Developing window behavior models for residential buildings using XGBoost algorithm

Hao Mo[1], Hejiang Sun[*1], Junjie Liu[1], Shen Wei[**2]

1 Tianjin Key Laboratory of Indoor Air Environmental Quality Control, School of Environmental Science and Engineering, Tianjin University, Tianjin, 300072, China

2 The Bartlett School of Construction and Project Management, University College London (UCL), 1-19 Torrington Place, London WC1E 7HB, UK

Corresponding author:

* Phone: +86-22-27403620, Email: sunhe@tju.edu.cn

** Phone: +44 7545479329, Email: shen.wei@ucl.ac.uk

Abstract

Buildings account for over 32% of total society energy consumption, and to make buildings more energy efficient dynamic building performance simulation has been widely adopted during the buildings' design to help select most appropriate HVAC (Heating Ventilation and Air Conditioning) systems. Due to the lack of good behavioral models in current simulation packages, many researchers have tried to develop useful behavioral models to improve simulation accuracy, including window behavior models, using field data collected from real buildings. During this work, many mathematical and machine learning methods have been used, and some level of prediction accuracy has been achieved.

XGBoost is a recently introduced machine learning algorithm, which has been proven as very powerful in modeling complicated processes in other research fields. In this study, this algorithm has been adopted to model and predict occupant window behavior, aiming to further improve the modeling accuracy from a globally accepted modeling approach, namely, Logistic Regression Analysis. Field data in terms of both occupant window behavior and relevant influential factors were collected from real residential buildings during transitional seasons. Both XGBoost and Logistic Regression Analysis were used to build window behavior models, after a feature selection work, and their prediction performances on an independent dataset were compared. The comparison revealed that XGBoost has solid advantages in modeling occupant window behavior, over Logistic Regression Analysis, and it is expecting that the same finding would be obtained for other behavioral types, such as blind control and air-conditioner operation.

Keywords: behavior modeling; window behavior; logistics regression; XGBoost algorithm; residential buildings

1. Introduction

Nowadays, buildings account for over 32% of total society energy consumption [1]. According to contribution, occupant behavior, such as opening windows and adjusting clothing insulation, has been widely acknowledged as a crucial aspect by many researchers [2], considering its impact on building energy [3, 4], building retrofitting [5], indoor thermal environment [6, 7], occupant thermal comfort [8, 9] and reducing building energy demand [4, 10, 11]. Occupant window behavior has been considered as having a great impact on both building energy consumption [12] and indoor environment [13], especially for non-air-conditioned buildings.

To select the most appropriate HVAC solutions during the design stage of a building, dynamic building performance simulation is being widely used. In real applications, however, engineers have realized a significant difference between the building's designed performance and actual performance, and this difference has been defined as performance gap [14]. Occupant behavior, including window behavior, has been considered as a major contributor to this performance gap [15]. Therefore, many studies have been carried out to better understand when occupants open/close their windows [16] and to develop useful mathematical models to improve prediction accuracy of dynamic building performance simulation [2, 17].

In existing studies modeling occupant window behavior, logistic regression has been widely adopted [18-23] Logistic regression is a probability-based two-classification algorithm, which is suitable for predicting binary outputs [24], such as window opening/closing. Stazi et al. [25] used logistic regression to model window states for school classrooms in Italy, achieved an accuracy rate of 71.9%. Haldi et al. [26] developed logistic regression models based on data collected from an office building in Switzerland, using outdoor temperature, indoor temperature, wind speed, relative humidity and rainfall as inputs. The model was used for predicting field monitored states of windows and an accuracy of 65% was obtained. Rijal et al. [27] also proposed logistic regression models to predict occupants' window operation in a residential building in Japan, with a final prediction accuracy of 70%. Yun et al. [28] have used this method and developed a model predicting window behavior of occupants in an office building in the UK. Using data collected from two hospital wards in Nanjing in China, Shi et al. [29] have developed a logistic window behavior model, and got a prediction accuracy of 70%.

In recent years, researchers have started to try some other mathematical algorithms or machine learning methods to generate window behavior models for buildings, aiming to get a better accuracy. Pan et al. [30] have used Gaussian distributions to model occupant window behavior, based on data collected from an office building located in Beijing, China. The model employed both outdoor and indoor temperatures as main drivers, and the prediction accuracy has been verified to be 74% based on ACC evaluation index. Wei et al. [31] compared three modeling methods for window behavior, namely, logistic regression model, Markov model and ANN model, and obtained prediction accuracies as 52%, 57% and 73%, respectively. Celi et al. [32] used logistic regression to identify influential factors on states of windows and used Markov chains for prediction. From the study, the most important factors affecting window opening were suggested as time of day and indoor CO_2 concentration, and the most important factors affecting window closing were outdoor temperature and time of day. To evaluate the reliability of models predicting window states, Fabi et al. [33] used data collected from 15 apartments to compare predicted probabilities from four models, one new model and three existing models, and suggested certain common characteristics for window behaviour models with high accuracy.

Barthelmes et al. [34] have adopted Bayesian Network Framework to describe window opening/closing behavior of occupants, and justified that this machine learning method could well capture the stochastic nature of occupants' use of windows. Haldi et al. [14] used generalized linear mixed models to study occupant behavior of opening windows, adjusting shading devices and turning on/off lighting simultaneously, and the mixed models have been validated using field measured data from buildings. Markovic et al. [35] used deep learning algorithm with 5 hidden layers to model window states, and the F1-score of prediction results was 0.53-0.74. Langevin et al. [36] measured temperature, humidity, wind speed and indoor occupancy rate to model window states using the agent-based method, and a BA (Balanced Accuracy, an index to evaluate accuracy) of 0.72 was achieved.

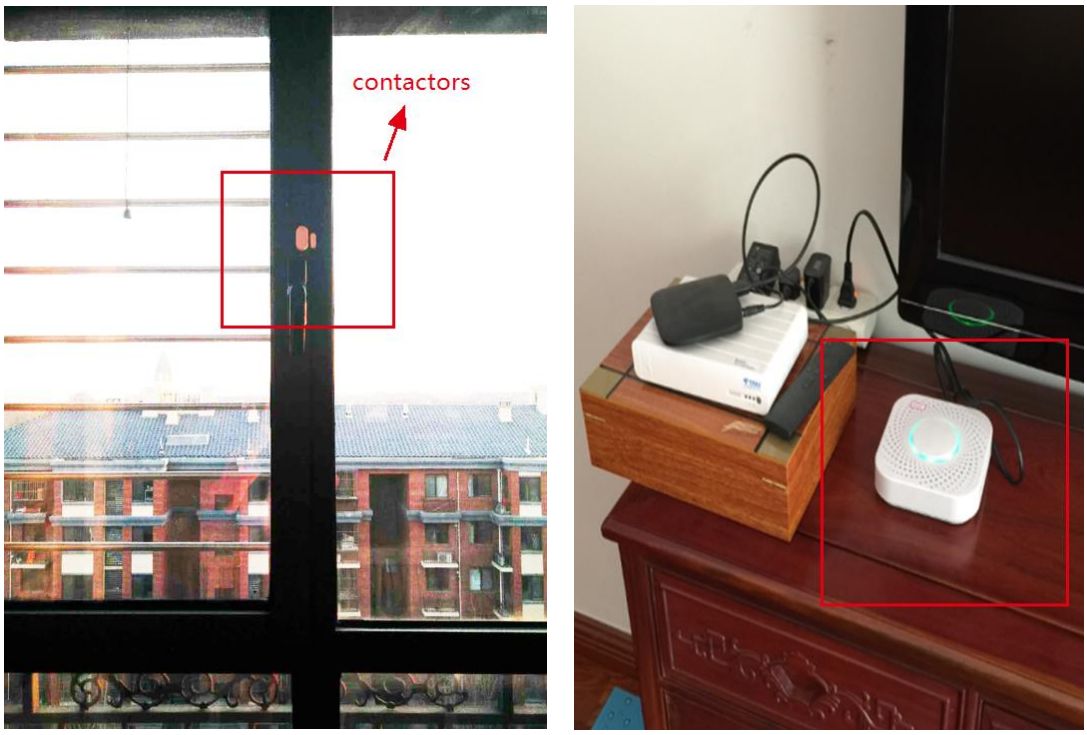
Based on the above review work, it could be summarized that in recent years modeling occupant window behavior using logistic regressions analysis has been challenged by some traditional machine learning methods, such as Gaussian distribution and ANN. This study tried to contribute to this research direction by using XGBoost to model occupant window behavior. As a new machine learning method, the XGBoost (eXtreme Gradient Boost) method was firstly introduced by Chen [37] in 2016, and has been used in many other applications, such as automotive manufacturing [38], predicting building cooling load [39] and fault detection for HVAC systems [40]. In existing studies, much

evidence was available about its advantages (stability, accuracy and efficiency) in modeling complex process over other conventional machine learning methods, such as SVM algorithm [41, 42], logistic regression method [43-49] and KNN/decision tree [50, 51]. This study, therefore, was designed to justify its contribution to modeling accuracy of occupant window behavior in buildings, mainly against the most conventional modeling approach, i.e. logistic regression analysis. Its advantages over other machine learning methods were properly discussed as well. The data used for this study was longitudinally collected from some residential apartments in China, lasting for 136 days during the transitional season. The findings from this study could also be used to guide the selection of modeling methods for other types of adaptive behavior in buildings, such as air-conditioning behavior and shading behavior.

2 Research Methods

2.1 Data collection

The data used in this study were collected from six residential apartments located in the hot summer and warm winter area of China, lasting for a total of 136 days during the transitional season, which was between November, 2017 and March 2018. All selected apartments were coming from the same climatic region of China, hence having similar ambient climate conditions. They all have similar floor areas (approximately 105m^2) and layout (1 living room; 2 bedrooms; 1 kitchen etc.). During the experiment, all of them were using natural ventilation to adjust indoor thermal environment and air quality, and all monitored windows were located in their bedrooms. During the measurement period, the major adaptive opportunity occupants can use to adjust their indoor thermal environment was their windows. In the study, window states were monitored by window contactors (Figure 1a) and several factors that may influence occupant window behavior have been monitored, including indoor temperature, indoor relative humidity, indoor CO_2 concentration, indoor PM2.5 concentration, AQI (Air Quality Index), outdoor temperature, outdoor relative humidity, outdoor PM2.5 concentration, outdoor PM10 concentration and rainfall. Indoor parameters were monitored by an iKair monitoring kit, as shown in Figure 1b, with detailed specifications summarized in Table 1. Outdoor parameters were collected by nearby public weather stations. The data used in this study have been published in key journals after analyzed from different angles to this study [52-54].



(a) (b)

Figure 1: Indoor monitoring sensors

Table 1: Main specifications of indoor monitoring devices

	Range	Accuracy
Temperature	-30ºC~125ºC	±0.3K
Humidity	0-100%RH	±3%RH
$PM_{2.5}$	1-1000ug/m^3	±10% ug/m^3
CO_2	400-10000ppm	±40ppm

2.2 Feature selection

Feature selection is very important for establishing accurate mathematical models [55], by filtering out unnecessary parameters from the modeling work. In this study, Pearson Correlation Coefficient has been used. It is a statistical value that can reflect the similarity between two variables or vectors, and is calculated by Equation 1,

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \cdot \sqrt{E(Y^2) - E^2(Y)}} \quad (1)$$

where, $\mathrm{cov}(X,Y)$ is the covariance; $\sigma_X \cdot \sigma_Y$ is the product of vector standard deviation. $\rho_{X,Y}$ is between -1 and 1, with a value closer to 1 meaning a stronger positive correlation between the two variables, and a value closer to -1 giving a stronger negative correlation. According to [56], the parameter selection was based on a threshold set up as 0.1. It means that when the absolute value of $\rho_{X,Y}$ is smaller than 0.1, the corresponding parameter is not statistically significantly correlated with the model output and therefore will be deleted from the model.

2.3 XGBoost algorithm

XGBoost is a boosting algorithm belonging to supervised learning, which is an ensemble algorithm based on gradient boosted trees [37]. It integrates predictions of "weak" classifiers (tree model) to achieve a "strong" classifier (tree model) via a serial training process. It can avoid over-fitting by adding a regularization term. Parallel and distributed computing makes the learning process faster to give a quicker modeling process. Figure 2 shows a schematic diagram of the computational process of XGBoost and y_i appeared in the process is calculated by Equation 2.

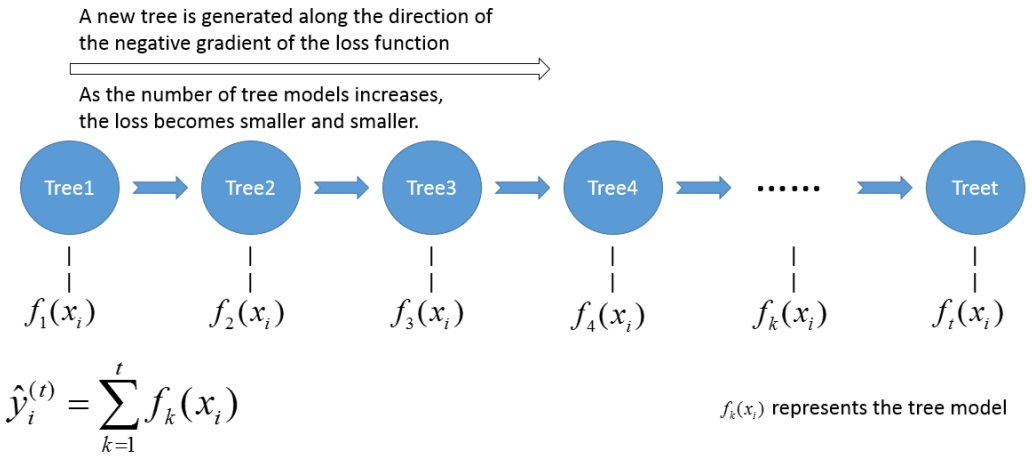


Figure 2. A schematic diagram of XGBoost algorithm

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (2)$$

where, $\hat{y}_i^{(t)}$ is the final tree model; $y_i^{(t-1)}$ is the previously generated tree model; $f_t(x_i)$

is the newly generated tree model, and t is the total number of base tree models.

For the XGBoost algorithm, both depth and number of trees are important parameters. The problem of finding the optimal algorithm was changed into finding a new classifier that can reduce the loss function, with the target loss function shown in Equation 3,

$$Obj^{(t)} = \sum_{i=1}^{t} L(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \tag{3}$$

where, y_i is the actual value; $\hat{y}_i^{(t)}$ is the predicted value; $L(y_i, \hat{y}_i^{(t)})$ is the loss function and $\Omega(f_i)$ is the regularization term.

Substituting Equation 2 into Equation 3 and then following some deduction steps (can be found in [37]), Equation 4 could be obtained,

$$Obj^{(t)} = \sum_{i=1}^{t} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant \tag{4}$$

The final target loss function was then converted into Equation 5, and the model was then trained according to this target loss function.

$$obj^{(t)} = \sum_{i=1}^{t} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{5}$$

where, $g_i = \partial_{y_i^{(t-1)}} l(y_i, y_i^{(t-1)})$ and $h_i = \partial^2_{y_i^{(t-1)}} l(y_i, y_i^{(t-1)})$ are the first and second order gradient statistics on the loss function.

The regularization term $\Omega(f_t)$ is calculated by Equation 6 to reduce the model's complexity and also improve its usability to other dataset.

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \tag{6}$$

where, T is the number of leaves; ω is the weight of the leaves; λ and γ are coefficients, with default values set as λ=1, γ=0.

The XGBoost algorithm can accept both continues variables and discrete variables as inputs but the output variable has to be discrete, including binary variables. In this study, the XGBoost algorithm was ran in Python 3.6 [37]. When using the XGBoost algorithm, Z-statistic is often used for testing the significance of each independent variable, with p-value given at 95% confidence interval [57]. It is calculated and given by the

computational package after running the XGBoost algorithm.

2.4 Window state prediction and evaluation criteria

2.4.1 Window state prediction

In this study, occupant window behavior was predicted using a popular stochastic algorithm adopted in existing studies [28, 58], including 4 steps:

Step 1: Set the initial state of windows as θ;

Step 2: Set closed windows as 0 and opened windows as 1;

Step 3: Calculate the probability, p, to determine state change;

Step 4: Generate a random value, U, following uniform distribution between 0 and 1, and, then compare p with the generated random value. If p>U, set θ = 1 (opened windows); otherwise θ = 0 (closed windows).

2.4.2 ACC

When only considering the state of windows, not the opening angle, occupant window behavior is a binary problem. When evaluating the accuracy on predicting binary problems, a confusion matrix is usually developed, as shown in Table 2, with columns representing predicted values and rows representing actual values.

Table 2: Confusion matrix for window state prediction

Predicted values / Actual values	Positive (window state = 1)	Negative (window state = 0)
True (window state = 1)	TP	FN
False (window state = 0)	FP	TN

where, TP is true positive, FN is false negative, FP is false positive, and TN is true negative.

Some existing studies have used the ACC to evaluate the model's prediction accuracy [30], as defined by Equation 7,

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

The use of a single ACC accuracy index, however, may not sufficient when the states of windows are not equally distributed.

2.4.3 F1-score

Recall and precision are two important indicators to evaluate the performance of classification [59]. Recall is also denoted as sensitivity and it is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Precision is also known as positive predictive value and has been defined as the fraction of relevant instances among the retrieved instances. Equation 8 and Equation 9 were used to calculate both Recall and Precision in this study,

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{9}$$

The F-β coefficient has been used for evaluating the model's predictive performance combining the results from both Recall and Precision [60]. The F-β coefficient is defined by Equation 10 and β can be selected according to the requirement of the application, namely, whether the study paying more attention to the recall rate or to the precision rate.

$$F_\beta = \frac{\left(1+\beta^2\right)\cdot \text{pre}cision \cdot recall}{\beta^2 \cdot precison + recall} \tag{10}$$

$$F_1 = \frac{2 \cdot \text{pre}cision \cdot recall}{precison + recall} \tag{11}$$

when β approaches to 0, $F_\beta \approx$ Recall;

when β approaches to 1, $F_\beta \approx F_1$;

when β approaches ∞, $F_\beta \approx$ Precision.

2.5 Model comparison

Logistic regression has been widely used in existing studies when modeling occupant window opening behavior in buildings [18-22]. It is derived from logistic distribution. The logistic regression builds a linear correlation between all influential factors, such as temperature, and the logit of the probability for event happening, such as window opening, as defined by Equation 12.

$$\log it(p) = \log(\frac{p}{1-p}) = w_1 x_1 + w_2 x_2 + ... + w_n x_n + b \tag{12}$$

where, p is probability of event happening, the coefficients w_1, …, w_n are constants estimated by the regression analysis through maximum likelihood estimation and x_1,…,x_n are the environmental parameters, and b is the intercept.

Solving Equation 12 gives the probability of event happening based on all influential factors, as defined by Equation 13,

$$P = \frac{\exp(w_1x_1 + w_2x_2 + ... + w_nx_n + b)}{1 + \exp(w_1x_1 + w_2x_2 + ... + w_nx_n + b)} \quad (13)$$

Logistic regression has no restrictions to the type of inputs, which can be both continuous or discrete, but the output variable has to be discrete/binary.

3 Results and Discussions

When using the data, it was randomly divided into two parts, with 80% used for developing window behavior models and 20% for model validation. This division has been popularly adopted in existing studies [45, 46, 49, 51, 61]. The whole process included feature selection (to filter out unnecessary parameters), model development (to build window behavior models), model validation (to justify the validity of the developed model on a new dataset) and model comparison (to demonstrate the benefits of the new model against the conventional model method, i.e. logistic regression analysis).

3.1 Selected features

Table 3 has listed the calculated Pearson Correlation Coefficient for each potential influential factor considered in this study.

Table 3: Pearson correlation coefficient for each potential influential factor

T_{in}	**T_{out}**	**RH_{in}**	**CO_2**	**Time of day**	
0.34	0.30	-0.30	-0.30	0.10	
$PM2.5_{in}$	**RH_{out}**	**PM10**	**AQI**	**$PM2.5_{out}$**	**Rainfall**
0.06	-0.03	-0.02	0.008	-0.007	-0.004

According to the 0.1 threshold, it could be found that indoor temperature, outdoor temperature, indoor humidity, indoor CO_2 concentration and time of day performed significant impact on the state of windows during the transitional season. Therefore,

they were remained in the modeling process and the other parameters were filtered out from the study.

3.2 Model development using XGBoost

Using the features selected above, the training dataset described in Section 2.1 has been used to develop the XGBoost model for predicting monitored states of windows. Table 4 has listed some important statistical values when developed the XGBoost model.

Table 4: Statistical values for XGBoost model development

	Correlation coefficient	S.E.	Z-statistic	p-value
T_{out}	0.302	0.018	238.180	0.000
T_{in}	0.345	0.016	-35.099	0.000
RH_{in}	-3.024	0.015	-250.524	0.000
CO_2	-3.03	0.032	-144.064	0.000
Time of day	0.102	0.006	-46.025	0.000

where, S.E. is the standard error, reflecting the degree of dispersion between sample means; Z-statistic is an independent variable test [57]; p-value is given at the 95% confidence interval, used for deciding whether the corresponding independent variable has statistically significant impact on the model's output.

When developing XGBoost models, hyper-parameters need to be determined to drive correlation establishment. Important parameters involved in this study were as followings:

- 'max_depth': the maximum depth of the base tree model, with higher values for more complicated base-tree models;
- 'n_estimators': the number of base tree models, with higher values for more iterations;
- 'min_child_weight': the minimum sum of child node weights, with higher values for more conservative models;
- 'gamma': minimum loss reduction required to make a further partition on a leaf node of the tree, with higher values for more conservative models;
- 'subsample': subsample ratio of training instances;

- ‘colsample_bytree’: subsample ratio of columns when producing new trees;
- ‘reg_lambda’: L2 regularization term on weights, with higher values for more conservative models.

Take max_depth as an example. Its F1-scores of different depth values are listed in Table 5.

Table 5: Values of max_depth for developing the XGBoost model

max depth	2	3	4	5	6	7	8	9	10
F1-score	0.687	0.708	0.714	0.720	0.751	0.777	0.802	0.795	0.791

According to Table 5, when the value of ‘max_depth’ was taken as 8, the performance achieved the best. Therefore, 8 has been selected for ‘max_depth’ in this study. Using the same method, the rest parameters have been determined as well and the results are shown in the Table 6.

Table 6: Values for hyper-parameters for developing the XGBoost model

n_estimators	min_child_weight	gamma
500	4	0.01
subsample	colsample_bytree	reg_lambda
0.8	0.9	0

When using the variables selected in Section 3.1 and the training dataset (80% of the overall dataset) to train the XGBoost model for states of windows, main statistical parameters described in Section 2.3 and the F1-score on the training dataset are listed in Table 7. The depth of the tree represents the maximum depth of the base tree model, with bigger values for more complicated base tree models. The number of trees means the number of base tree models, with bigger values for more iterations.

Table 7: Main parameters and training results of the XGBoost model

Variables	Depth of the tree	Number of trees	F1-score
T_{outt+} T_{in+} RH_{in+} CO_{2+} Time of day	8	500	0.814

3.3 Model validation using independent dataset

To validate the model developed above, it was used for predicting the monitored states of windows in the validation dataset with 20% of overall data, which is independent to the training dataset. Five different sets of validation datasets were taken using validation. The two evaluation parameters, namely, ACC and F1-score, introduced in Section 2.4 were used to reflect the model's prediction accuracy. Due to the stochastic nature of this method, a total of five predictions were tested, with the average of them used for evaluating the overall performance of the model in different validation dataset. Table 8 has listed the results at each prediction.

Table 8: Predicting results of the XGBoost model using the validation dataset

Group number	Precision	Recall	F1-score	ACC
Group 1	0.801	0.815	0.811	0.826
Group 2	0.803	0.807	0.805	0.810
Group 3	0.799	0.804	0.801	0.807
Group 4	0.810	0.817	0.814	0.821
Group 5	0.804	0.808	0.805	0.811
Average	0.811	0.810	0.807	0.815

The results listed in Table 6 reflects that both evaluation parameters gave consistent evaluation results and both values were around 0.81, meaning 81% predictions were correct. The average values of both parameters can well reflect the model's accuracy, and for F1-score it was 0.807, and for ACC it was 0.815, over 5% higher than existing literatures using other modelling methods, such as logistic regression (0.719 [25], 0.65 [26], 0.70 [27]), Markov (0.572 [31]) and Gaussian distribution method (0.74 [30]).

3.4 Comparison with logistic regression

For comparison, the same input variables and training data were again used for the development of a logistic regression model, with important statistical results listed in Table 9 and the final model defined by Equation 13.

Table 9: Important model coefficients and the F1-score on training dataset

$\theta_{T_{out}}$	$\theta_{T_{in}}$	$\theta_{RH_{\text{in}}}$	θ_{CO_2}	θ_{Time}	Intercept α	F1-score on training dataset
2.587	1.166	-2.240	-3.668	-0.297	-0.006	0.552

$$\log it(p) = \log(\frac{p}{1-p}) = 2.587\theta_{T_{out}} - 1.166\theta_{T_{in}} - 2.240\theta_{RH_{in}} - 3.668\theta_{CO_2} - 0.297\theta_{\text{Time}} - 0.006 \quad (13)$$

To justify the advantage of XGBoost over logistic regression, the developed logistic regression model was used to predict the monitored window states in the validation dataset, and its modeling accuracy was compared with the one obtained from the XGBoost model, as used in the last section. Table 10 has listed the prediction accuracies for all five rounds of predictions, with the average accuracy listed at the end of the table. From the calculated F1-scores of both models, it could be obviously observed that the XGBoost model provided much better prediction results than the common logistic regression model, with the former one around 0.8 and the latter one around 0.58. The other parameters, namely, the Precision, the Recall and ACC scores also confirmed this finding. Figure 3 shows both predicted states of windows by the two mathematical methods, as well as the monitored states of windows in a typical day, i.e. 9th February, 2018, at a time interval of one minute. XGBoost achieved an accuracy rate of 84%, while Logistic Regression has an accuracy rate of 63%.

Table 10: Comparison between the XGBoost and the logistic regression models

		Precision	Recall	F1-score	ACC
Group 1	XGboost	0.801	0.811	0.815	0.826
	Logistic	0.552	0.541	0.546	0.593
Group 2	XGboost	0.803	0.805	0.807	0.810
	Logistic	0.554	0.541	0.547	0.593
Group 3	XGboost	0.799	0.801	0.804	0.807
	Logistic	0.552	0.539	0.546	0.592
Group 4	XGboost	0.810	0.814	0.817	0.821
	Logistic	0.551	0.540	0.545	0.591
Group 5	XGboost	0.804	0.805	0.808	0.811
	Logistic	0.554	0.542	0.548	0.594
Average	XGboost	0.811	0.807	0.810	0.815
	Logistic	0.553	0.541	0.546	0.593

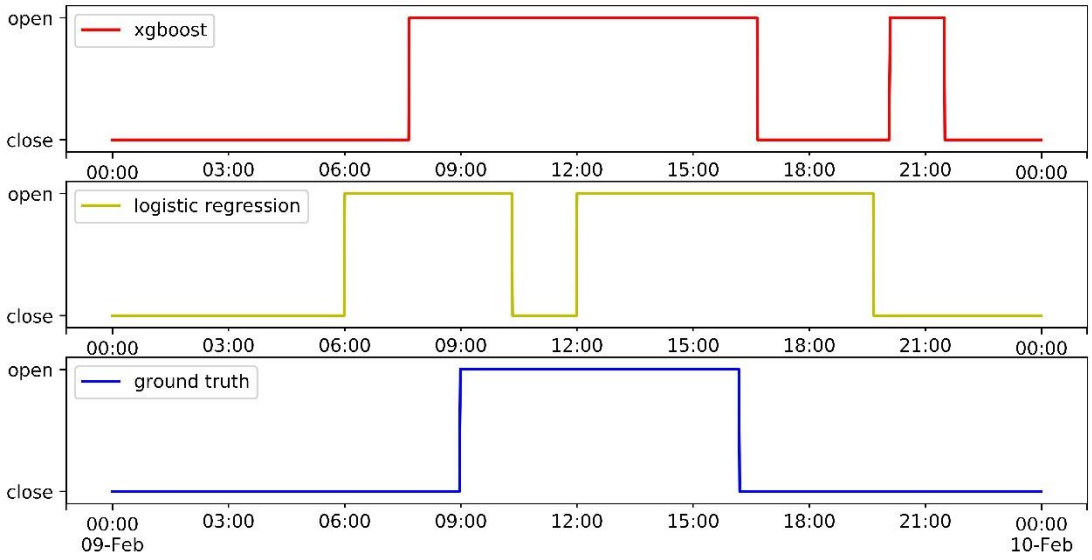


Figure 3. Predicted and monitored states of windows in a typical day

4 Conclusions

Nowadays, it has been widely accepted that occupant window behavior has a significant impact on buildings' energy performance and indoor environment. Accurate modeling of occupant window behavior can provide great support on dynamic building performance simulation to achieve more reliable computer-aided design of buildings. In the past several decades, many modeling algorithms, such as logistic regression and Gaussian distribution methods, have been used for modeling occupant window behavior in buildings, and these modeling techniques have achieved some level of accuracy. This study, has tested a new machine learning method, namely, XGBoost (eXtreme Gradient Boost) algorithm, which is a recently introduced machine learning algorithm that has achieved great achievements in modeling complicated processes in other research fields. Field data including both states of windows and relevant influential factors have been collected from six apartments, lasting for 136 days. Before modeling, feature selection was performed to pick out most useful influential factors for the model output. Both XGBoost, the new method, and Logistic regression, a commonly adopted method, have been used for modeling these monitored states of windows and both models were validated and compared using an independent dataset. Main findings from this study are listed as followings:

1) Feature selection methods can help to simply modeling work by reducing number of inputs for model development;
2) The XGBoost method, as a newly emerged machine learning method, can provide high accuracy on modeling occupant window behavior in buildings, with prediction accuracy around 80%;
3) Comparing to a more common method, namely Logistic Regression Analysis, the XGBoost method has a solid advantage in terms of modeling accuracy (80% vs. 60%) for occupant window behavior.

This study provides a successful preliminary exploration of using XGBoost algorithm for modeling occupant window behavior in buildings, and a great success has been realized. This work expands the use of XGBoost in building applications, especially for occupant behavior modeling, which is a hot research topic in recent years, according to both IEA Annex 66 and Annex 79. The limited sample size in terms of monitored households restricts the reflection of individual variations between people during the modeling process, and this is a common issue of studies monitoring significant

parameters using electronic devices [62], such as 3 offices for Yun and Steemers [28], 5 offices for Li et al. [63] and 3 apartments for Schweiker et al. [64]. However, as the main aim of this study was to test different modeling methods, the number of households being monitored should not affect the research findings mentioned above. In future studies, data from more households will be collected to enhance the sample size to better reflect behavioral variations between people in the modeling work. It is expecting that this newly developed machine learning method could also be used for modeling other types of adaptive behavior in buildings, such as blind control and air-conditioner operation.

Acknowledgement

The research was supported financially by the national key project of the Ministry of Science and Technology, China on “Green Buildings and Building Industrialization” through Grant No. 2016YFC0700500.

Thanks to Dr. Ren Jun, Dr. Gao Yao and Dr. Chen Ximing of Shenzhen Institute of Building Research Co.,Ltd as well as the selected families from Guangdong and Guangxi Province for their assistance with the work.

References

[1] D. Uerge-Vorsatz, L.F. Cabeza, S. Serrano, C. Barreneche, K. Petrichenko, Heating and cooling energy trends and drivers in buildings, Renewable & Sustainable Energy Reviews, 41 (2015) 85-98.

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

[3] M.A.R. Lopes, C.H. Antunes, N. Martins, Energy behaviours as promoters of energy efficiency: A 21st century review, Renewable & Sustainable Energy Reviews, 16 (6) (2012) 4095-4104.

[4] M.M. Agha-Hossein, R.M. Tetlow, M. Hadi, S. El-Jouzi, A.A. Elmualim, J. Ellis, M. Williams, Providing persuasive feedback through interactive posters to motivate energy-saving behaviours, Intelligent Buildings International, 7 (1) (2015) 16-35.

[5] S. Wei, T.M. Hassan, S.K. Firth, F. Fouchal, Impact of occupant behaviour on the

energy-saving potential of retrofit measures for a public building in the UK, Intelligent Buildings International, 9 (2) (2017) 97-106.
[6] T. Peffer, M. Pritoni, A. Meier, C. Aragon, D. Perry, How people use thermostats in homes: A review, Building and Environment, 46 (12) (2011) 2529-2541.
[7] Z.B. Brown, H. Dowlatabadi, R.J. Cole, Feedback and adaptive behaviour in green buildings, Intelligent Buildings International, 1 (4) (2009) 296-315.
[8] R.J. de Dear, G.S. Brager, Thermal comfort in naturally ventilated buildings: revisions to ASHRAE Standard 55, Energy and Buildings, 34 (6) (2002) 549-561.
[9] M. Luo, B. Cao, X. Zhou, M. Li, J. Zhang, Q. Ouyang, Y. Zhu, Can personal control influence human thermal comfort? A field study in residential buildings in China in winter, Energy and Buildings, 72 (2014) 411-418.
[10] K. Steemers, G.Y. Yun, Household energy consumption: a study of the role of occupants, Building Research and Information, 37 (5-6) (2009) 625-637.
[11] H. Darby, A. Elmualim, D. Clements-Croome, T. Yearley, W. Box, Influence of occupants' behaviour on energy and carbon emission reduction in a higher education building in the UK, Intelligent Buildings International, 8 (3) (2016) 157-175.
[12] L. Wang, S. Greenberg, Window operation and impacts on building energy consumption, Energy and Buildings, 92 (2015) 313-321.
[13] S.M. Porritt, P.C. Cropper, L. Shao, C.I. Goodier, Ranking of interventions to reduce dwelling overheating during heat waves, Energy and Buildings, 55 (2012) 16-27.
[14] F. Haldi, D. Cali, R.K. Andersen, M. Wesseling, D. Mueller, Modelling diversity in building occupant behaviour: a novel statistical approach, Journal of Building Performance Simulation, 10 (5-6) (2017) 527-544.
[15] V. Fabi, R.V. Andersen, S. Corgnati, B.W. Olesen, Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models, Building and Environment, 58 (2012) 188-198.
[16] W. O'Brien, K. Kapsis, A.K. Athienitis, Manually-operated window shade patterns in office buildings: A critical review, Building and Environment, 60 (2013) 319-338.
[17] A. Roetzel, A. Tsangrassoulis, U. Dietrich, S. Busching, A review of occupant control on natural ventilation, Renewable & Sustainable Energy Reviews, 14 (3) (2010)

1001-1013.
[18] H. Kim, T. Hong, J. Kim, Automatic ventilation control algorithm considering the indoor environmental quality factors and occupant ventilation behavior using a logistic regression model, Building and Environment, 153 (2019) 46-59.
[19] R. Andersen, V. Fabi, J. Toftum, S.P. Corgnati, B.W. Olesen, Window opening behaviour modelled from measurements in Danish dwellings, Building and Environment, 69 (2013) 101-113.
[20] R.V. Jones, A. Fuertes, E. Gregori, A. Giretti, Stochastic behavioural models of occupants' main bedroom window operation for UK residential buildings, Building and Environment, 118 (2017) 144-158.
[21] B. Jeong, J.-W. Jeong, J.S. Park, Occupant behavior regarding the manual control of windows in residential buildings, Energy and Buildings, 127 (2016) 206-216.
[22] S. Wei, R. Buswell, D. Loveday, Factors affecting 'end-of-day' window position in a non-air-conditioned office building, Energy and Buildings, 62 (2013) 87-96.
[23] D. Yan, T. Hong, Definition and simulation of occupant behavior in buildings, International Energy Agency, EBC Annex 66, 2018.
[24] S.L. Gortmaker, D.W. Hosmer, S.J.C.S. Lemeshow, Applied Logistic Regression, 23 (1) (1994) 159.
[25] F. Stazi, F. Naspi, M. D'Orazio, Modelling window status in school classrooms. Results from a case study in Italy, Building and Environment, 111 (2017) 24-32.
[26] F. Haldi, D. Robinson, Interactions with window openings by office occupants, Building and Environment, 44 (12) (2009) 2378-2395.
[27] H.B. Rijal, M.A. Humphreys, J.F. Nicol, Development of a window opening algorithm based on adaptive thermal comfort to predict occupant behavior in Japanese dwellings, Japan Architectural Review, 1 (3) (2018) 310-321.
[28] G.Y. Yun, K. Steemers, Night-time naturally ventilated offices: Statistical simulations of window-use patterns from field monitoring, Solar Energy, 84 (7) (2010) 1216-1231.
[29] Z. Shi, H. Qian, X. Zheng, Z. Lv, Y. Li, L. Liu, P.V. Nielsen, Seasonal variation of window opening behaviors in two naturally ventilated hospital wards, Building and Environment, 130 (2018) 85-93.

[30] S. Pan, Y. Han, S. Wei, Y. Wei, L. Xia, L. Xie, X. Kong, W. Yu, A model based on Gauss Distribution for predicting window behavior in building, Building and Environment, 149 (2019) 210-219.
[31] Y. Wei, H. Yu, S. Pan, L. Xia, J. Xie, X. Wang, J. Wu, W. Zhang, Q. Li, Comparison of different window behavior modeling approaches during transition season in Beijing, China, Building and Environment, 157 (2019) 1-15.
[32] D. Cali, R.K. Andersen, D. Mueller, B.W. Olesen, Analysis of occupants' behavior related to the use of windows in German households, Building and Environment, 103 (2016) 54-69.
[33] V. Fabi, R.K. Andersen, S. Corgnati, Verification of stochastic behavioural models of occupants' interactions with windows in residential buildings, Building and Environment, 94 (2015) 371-383.
[34] V.M. Barthelmes, Y. Heo, V. Fabi, S.P. Corgnati, Exploration of the Bayesian Network framework for modelling. window control behaviour, Building and Environment, 126 (2017) 318-330.
[35] R. Markovic, E. Grintal, D. Woelki, J. Frisch, C. van Treeck, Window opening model using deep learning methods, Building and Environment, 145 (2018) 319-329.
[36] J. Langevin, J. Wen, P.L. Gurian, Simulating the human-building interaction: Development and validation of an agent-based model of office occupant behaviors, Building and Environment, 88 (2015) 27-45.
[37] T. Chen, C. Guestrin, M. Assoc Comp, XGBoost: A Scalable Tree Boosting System, 2016.
[38] K. Chen, H. Chen, L. Liu, S. Chen, Prediction of weld bead geometry of MAG welding based on XGBoost algorithm, International Journal of Advanced Manufacturing Technology, 101 (9-12) (2019) 2283-2295.
[39] C. Fan, F. Xiao, Y. Zhao, A short-term building cooling load prediction method using deep learning algorithms, Applied Energy, 195 (2017) 222-233.
[40] D. Chakraborty, H. Elzarka, Early detection of faults in HVAC systems using an XGBoost model with a dynamic threshold, Energy and Buildings, 185 (2019) 326-344.
[41] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, Y. Xiang, Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global

solar radiation using temperature and precipitation in humid subtropical climates: A case study in China, Energy Conversion and Management, 164 (2018) 102-111.
[42] M. Nishio, M. Nishizawa, O. Sugiyama, R. Kojima, M. Yakami, T. Kuroda, K. Togashi, Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization, Plos One, 13 (4) (2018).
[43] J.A. Hernesniemi, S. Mahdiani, J.A. Tynkkynen, L.-P. Lyytikainen, P.P. Mishra, T. Lehtimaki, M. Eskola, K. Nikus, K. Antila, N. Oksala, Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome - the MADDEC study, Annals of Medicine, 51 (2) (2019) 156-163.
[44] X. Zeng, J. An, R. Lin, C. Dong, A. Zheng, J. Li, H. Duan, Q. Shu, H. Li, Prediction of complications after paediatric cardiac surgery, European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery, (2019).
[45] J. Taninaga, Y. Nishiyama, K. Fujibayashi, T. Gunji, N. Sasabe, K. Iijima, T. Naito, Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study, Scientific Reports, 9 (2019).
[46] B.J. Mortazavi, E.M. Bucholz, N.R. Desai, C. Huang, J.P. Curtis, F.A. Masoudi, R.E. Shaw, S.N. Negahban, H.M. Krumholz, Comparison of Machine Learning Methods With National Cardiovascular Data Registry Models for Prediction of Risk of Bleeding After Percutaneous Coronary Intervention, Jama Network Open, 2 (7) (2019).
[47] X. Ji, W. Tong, Z. Liu, T. Shi, Five-Feature Model for Developing the Classifier for Synergistic vs. Antagonistic Drug Combinations Built by XGBoost, Frontiers in Genetics, 10 (2019).
[48] Z. Qiao, N. Sun, X. Li, E. Xia, S. Zhao, Y. Qin, Using Machine Learning Approaches for Emergency Room Visit Prediction Based on Electronic Health Record Data, Studies in health technology and informatics, 247 (2018) 111-115.
[49] L. Torlay, M. Perrone-Bertolotti, E. Thomas, M. Baciu, Machine learning-XGBoost analysis of language networks to classify patients with epilepsy, Brain informatics, 4 (3) (2017) 159-169.
[50] Y.-H. Chen, J.-L. Chen, AI@ntiPhish - Machine Learning Mechanisms for Cyber-

Phishing Attack, Ieice Transactions on Information and Systems, E102D (5) (2019) 878-887.
[51] L. Yao, M. Cai, Y. Chen, C. Shen, L. Shi, Y. Guo, Prediction of antiepileptic drug treatment outcomes of patients with newly diagnosed epilepsy by machine learning, Epilepsy & Behavior, 96 (2019) 92-97.
[52] J. Pei, C. Dong, J. Liu, Operating behavior and corresponding performance of portable air cleaners in residential buildings, China, Building and Environment, 147 (2019) 473-481.
[53] D. Lai, S. Jia, Y. Qi, J. Liu, Window-opening behavior in Chinese residential buildings across different climate zones, Building and Environment, 142 (2018) 234-243.
[54] D. Lai, Y. Qi, J. Liu, X. Dai, L. Zhao, S. Wei, Ventilation behavior in residential buildings with mechanical ventilation systems across different climate zones in China, Building and Environment, 143 (2018) 679-690.
[55] R. Zhang, F. Nie, X. Li, X. Wei, Feature selection with multi-view data: A survey, Information Fusion, 50 (2019) 158-167.
[56] M.R. Spiegel, L.J. Stephens, Statistics, McGraw-Hill Education, New York, 2018.
[57] G.Y. Yun, K. Steemers, Time-dependent occupant behaviour models of window control in summer, Building and Environment, 43 (9) (2008) 1471-1482.
[58] S. Wei, R. Jones, P. de Wilde, Driving factors for occupant-controlled space heating in residential buildings, Energy and Buildings, 70 (2014) 36-44.
[59] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and F-score, with implication for evaluation, in: D.E. Losada, J.M. FernandezLuna (Eds.) Advances in Information Retrieval, Vol. 3408, 2005, pp. 345-359.
[60] H. Huang, J. Wang, H. Abudureyimu, A. International Speech Communications, Maximum F1-Score Discriminative Training for Automatic Mispronunciation Detection in Computer-Assisted Language Learning, 2012.
[61] K.K. Dobbin, R.M. Simon, Optimally splitting cases for training and testing high dimensional classifiers, Bmc Medical Genomics, 4 (2011).
[62] J.S. Weihl, P.M. Gladhart, Occupant behavior and successful energy conservation: finding and implications of behavioral monitoring, (1990).

[63] N. Li, J. Li, R. Fan, H. Jia, Probability of occupant operation of windows during transition seasons in office buildings, Renewable Energy, 73 (2015) 84-91.
[64] M. Schweiker, F. Haldi, M. Shukuya, D. Robinson, Verification of stochastic models of window opening behaviour for residential buildings, Journal of Building Performance Simulation, 5 (1) (2012) 55-74.