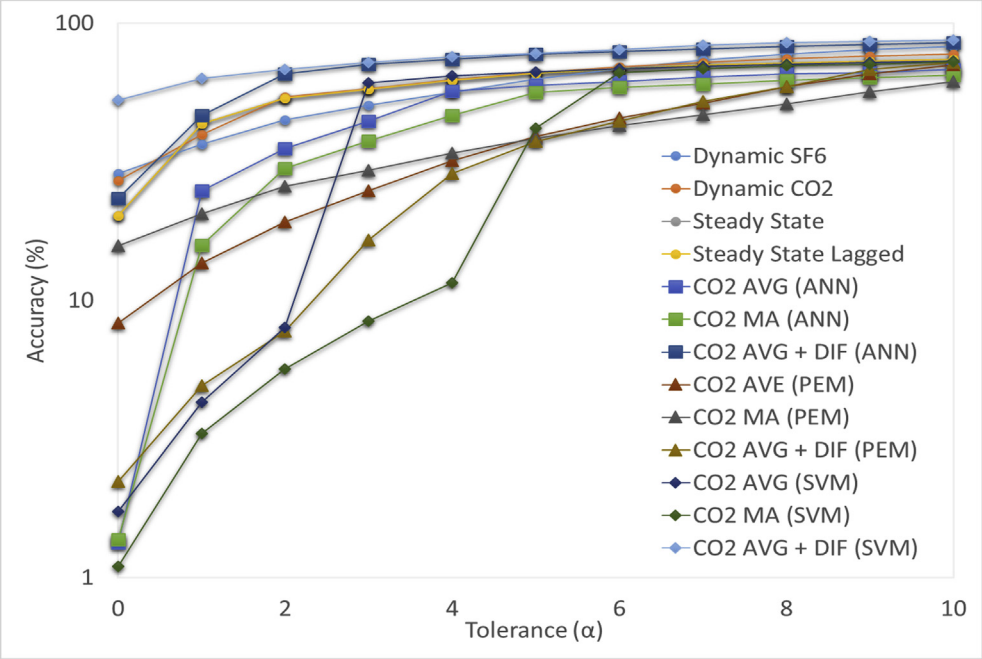

Fig. 8. Accuracy of the various physical and statistical models.

0-tolerance accuracies from SVM and ANN of around 75% to predict occupancy of room filled up to 4 occupants. Other researchers noted that predictive errors for occupant presence increased when the test beds moved from a small room to a larger one [8,25]. Indeed, obtained results from this study and those reported elsewhere highlight the necessity to derive better methods to accurately predict large occupant presence in a space. This is especially important while evaluating the predictive potential of an occupancy model, if the final intent is to be used for operational purposes (i.e., predictive control) in buildings.

3.3. *Factors affecting the models*

An analysis was next performed to provide information on how certain parameter or assumption may affect the prediction results (Table 2).

For the physical model, predictions are dependent on the air mixing assumption as well as CO_2 generation rates. Several researchers have described the problem of CO_2 spatial heterogeneity in indoor air and how the model can be invalidated if the air mixing is an issue [12,19,26,43,44]. To rule out this problem, we used the three sensor locations (Fig. 1) to predict occupancy counts and determine RMSE and regression slope variations for comparison. We noted small variations between −1.5–4.9% to confirm that air mixing was not a problem in this study. For CO_2 generation rates, it has been reported that the amount of CO_2 that is produced by human beings depends on the diet, rate of oxygen consumption, physical activity, age and weight of the occupants [26,43,44]. Other researchers have normally used a value of 0.31 L min^{-1} for an average sized adult engaged in light office work (1.2 met) (see Ref. [12]). A Chinese study has documented CO_2 generation rates of 0.205 L min^{-1} which is about 80% of the value predicted by common empirical equation [28]. Additionally, a study has reported that generation rates vary between 0.17 and 0.21 L min^{-1} depending on environmental conditions that subjects are exposed to [39]. In our study, we have used a value of 0.21 L min^{-1} obtained from a Singapore study involving subjects of the same age group at similar activity levels as this study and where the occupants are exposed to

Table 2
Variations caused by model factors and assumptions.

Factor and Assumption	Description	Variation from RMSE (slope) values[a]
Physical Model – dynamic CO_2		
Air mixing	3 CO_2 sample locations data compared with average values	Sensor a: 0.2 (−1.5) % Sensor b: 0.4 (1.9) % Sensor c: 4.9 (0.4) %
CO_2 generation rates	Singapore subjects, Chinese subjects, ASHRAE values compared with average value	ASHRAE: 27.0 (−32.0) % Chinese: 0.4 (2.8) % Singapore: 7.5 (24.0) %
Statistical Model – PEM, SVM and ANN utilizing CO_2 AVG + DIF parameter		
Continuity data problem	Complete data set compared with unstable information removed from data set	ANN: −4.0 (1.5) % PEM: 0.9 (−0.2) % SVM: −15.8 (1.5) %
Number of hidden layer neuron for ANN	5 and 30 hidden layers compared with 10	5: −1.2 (0.0) % 20: 42.5 (3.6) % 30: 172.3 (−15.9) %
PEM	2^{nd} order and 4th order compared with average value (6th order)	2^{nd} order: 26.4 (12.4) % 4^{th} order: 41.5 (22.7) %

[a] Compared to values reported in Table 1.

similar environmental conditions in the LT [39]. What we found was a variation between −0.4–27% RMSE and −32 to 27% slope caused by differences in used generation rates.

There are occasions when CO_2 data did not merge seamlessly considering we did not include data from weekends and public holiday. For example, a Friday may have ended with a high CO_2 concentration but the subsequent Monday's data began with a low CO_2 concentration. Due to this, the statistical models exhibited momentary and unrealistic high predictive values which subsequently stabilized two or three steps later. To overcome this, we removed these sources of instability and evaluated the effects on the prediction performances. We observed that the SVM models were mostly affected by the absence of seamless data merge. Concerning the PEM model, it is possible to come up with a series of models with varying degree of successes. The main differentiating factor between these models is the model order, i.e. the order of the matrices that describe the relationship between inputs and output. In this case, the greater the order, the more complex the description of the system dynamics. It is generally preferred to maintain as low order as possible without sacrificing model accuracy. For our study, we have determined that the best model was a sixth order model. When this model is compared with a simpler second and fourth order models, the RMSE variations were 26 and 42% respectively demonstrating the impact the model order choice. Sensitivity of the ANN model was next evaluated for 5–30 number of hidden layer neurons compared to 10 that was used in this study. The number of hidden layer neurons vary from problem to problem and depends on the number and quality of training patterns [20]. If too few neurons in the hidden layers are selected, the network can result in under-fitting while too many neurons would result in overfitting. Some researchers have recommended that this number should be equal to one more than twice the number of input neurons [16]. Others reported that the number of neurons obtained using this approach may not guarantee network generalization [29]. Our study revealed that decreasing the number of neurons will not make a significant difference in the performance of the models. However, increasing them the number of neurons caused huge variations in the RMSE and regression slope values.

3.4. Benefits and drawbacks of models

It can been seen that the statistical models are the best performing ones in terms of predicting occupants and the most computationally efficient. Indeed, statistical models are data-driven which has demonstrated durability and data robustness with the added benefits of not being restricted to physical properties of the building space. On the other hand, statistical models need detailed prior information to train the data extensively and get it validated. This means that they cannot be applied immediately after sensors are installed in the building and the models that have been tested can only be limited to the specific environment from which the method was derived. Finally, considering how building physics knowledge are not required for statistical models to predict occupancy counts, there could be plausible issues for building owners to identify specific problems and diagnose non-performing systems.

The use of the physical model offered the several advantages for estimating occupant numbers. Unlike machine learning techniques, physical model does need large amount of training sample data. It require simple a priori information related to AER measurements and did not require detailed building data. Thus, the dynamic model can be set up manually for immediate application. Significantly, physical models are generic in nature and can be applicable for other ventilated rooms. However, it needs to be pointed out that the mass balanced model used in this application are restricted to single-zone space. The model will not be applicable for use in buildings where several air handling units serve the same zone [31,36] or has multi-zone spaces [24]. One also has to confirm that the indoor air is well mixed for the model assumptions to be upheld. Lastly, the steady state version of the physical models do not consider the effects of temperature sensors that could affect the AER dynamics of the space.

Both the physical and statistical models suffer from the effect of residual CO_2 concentration after occupants have left the room. One way of reducing this effect is to combine the CO_2 measurements with other sensors that immediately detect the zero occupancy. A motion sensor such as passive infra-red sensor (PIR) can indicate that the LT has been vacated even though the corresponding short-term CO_2 response is still high [4]. The use of beam-break sensors on doors to detect door opening with certain assumptions of entrance-exit proportions can also be considered [12]. Future work would extend the results of the current findings to include other sensors or different sensor combinations with the goal of determining the best model.

4. Conclusions

This is the first study to report occupancy count prediction in a room up to 200 people using both types of models relying on CO_2 data. The physical models include dynamic and steady state mass balance models while the statistical models utilizes ANN, PEM and SVM machine learning techniques. Using standardized performance metrics recommended by ASTM, performance of various models were evaluated. It was found that that the dynamic physical models and SVM and ANN models utilizing a combination of average and first order differential CO_2 concentrations performed the best. ANN and SVM models showed higher predictive performance compared to the physical model. RMSE values for the corresponding models were 12.8, 12.6 and 12.1 respectively and correlation coefficients were all greater than 0.95. The relatively good agreement between dynamic physical model predictions and ground truth shows that the dynamic mass balanced model is adequate for predicting occupancy counts provided that the AER measured are accurate. Despite the large number of occupants in this study, average accuracies across all tolerance between 70 and 76% demonstrate relatively good performance. The importance of this study is to demonstrate the accurate prediction of occupancy counts in a large space using sensors without significant capital investments. This is significant considering occupancy predication and management has the potential for optimal energy saving and improvement of HVAC operations. Future work is needed to investigate how the models can be improved to enable accurate detection of occupant presence intended to be used for operational purposes (i.e., predictive control) in buildings.

Acknowledgements

The study was supported by the National University of Singapore under CiBEST (BEE Hub) Grant No: C-296-000-029-001.

References

[1] ASHRAE 62.1, ASHRAE standard 62.1 Ventilation for Acceptable Indoor Air Quality, Atlanta, 2013.
[2] ASTM D 6245, ASTM Standard D 6245-12 Standard Guide for Using Indoor Carbon Dioxide Concentrations to Evaluate Indoor Air Quality and Ventilation, American Society for Testing and Materials, West Conshohocken, PA, 2012.
[3] ASTM D 5157, ASTM Standard D 5157-97. Standard Guide for Statistical Evaluation of Indoor Air Quality Models, American Society for Testing and Materials, West Conshohocken, PA, 1997.
[4] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, T. Weng. Occupancy-driven energy management for smart building automation. Proceedings of the 2nd

ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building. ACM 1–6 (2010).
[5] M. Amayri, A. Arora, S. Ploix, S. Bandhyopadyay, Q.-D. Ngo, V.R. Badarlaba, Estimating occupancy in heterogeneous sensor environment, Energy Build. 129 (2016) 46–58.
[6] K.E. Anders Ohlsson, B. Yang, A. Ekblad, C. Boman, R. Nyström, T. Olofsson, Stable carbon isotope labelled carbon dioxide as tracer gas for air change rate measurement in a ventilated single zone, Build Environ. 115 (2017) 173–181.
[7] G.E.P. Box, G. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day, 1976.
[8] D. Calì, P. Matthes, K. Huchtemann, R. Streblow, D. Müller, CO2 based occupancy detection algorithm: experimental analysis and validation for office and residential buildings, Build. Environ. 86 (2015) 39–49.
[9] A.K. Chandran, A. Subramanian, W.C. Wong, J. Yang, K.A. Chaturvedi, A PTZ Camera Enhanced People-Occupancy Estimation System (PCEPOES), The 15th IAPR Conference on Machine Vision Applications, Nagoya, May 8–12, 2017.
[10] C.C. Chang, C.J. Lin, LIBSVM: a Library for Support Vector Machines, 2001. Software available at, http://www.csie.ntu.edu.tw/cjlin/libsvm.
[11] Z. Chen, M.K. Masood, Y.C. Soh, A fusion framework for occupancy estimation in office buildings based on environmental sensor data, Energy Build. 133 (2016) 790–798.
[12] S. Dedesko, B. Stephens, J.A. Gilbert, J.A. Siegel, Methods to assess human occupancy and occupant activity in hospital patient rooms, Build. Environ. 90 (2015) 136–145.
[13] B. Dong, B. Andrews, K.P. Lam, M. Höynck, Y.S. Chiou, R. Zhang, D. Benitez, An information technology enabled sustainability testbed (ITEST) for occupancy detection through an environmental sensing network, Energy Build. 42 (7) (2010) 1038–1046.
[14] M. Gruber, A. Trüschel, J.-O. Dalenbäck, CO2 sensors for occupancy estimations: potential in building automation applications, Energy Build. 84 (2014) 548–556.
[15] M.S. Gul, S. Patidar, Understanding the energy consumption and occupancy of a multi-purpose academic building, Energy Build. 87 (2015) 155–165.
[16] R. Hecht-Nielsen, Theory of the back propagation neural network, Int. Jt. Conf. Neural Networks-IJCNN 1 (1989) 593–605.
[17] P. Hoes, J. Hensen, M. Loomans, B. De Vries, D. Bourgeois, User behavior in whole building simulation, Energy Build. 41 (3) (2009) 295–302.
[18] C. Jiang, M.K. Masood, Y.C. Soh, H. Li, Indoor occupancy estimation from carbon dioxide concentration, Energy Build. 131 (2016) 132–141.
[19] Y.-P. Ke, S.A. Mumma, Using carbon dioxide measurements to determine occupancy for ventilation controls, ASHRAE Trans. (1997) 365–374.
[20] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Predicting hourly cooling load in the building: a comparison of support vector machine and different artificial neural networks, Energy Convers. Manag. 50 (1) (2009) 90–96.
[21] L.J. Lo, A. Novoselac, Localized air-conditioning with occupancy control in an open office, Energy Build. 42 (7) (2010) 1120–1128.
[22] A. Mahdavi, F. Tahmasebi, Predicting people's presence in buildings: an empirically based model performance analysis, Energy Build. 86 (2015) 349–355.
[23] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (1943) 115–133.
[24] L.C. Ng, J. Wen, Estimating building airflow using CO2 measurements from a distributed sensor network, HVAC&R Res. 17 (3) (2011) 344–365.
[25] T.H. Pedersen, K.U. Nielsen, S. Petersen, Method for room occupancy detection based on trajectory of indoor climate sensor data, Build. Environ. 115 (2017) 147–156.
[26] A. Persily, L. de Jonge, Carbon dioxide generation rates for building occupants, Indoor Air (2017). In press.
[27] S. Pigg, M. Eilers, J. Reed, Behavioral aspects of lighting and occupancy sensors in private offices: a case study of a university office building, Proc. 1996 ACEEE Summer Study Energy Effic. Build. 8 (1996) 161–170.
[28] M.W. Qi, X.F. Li, L.B. Weschler, J. Sundell, CO2 generation rate in Chinese people, Indoor Air 6 (2014) 559–566.
[29] M. Rafiq, G. Bugmann, D. Easterbrook, Neural network design for engineering applications, Comput. Struct. 79 (17) (2001) 1541–1552.
[30] C.-A. Roulet, F. Foradini, Simple and cheap air change rate measurement using CO2, Int. J. Vent. 1 (2002) 39–44.
[31] C.-A. Roulet, M.S. Zuraimi, S.C. Sekhar, K.W. Tham, Tracer gasmMeasurement of airflow rates in spaces with several air-handling units, recirculation, or large time constants, HVAC&R Res. 12 (3) (2005) 477–496.
[32] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.
[33] S.H. Ryu, H.J. Moon, Development of an occupancy prediction model using indoor environmental data based on machine learning techniques, Build. Environ. 107 (2016) 1–9.
[34] P.N. Smith, Determination of ventilation rates in occupied buildings from metabolic CO2 concentrations and production rates, Build. Environ. 23 (2) (1988) 95–102.
[35] Z.W. Sun, S.W. Wang, Z.J. Ma, In-situ implementation and validation of a CO2-based adaptive demand-controlled ventilation strategy in a multi-zone office building, Build. Environ. 46 (2011) 124–133.
[36] K.W. Tham, S.C. Sekhar, D. Cheong, Indoor air quality comparison of two air-conditioned zones served by the same air-handling unit, Build. Environ. 37 (10) (2002) 947–960.
[37] P. Van Overschee, B. De Moor, Subspace Identification for Linear Systems, Kluwer Academic Publishers, 101 Philip Drive, Assinippi Pard, Nowell, Kluwer Academic Publishers, Massachusetts, 1999.
[38] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
[39] H.C. Willem, Thermal and Indoor Air Quality Effects on Physiological Responses, Perception and Performance of Tropically Acclimatized People, PhD Dissertation, National University of Singapore, 2006.
[40] D. Yan, W. O'Brien, T. Hong, X. Feng, H. Burak Gunay, F. Tahmasebi, A. Mahdavi, Occupant behavior modeling for building performance simulation: current state and future challenges, Energy Build. 107 (2015) 264–278.
[41] L. Yang, H. Yan, J.C. Lam, Thermal comfort and building energy consumption implications – a review, Appl. Energy 115 (2014) 164–173.
[42] J. Yang, M. Santamouris, S.E. Lee, C. Deb, Energy performance model development and occupancy number identification of institutional buildings, Energy Build. 123 (2016) 192–204.
[43] S. Wang, X. Jin, CO_2-based occupancy detection for on-line outdoor air flow control, Indoor Built Environ. 7 (1998) 165–181.
[44] S. Wang, X. Burnett, J. Chong, Experimental validation of CO_2 -based occupancy detection for demand-controlled ventilation, Indoor Built Environ. 8 (1999) 377–391.