DATA-DRIVEN APPROACHES FOR ANALYSIS OF BUILDING ENERGY CONSUMPTION AND INDOOR OCCUPANCY BEHAVIOUR

YIXUAN WEI

Ph. D

University of Nottingham

2019

University of Nottingham

Research Centre for Fluids and Thermal Engineering

Data-driven Approaches for Analysis of Building Energy Consumption and Indoor Occupancy Behaviour

Yixuan Wei

A thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

CONTENTS INDEX

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Yixuan Wei

Research Centre for Fluids and Thermal Engineering

University of Nottingham

Ningbo, China

February, 2019

ABBREVIATIONS

AC	air-conditioning
ACC	accuracy
AI	artificial intelligence
ANNs	artificial neural networks
ARIMA	autoregressive, integrated and moving average
BAS	building automation system
BES	Building Energy System
BPNN	back propagation neutral network
BSI	blind system identification
CART	Classification and Regression Trees
CBECS	Commercial Building Energy Consumption Survey
CDA	conditioned demand analysis
CV	coefficient of variation
DAT	daily ambient temperature
DR	demand response/rate
DSM	demand side management
DT	decision tree
DTW	dynamic time warping
EC	energy consumption
EE	energy efficiency
ELA	equivalent leak area
ELM	extreme learning machine
EMD	empirical mode decomposition
EUI	energy use intensity
FFNN	feed-forward neutral network
FCM	fuzzy C-means
FP	frequent pattern

GA	genetic algorithm
GBM	Gradient Boosting Machines
GDP	gross domestic product
GNR	Gauss-Newton regression
GIS	geographic Information System
GPS	global positioning system
HANFIS	hierarchical adaptive network-based fuzzy inference system
HMM	Hidden Markov Models
HLC	heat loss coefficient
HPB	high performance buildings
HVAC	heating, ventilation and air conditioning
KDD	Knowledge Discovery in Databases
LDA	Linear Discriminant Analysis
LRF	local receptive fields
LSTM	long short-term memory
MAPE	mean absolute percentage error
MISO	multiple-input and single-output
ML	Maximum likelihood
MLR	multiple linear regression
MSE	Mean Squared Error
NB	Naïve Bayesian
NN	Neural Networks
NRMSE	normalized root-mean-square error
PCA	principle component analysis
PIR	pyroelectric infrared
RC	retrofit cost
RECS	Residential Energy Consumption Survey
RF	Random Forest
RMSE	root-mean-square error

RMSD	root-mean-square deviation
RNN	recurrent neutral network
SARMA	seasonal auto regressive moving average
SSE	sum of squared errors
SOM	self-organizing map
STLF	Short term load forecasting
STMLF	short term multiple load forecasting
SVM	support vector machine
SVR	support vector regression
TDH	thermal discomfort hours
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
VBD	virtual building database
VRV	variable-refrigerant-volume
WLAN	wireless local area network

PUBLICATION

Journal Publications

- **Wei Yixuan**, Xingxing Zhang, Yong Shi, Liang Xia, Song Pan, Jinshun Wu, Mengjie Han, and Xiaoyun Zhao. "A review of data-driven approaches for prediction and classification of building energy consumption." Renewable and Sustainable Energy Reviews 82 (2018): 1027-1047. (Chapter 2)
- **Wei Yixuan**, Liang Xia, Song Pan, Jinshun Wu, Xingxing Zhang, Mengjie Han, Weiya Zhang, Jingchao Xie, and Qingping Li. "Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks." Applied Energy 240 (2019): 276-294. (Chapter 5 and chapter 6)
- Pan, Song, Xinru Wang, **Yixuan Wei**, Xingxing Zhang, Csilla Gal, Guangying Ren, Da Yan et al. "Cluster analysis for occupant-behaviour based electricity load patterns in buildings: A case study in Shanghai residences." In Building Simulation, vol. 10, no. 6, pp. 889-898. Tsinghua University Press, 2017. (Chapter 7)
- Pan, Song, Yiye Han, Shen Wei, **Yixuan Wei**, Liang Xia, Lang Xie, Xiangrui Kong, and Wei Yu. "A model based on Gauss Distribution for predicting window behaviour in building." Building and Environment 149 (2019): 210-219. (Chapter 4)
- Pan, Song, Fei Pei, **Yixuan Wei**, Hongwei Wang, Jiaping Liu, Xingxing Zhang, Guoqing Li, and Yaxiu Gu. "Design and experimental study of a novel air conditioning system using evaporative condenser at a subway station in Beijing, China." Sustainable cities and society 43 (2018): 550-562.

ACKNOWLEDGEMENTS

I would like to express my sincere and heartfelt gratitude to my supervisors, Professor Liang Xia, Song Pan, Xingxing Zhang and Jinshun Wu, for their continuing support and guidance, as well as their patience, motivation, enthusiasm, immense knowledge and enthusiastic involvement throughout the whole process of my PhD research. I would also like to thank my external supervisor, Professor Yanpeng Wu, for her support and suggestions.

I wish to thank the University of Nottingham, for their financial support of this project. I am also grateful to Beijing Institute of Residential Building Design and Research Co., LTD for providing sufficient operating data from practical buildings. Without their precious support it would not be possible to conduct this research.

Also I thank my friends Bo Pang, Lany Zhang, Shurong Lei, Xuchen Wang, Haowei Yu, Linjun Xie and Manxuan Xiao, for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last several years. They extended their support in a very special way, and I gained a lot from them.

Finally, I wish to thank and dedicate this thesis to my parents for their endless love and guidance. I am so lucky and grateful that both of my parents are senior professors in the direction of HVAC, who unselfishly provide insightful suggestions and encouragements to widen my research from various perspectives. I could never have accomplished so much without them.

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

A, B	Regression parameters	m_y	expected value
b	bias	O	number of occupants
C	Cluster	o	number of occupants
c	Constant of Lagranian	o_{max}, o_{min}	maximum and minimum number of occupants
d	Squared Euclidean distance	$\dot{Q}^{vent,sup}$	supply fresh air rate
D	Distance of clusters	$\dot{Q}^{vent,exh}$	exhaust mechanical ventilation rate
E	Information entropy	ε	threshold
e	Orthogonal eigenvector	$\alpha, \tilde{\alpha}, \beta, \tilde{\beta}$	Regression parameters
g	Learning rate	λ	Eigenvalue
K	number of clusters	μ	Cluster centre
$i, j, k, l, m,$ n, N, p	Number	$\dot{Q}^{leak,in}$	inflow of leakages
H	distance matrix	$\dot{Q}^{leak,out}$	outflow of leakages
Q	Statistic value	t	time
Q_a	Statistical threshold	T	number of the samples in a day
P	Probability	u	actual level of fresh air system
O	Variance-covariance matrix	V	volume of considered room
R^2	Coefficient of determination	y	measured CO_2 concentration
$f(\cdot)$	Activation function	y_s	smoothed CO_2 data
$F(\cdot)$	Decision function	Y, U, O	Toeplitz matrix of **y**, **u** and **o**

$h(\cdot)$	Fitness function	σ	variance
$J(\cdot)$	Squared error function	ω	regular factor
$K(\cdot,\cdot)$	Kernel function	$\lambda_y, \lambda_u, \lambda_o$	scaling factor
$L(\cdot)$	Lagranian function	θ	unknown parameters
$W(\cdot)$	Dual optimization of Lagranian	$\beta y, \beta u, \beta o$	shaping parameter
$var(\cdot,\cdot)$	Variance	μ	configuration parameter
$cov(\cdot,\cdot)$	Covariance	Δ	dimensional identify matrix
$\varphi(\cdot)$	Non-linear function	Δy_k	difference between y_k and y_{k-1}
T	Temperature	∇	gradient
$w, w^{(\cdot)}, \widetilde{w}$	weights	$E(\cdot)$	energy function
X	Chromosome	$m(\cdot)$	average value
x	Input	$\mathcal{N}$	normal distribution
Y	Target	$\sigma(\cdot)$	standard deviation
y	Output	$L(\cdot)$	likelihood function
Z	Principle component	$\log L(\cdot)$	log-likelihood function
a, a^*, η, η^*	Lagranian multiplier	cov	variance value
ξ, ξ^*	Slack variables	$\hat{\bullet}$	estimated value
C	outdoor CO_2 concentration	$P(\cdot)$	posterior distribution
$\bar{C}$	indoor CO_2 concentration	$T_n(\cdot)$	Toeplitz matrix
e	measurement error	$\|\cdot\|_2^2$	l_2-norm
g	CO_2 generation rate per person	$\mathbb{I}(\cdot)$	indicator function
g_y, g_u, g_o	transfer function	$\mathbb{x}(\cdot)$	indicator function
k	Discrete time domain	$\tau(\cdot)$	x-tolerance accuracy

$K_{\beta y}$, $K_{\beta u}$, $K_{\beta o}$	covariance matrix	$\mathbb{R}$	natural number

ABSTRACT

A recent surge of interest in building energy consumption has generated a tremendous amount of energy data, which boosts the data-driven algorithms for broad application throughout the building industry. In addition, occupancy behaviour is an important influencer of energy consumption in building. Currently the shallow understanding of occupancy has leaded to considerable performance gap between prediction and measurements of energy use. In this work, data-driven approach, mathematical approach and blind system identification model are developed to investigate building energy consumption and indoor occupancy behavior.

As for the occupants' "active" influence in building energy, we characterize residential appliance usage utilizing the K-means clustering approach through case studies and present the complex residential electricity behaviors in Shanghai. Similarly, the occupant's "passive" role also has impact on the building performance. Among different passive occupant behaviors, window opening action and occupant profile have been deeply investigated in our research. Furthermore, in order to identify the impact of occupancy behaviors on building energy, prediction models based on the artificial neural network (ANN) are established to predict the electricity consumption of the air-conditioning system at the next time step, the superiority of the ANN model with the supplementary input of estimated current occupancy is verified by comparing the ANN model results without the input occupancy.

In summary, the proposed approaches provide a new and detailed way for engineers and building operators to better understand occupant behaviors and their impacts on building performance. Therefore, dedicated energy-prediction models with

consideration of occupancy provide an opportunity to couple the electric grid and the building's control actions, and to be utilised by buildings and utility companies to simultaneously optimise their performance.

CHAPTER 1: INTRODUCTION

1.1 Research background

The global contribution from buildings towards energy consumption has steadily increased reaching figures between 20% and 40% in developed countries and about 1/3 of greenhouse gas emission. The case of China is particularly striking: the country only takes two decades to double its building energy consumption at an average growing rate of 3.7% [1.1]. These facts demonstrate that to facilitate energy efficiency of building is a cost-effective resource for reducing energy consumption and carbon emission from building [1.2]. Also, large potential saving in economy has been anticipated by a large variety of previous studies. For instance, Nikolaidis, *et al*. have indicated that among various energy saving measures for common building types, isolation of roof constitutes the most superiority nearly €5,000 economic benefit during past 30 years [1.3]. As the central approaches transmitting to energy efficiency, prediction and classification of energy consumption in building are significantly necessary with the aim to improve building performance, reduce environmental impact, and estimate economical potential for further energy conservation and renewable energy program [1.4].

Energy consumption in building has been extensively analyzed by substantial studies for the entire building lifecycle, with different focuses on identifying the energy use for different sub-components at the building level [1.5, 1.6] or measuring energy performance in a nationwide analysis [1.7-1.10]. This comprehensive set of analysis on different levels could help us not only optimize the energy use of a

particular dwelling through appropriate retrofit in building envelop or inclusion of state-of-the-art renewable energy technologies (at the microscale), but also explore possible energy reduction opportunities and establish better urban-sustainability strategies (at the macroscale).

Management and optimization of building energy consumption call for a full understanding of building performance, which should first identify energy resources and major end-uses of a building. Energy resources in a building usually refer to electricity, natural gas and district heating supply. The corresponding major end-uses include heating, ventilation and air-conditioning (HVAC) system, domestic hot water, lighting, plug-loads, elevators, kitchen equipment, ancillary equipment and appliances. **Fig. 1-1** illustrates a representative classification of building energy use adopted in ISO Standard 12655:2013 [1.11]. Note that on top of the above building energy resources and major end-uses, HVAC operation schedule and indoor/outdoor conditions are also two important contributing factors to be considered in a building performance analysis.

Generally, reliability of a building performance analysis counts heavily on the qualified datasets used in analysis, which should contain sufficient energy consumption information of the buildings under investigation. Facility managers or research institutes always ask for utility bills for electricity and natural gas from power supply companies as these are the common type of databases of building energy consumption. They also collect information via survey and questionnaire for large-scale buildings, such as the residential sector (Residential Energy Consumption Survey (RECS), EIA, 2009) and commercial buildings (Commercial Building Energy

Consumption Survey (CBECS), EIA, 2012) [1.12]. In addition, in today's building performance analysis, virtual building database (VBD) developed from simulation software (e.g. TRNSYS and EnergyPlus) and energy disclosure laws [1.13, 1.2] (e.g., US Energy Information Administration database) are the other two possible data resources. It is particularly worth mentioning that the empirical datasets taking advantage of smart meters and building energy system have emerged in recent years. These databases substantially improved accuracy and reliability of the related analyses [1.14] despite their expensive costs and technical complexity involved for many practical commercial-uses.

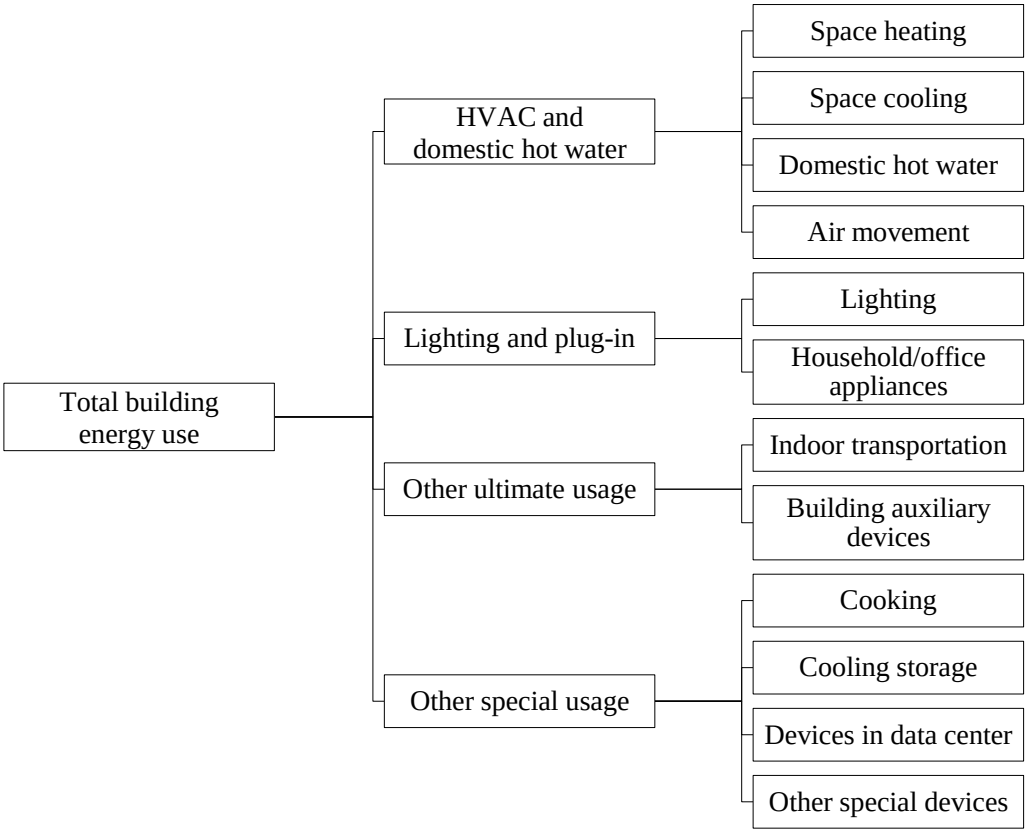


Fig. 1-1: Usage of energy in buildings [1.11]

It is a challenging task to precisely describe energy consumption in a building as such an energy performance depends on a wide range of factors, such as weather condition, thermal properties of building envelope, occupancy behaviour, sub-level components' (lighting, HVAC and plug equipment) performance and schedules [1.4]. A large number of efforts have been paid in the literature to ascertain the complexity pertinent to building energy consumption and strive to a precise depiction of building energy performance. Currently, these approaches used for building energy simulation are categorized roughly as: (1) white-box based approaches, (2) grey-box based approaches and (3) black-box based approaches, whose main features are summarized in **Table 1-1**.

White-box based approaches are physical-based approaches, which require detailed information of complex physical process taking place in building and building energy consumer. This basic characteristic makes their simulations computationally expensive. Recently, a series of attempts have been made to simplify the white-box based approaches. However, these simplifications are error-prone and usually overestimate energy-saving of buildings [1.16, 1.17]. Grey-box based approaches are a modification of these white-box based approaches through use of statistical methods combining the simplified physical information with historical data to simulate building energy. One primary issue in current grey-box version is computational inefficiency (e.g. complexed establishment of model and detailed input parameters) as the approaches involve using uncertain inputs and complex interactions among elements and stochastic occupant behaviours [1.18, 1.19]. To circumvent the above shortfalls of white- and grey- box based approaches, black-box based approaches are

developed which are capable of a building energy consumption analysis only based on historical data without the detailed knowledge of on-site physical information. This essential change enable black-box based approaches fast calculations in high accuracy in comparison to their white- and grey-box counterparts [1.4]. In many practical scenarios, the black-box based approaches are also called as data-driven approaches due to the statistical algorithm structures and a large amount of data in use. We will follow this convention and use the data-driven approaches throughout the following discussion in this review.

Table 1-1: Comparison among white-box, grey-box and black-box approaches for building energy consumption

APPROACHES FOR BUILDING ENERGY SIMULATION		INPUTS NEEDED	TYPICAL SOFTWARE & METHODS	EASY TO USE	RUNNING SPEED	ACCURACY
White-box based	Elaborated simulation	Detailed physical information	EnergyPlus, TRYSYS, DeST, ESP-r	No	Low	High
	Simplified simulation		Degree day method, temperature frequency method, residential load factor method	Yes	High	Fairly high
Grey-box based		Physical information & historical data	RC network	No	Low	Fairly high
Black-box based		Historical data	ANNs, SVMs, statistical regression, GA, cluster	No	High except	High except regressio

algorithms SVM n

The energy consumption of buildings is mainly affected by six factors: meteorological conditions, building envelope, building equipment, indoor environmental parameters, operation management and occupant behaviour. In the past decade, significant progresses and advanced technologies have been developed in terms of above aspects, expect occupant behaviour [1.20, 1.21]. Some methods, such as the survey-based approach [1.22-1.27], Data-driven approaches [1.28, 1.29], and building performance simulation (BPS) [1.30] have been used to evaluate the impact of occupant behaviour on buildings performance. Recently, building industry steps up to facilitate energy-efficient buildings with comfortable and healthy indoor environment for the occupants. Recently, numerous studies [1.23-1.27, 1.30-1.46] have confirmed that occupant's interaction with building systems would attributes to sizeable variation in building energy consumption. For example, Takasu et al. [1.27] conducted questionnaire based field surveys to record thermal comfort responses of occupants and found that find behavioural adaptation related to window-opening leading to variation in the comfort temperature across different seasons. However, in conventional simulation packages, occupant behaviour is described in the form of either fixed schedules or rule-based methods, which fail to capture the stochastic nature of occupant behaviour. This simplification of occupant behaviour will significantly reduce the reliability and accuracy of results from building performance simulation [1.42-1.46]. On the other hand, many simulated measurements or retrofits [1.47-1.52] with significantly energy-saving potential often fail to reach expected performance in real situations. Sometimes

the situation would be even worse that energy consumption of real buildings is increased after simulated energy-efficient measurements are adopted [1.53-1.57], one of the important reasons for this phenomenon is the huge deviation between the behavioural modeling in simulation and actual occupants behaviour. Therefore, designers should consider the complexity and variation of occupant behaviour when designing buildings. Undoubtedly, inaccurate descriptions concerning occupant behaviour would inevitably result in great deviation between building design and operation, and this deviation is often referred as "performance gap" [1.58-1.59]. Therefore, better understanding and more accurate modeling of occupant behaviour in buildings is vital to bridge the gap between simulation result and actual building performance, especially for those buildings that largely depend on passive design features and occupancy controlled technologies [1.60-1.61].

1.2 Research objective

As effective and useful techniques providing profound insights and possible strategic solutions in policy and management of building energy consumption, data-driven approaches have been deemed as favorable means for facilitate future in-depth studies on building energy efficiency. However, other factors, such as occupancy behaviour and equipment energy-performance coefficient, are also equally important. This indicates an ideal data-driven model should make use of multiple indexes to provide a comprehensive analysis of building performance, instead of the current single output of energy consumption or heating/cooling loads. Significantly, apart from the basic functions (i.e., prediction and classification), significant outlook of data-driven techniques targeting decision-making machine is important, such as

occupancy behaviour recommender and equipment operation instructor. These data-driven based developments in building industries would offer real-time on-site information for thermo-comfortable accommodation with minimum energy consumption. To achieve this goal, the research set out five interlinked objectives, as below:

(1) To carry out an extensive literature study of data-driven models for building energy analysis and occupant behaviour, identify the existing challenges and suggest potential solutions.

(2) To establish load profiling in residential buildings and understand complex electricity behaviour by exploring the situation in a developing country utilizing long-term electricity consumptions.

(3) To fully investigate the occupant window opening models based on mathematical approaches and data-driven approaches, and to increase the prediction accuracy of the proposed models.

(4) To providing a non-intrusive and accurate model to estimate the number of occupants in office buildings. Different from most of works which depended on training sets, this model could be used to blindly compute the number of occupants.

(5) To develop a prediction model of the electricity consumption of an AC system based on occupancy information.

To be specific, as for the occupants' "active" influence in building energy, we characterize residential appliance usage utilizing the K-means clustering approach through case studies and present the complex residential electricity behaviors in

Shanghai. Similarly, the occupant's "passive" role also has impact on the building performance. Among different passive occupant behaviors, window opening action and occupant profile have been deeply investigated in our research. Furthermore, in order to identify the impact of occupancy behaviors on building energy, prediction models based on the artificial neural network (ANN) are established to predict the electricity consumption of the air-conditioning system at the next time step, the superiority of the ANN model with the supplementary input of estimated current occupancy is verified by comparing the ANN model results without the input occupancy.

1.3 General Description of the Research Concept

The overall concept, as shown in **Fig. 1-2**, is to use data-driven approach, mathematical approach and blind system identification model to investigate building energy consumption and indoor occupancy behaviour. The approaches to processing the scientific and technological works were as follows: (1) identification of current R&D status of data-driven models and practical applications while pointing out existing challenges in their development for prediction of building energy consumption and occupancy behaviour (for objective 1); (2) characterize residential electricity load patterns utilizing the standard K-means clustering approach through case studies and present the complex residential electricity behaviours in Shanghai (for objective 2); (3) indoor occupant behaviour analysis including window opening behaviour and occupant number estimation (for objective 3 and 4), model training and model validation are carried out after model establishment; (4) to establish prediction

models based on the artificial neural network (ANN) model to predict the electricity consumption of the AC system at the next time step, the superiority of the ANN model with the supplementary input of estimated current occupancy is verified by comparing the ANN model results without the input occupancy (for objective 5).

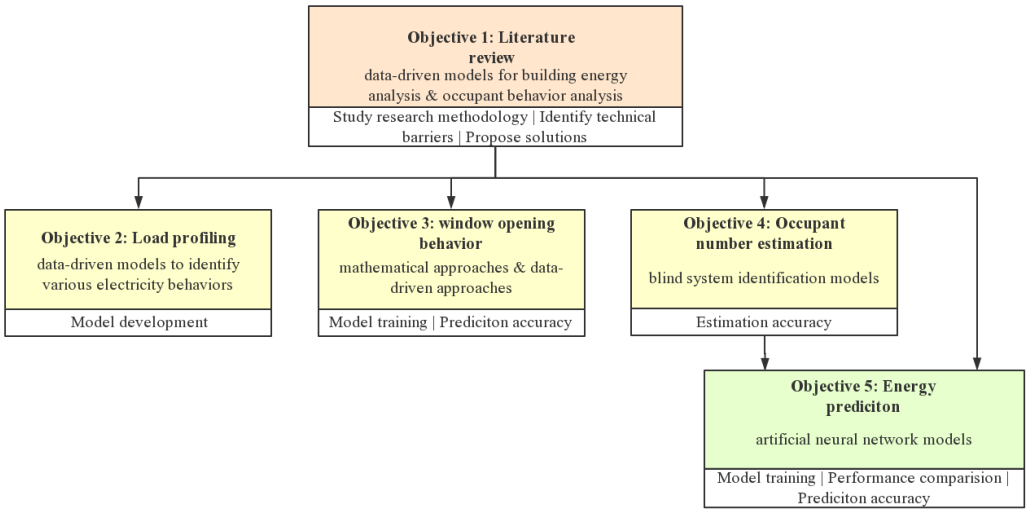


Fig. 1-2: Schematic of the research concept

1.4 Research novelty

In brief, the research has the following identifiable novel aspects:

(1) This research aims to fill the research gap of load profiling in residential buildings by exploring the situation in a developing country utilizing long-term electricity. At the initial investigation stage, this study characterized residential electricity load pattern utilizing the standard clustering approach through case studies in Shanghai, China. This research presents the complex electricity behaviour of two residential communities in Shanghai in a

meaningful way and provides recommendations for different stakeholders based on the findings.

(2) This research aims to fully investigate occupant window opening behavior during transition season based on mathematical approach and data-driven approach, which is proposed to explore the application and optimization of ANN algorithm and Gauss distribution model. Moreover, $PM_{2.5}$concentration is considered as an influencing factor to build window opening model of office building in China area.

(3) Furthermore, an occupancy model is proposed with the aim to providing a non-intrusive and accurate algorithm to estimate the number of occupants based on CO_2 concentration in office room. Unlike existing occupancy-estimation models, this model requires no prior knowledge of people-counting data or extra training steps. In addition, the estimation result is more accurate than that by using the analytical method because the measured error is eliminated and prior experience parameters are considered during calculation.

(4) Different data-driven models are constructed as the prediction performance of building energy consumption. Then, we compare the performance of data-driven models with the different supplementary inputs, i.e., true occupancy, estimated occupancy, and without inputs of occupancy. Overall, this research focuses on bridging the gap between energy-prediction models and the dynamic occupancy profile estimated from indoor CO_2 concentration.

1.5 Thesis structure

Chapter 1 – Introduction: we briefly describe the research background, objectives, research concept and novelty.

Chapter 2 – Literature review: we review the prevailing data-driven approaches used in building energy analysis under different archetypes and granularities including those for prediction and those for classification.

Chapter 3 – Load profiling model: we extract occupant-behaviour related electricity load patterns using classical clustering approach at the initial investigation stage. Smart-metering data from a case study in Shanghai, China, was used for the load pattern analysis. The electricity load patterns of occupants were examined on a daily/weekly/seasonal basis.

Chapter 4 – Window behaviour model: we compare different models of occupants' window behaviour including logistic regression, Markov process, Gauss distribution model, and ANN model, which is proposed to explore the application and optimization of ANN algorithm under less samples condition. Moreover, outdoor $PM_{2.5}$ concentration has been considered as an influencing factor to build window opening model of office building during transition season in China area.

Chapter 5 – Occupancy estimation model: we estimate the number of occupants explicitly; a combination of multiple common measurements is used, including real-time CO_2 concentration, electricity consumption due to indoor appliances and fresh air

system. The newly-developed parameter estimation models could be used to compute the occupancy level blindly.

Chapter 6 – Energy prediction model: we establishe the prediction model of the electricity consumption of the air-conditioning system by using different data-driven models. To analyse some aspects of the benchmark test for identifying the effect of structure parameters and input-selection alternatives, three studies are conducted on 1) the effect of predictor selection based on principal components analysis (PCA), 2) the effect of the estimated occupancy as the supplementary input, and 3) the effect of the neural network ensemble.

Chapter 7 – Conclusion and future work: we present the conclusion of this thesis and proposed future.

CHAPTER 2: LITERATURE REVIEW

This chapter will carry out a critical review of R&D progress and the practical application of data-driven models in building energy consumption and occupant behaviour. The major aims are briefly given as follows:

(1) Introduce the most popular data-driven prediction models as well as the classification models.

(2) Present the applications of these data-driven models including load prediction, energy pattern profile of specific use-cases, regional energy consumption mapping, energy benchmark for building stock, retrofit strategies and guideline making.

(3) Illustrate a comprehensive literature review into the R&D works of window behaviour models.

(4) Present several occupancy detected methods and various models which are proposed to estimate the number of occupants.

(5) Identify the performance gap between the predicted and measured building energy use when the occupancy presence and behaviour are neglected.

(6) Identify the research inadequacy of residential occupant behaviour which is more complex and characterized by randomness.

(7) Discuss the opportunities for further development of data-driven models for building energy consumption.

2.1 Data-driven approaches for prediction and classification of building energy consumption

2.1.1 Data-driven approaches

Data-driven models are constructed based on a group of datasets consisting of historical data records. These historical data will be used as benchmarks to justify the model's performance and guide its algorithm design. To be specific, all the parameters in a data-driven model will be carefully selected and modified through systematical comparisons between the model outputs and the historical data. This is the so-called learning process and only when the output errors fall within the required threshold, the corresponding data-driven models are deemed to be qualified for practical applications with fresh input data. Currently, the data-driven models is very prevailing in medical diagnosis [2.1], political campaigns [2.2] and commerce [2.3] because of their low costs with no need of expensive equipment and audit activity. As to the building energy consumption studies, data-driven models are widely applied to either estimate the building energy demands (i.e., data-driven prediction models) or profile the energy consumption patterns (i.e., data-driven classification models), which are grouped in **Fig. 2-1**.

Among the most popular data-driven prediction models are artificial neural networks (ANNs), support vector machine (SVM), statistical regression, decision tree (DT) and genetic algorithm (GA). This subsection will introduce each of these models.

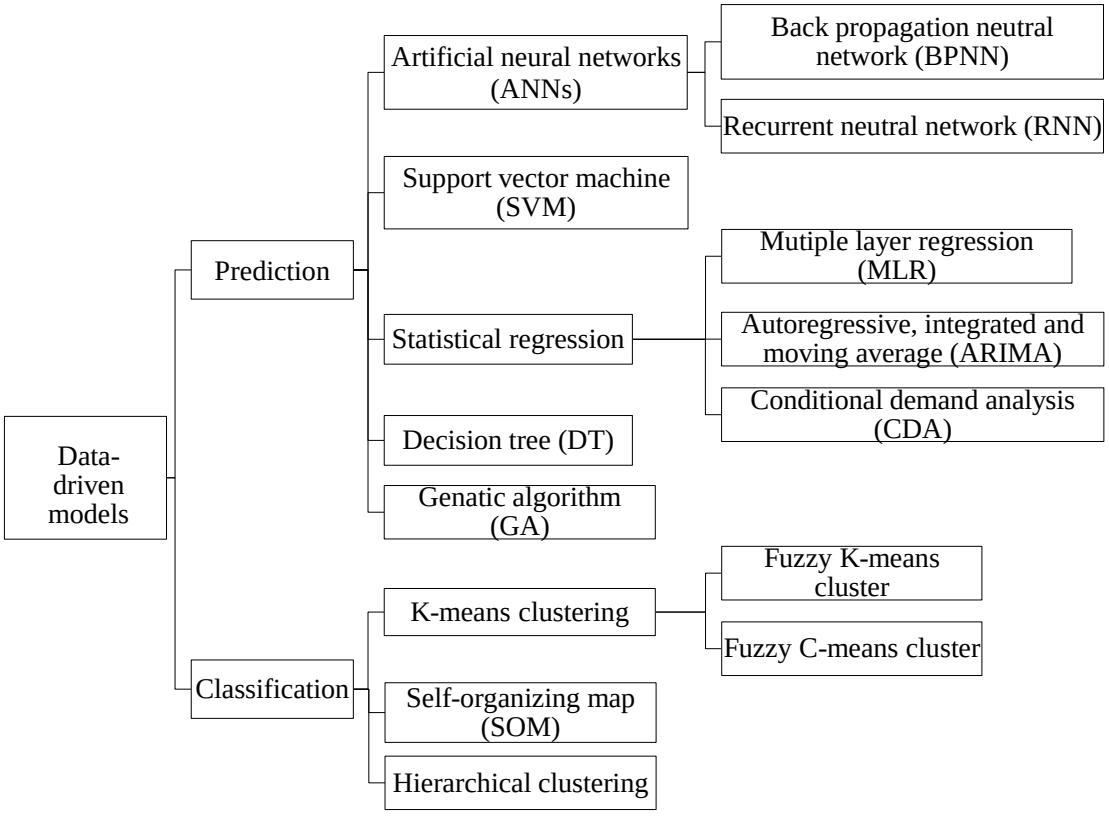


Fig. 2-1: Different data-driven models for building energy consumption

a. Artificial neural networks

ANNs are designed mimicking the basic architecture of human brain, whose basic element is called as processing unit modelling a biological neuron. The network consists of a large number of these process units arrayed in layers, and process units in different layers are connected with one another via connections, shown in **Fig. 2-2**.

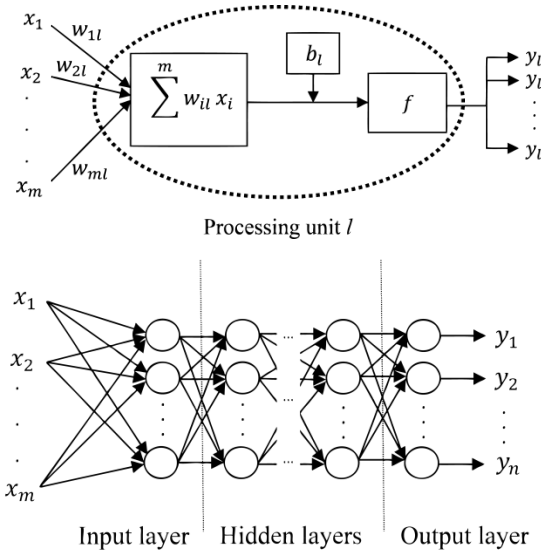


Fig. 2-2: Schematic of ANN

(a) a single process unit; (b) artificial neural networks

Each process unit, say l, deals with signals, x_{il} $(i = 1, 2, \dots\dots, m)$, from units connected with it in the other layers. These signals are input through the incoming connections with a weight w_{il} $(i = 1, 2, \dots\dots, m)$. The process unit then takes two basic operations on the input signals: summation and activation, and delivers an output y_l [2.4]

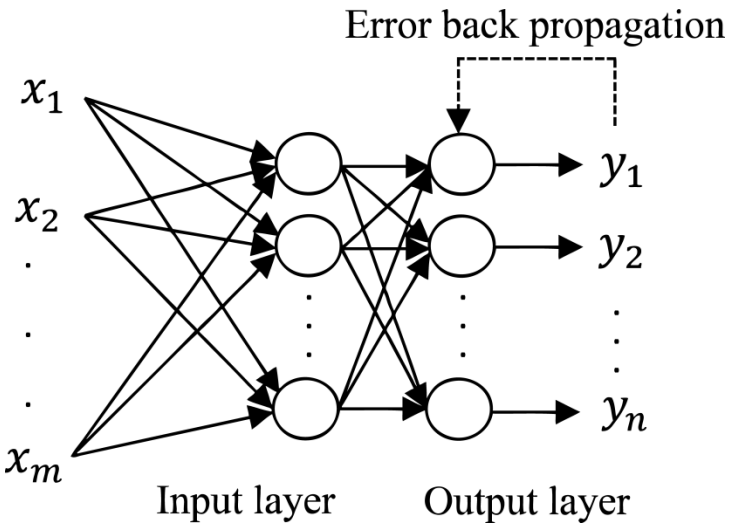
$$y_l = f(\sum_{i=1}^{m} w_{il}\, x_i + b_l) \tag{2.1}$$

where b_l is a bias set specifically for each process unit and f is the activation function, commonly defined as the sigmoid function [25]

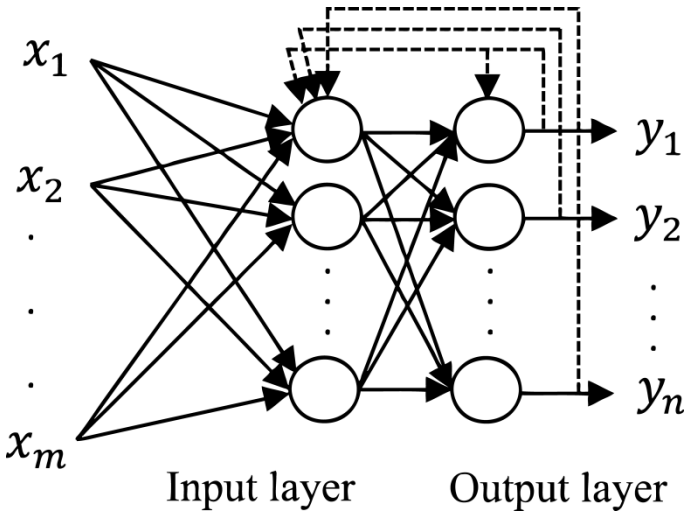
$$f(x) = \frac{1}{1+e^{-x}} \tag{2.2}$$

The output y_l will be used as an input signal for the process units in the next layer connecting to the process unit l.

As we discussed, all the process units in ANNs are arranged in a layer-structure and process units in different layers are interconnected based on a designed architecture. **Fig. 2-2 (b)** shows a simple example: feed-forward ANNs where process units are arrayed in the input, hidden and output layers and the information flows in one direction throughout these layers. In today's ANNs studies, ANNs models also take other architectures to more effectively approximate human brain activities. Two representative are back-propagation neutral network (BPNN) and recurrent neutral network (RNN), see **Fig. 2.3**. The former computes the error of output every time, and then propagates this information as a negative feedback to tune the incoming connection weight and bias. This manipulation offers flexibility to modify the output error to a minimum, and thus improving accuracy of ANN calculation. As to RNN, it involves the backward connections feeding back the outputs themselves as the inputs to the process units in the former-layer or even the current unit to capture tempore behaviours. Such a recurrent design makes RNN deal with time series datasets without random data, which leads it to being particularly welcome for sequence events [2.6].



(a)



(b)

Fig. 2-3: Schematic of (a) two-layer BPNN and (b) two-layer RNN

White cycles: process units in different layers. Solid arrows: connections; Dashed arrows: feedbacks

No matter what kind of network architecture is in use, an ANNs model must experience a training (learning) process to specify all needed connection weights and

biases before real applications. This training process will take advantage of available historical data records, which will used as benchmarks to cultivate the proper response of the ANNs model for given inputs. Therefore, ANNs are capable of learning the relationship among input signals, and capturing key information through a training process based on historical data records. On top of that, it also possesses a number of other advantages, such as fault tolerance, robustness and noise immunity. Thanks to these favourable features, ANNs have achieved great success in solving non-linear problems so far. On the other hand, meanwhile, it should be also pointed out that the architecture choice and learning-rate optimization in the current ANNs are still developed on an *ad hoc* base. This implies ANNs applications are usually case-dependent nonetheless. They have to be designed and validated for each time for different applications [2.5].

b. Support vector machine

Supported vector machine (SVM) is another popular artificial intelligent method [2.7], which deals with n data records, i.e., $\{(x_i, Y_i)\}_{i=1}^{n}$, with the input $x_i \in R^N$ and the target $Y_i \in R$. (Note that Y_i could also be in binary for some applications [2.8]). Nowadays, this method has been widely applied to solve regression problems to estimate an underlying relationship between the nonlinear inputs to the continuous real-valued target. The SVM used for regression is called as support vector regression (SVR), which has become a particularly important data-driven approach for predicting building energy consumption.

The core task in SVR is to construct a decision function, $F(x_i)$, by use of a training process based on historical data. It is required that for a given input x_i, the result estimated by this function should not deviate from the actual target Y_i larger than the predefined threshold ε. In SVR, such a function is usually assumed in the form of

$$F(x_i) = \langle w, \varphi(x_i) \rangle + b \tag{2.3}$$

where the bias $b \in R$. $\langle \cdot, \cdot \rangle$ and w represent the dot product and weight defined in R^N. $\varphi(x_i)$ is a non-linear mapping of the input space to a high-dimensional feature space [2.9]. w and b are two unknown in Eq. (3), and need to be estimated through minimizing the regularized risk function [2.9]. In SVM theory, the latter is easily solved in its dual formulation by an introduction of a Lagranian L [2.9],

$$L \coloneqq \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*)$$

$$-\sum_{i=1}^{n} a_i(\varepsilon + \xi_i - \mathrm{y}_i - \langle w, \varphi(x_i) \rangle - b)$$

$$-\sum_{i=1}^{n} a_i^*(\varepsilon + \xi_i^* - \mathrm{y}_i - \langle w, \varphi(x_i) \rangle - b) \tag{2.4}$$

where $\{a_i, a_i^*, \eta_i, \eta_i^* \geq 0\}$ are the Lagrange multiplier. $\|w\|$ is the Euclidean norm. $\{\xi_i, \xi_i^* \geq 0\}$ are two slack variables to copy with some infeasible optimization constraints. The constant $c > 0$ is defined to determine the trade-off between the training error (over-fitting) and model flatness (under-fitting). It should be noted that the Lagrange multipliers are all independent. They are $\eta_i = c - a_i$ and $\eta_i^* = c - a_i^*$, and $\{a_i, a_i^*\}$ can be determined by the corresponding dual optimization [2.9],

Maximize $W(a_i, a_i^*) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_i^*)(a_j - a_j^*)(\varphi(x_i)\cdot\varphi(x_j))$

$$+\sum_{j=1}^{n}(a_i - a_i^*)y_i - \varepsilon\sum_{j=1}^{n}(a_i + a_i^*)$$

subject to $\begin{cases} \sum_{j=1}^{n}(a_i - a_i^*) = 0 \\ a_i, a_i^* \in [0, c] \end{cases}$ (2.5)

With the computed a_i, a_i^*, the weight w can be written a function of $\{a_i, a_i^*, x_i\}_{i=1}^{n}$. This gives rise to the decision function in SVR

$$F(x) = \sum_{x_i \in SV}(a_i - a_i^*)K(x, x_i) + b$$

(2.6)

where $K(x, x_i) = \varphi(x)\cdot\varphi(x_i)$. In SVR, this is called as the kernel function, having different formulas for various applications in the literature, e.g., $K(x, x_i) = \exp(-\gamma\|x - x_i\|^2)$. It should be pointed out the sum in Eq. (1.6) does not cover all inputs. Instead, only those (i.e., support vectors $x_i \in SV$) corresponding to $(a_i - a_i^*) \neq 0$ are included. Moreover, the bias b in Eq. (6) is also computed by these support vectors

$$b = \frac{1}{N_1}\left\{\sum_{a_i \in (0,C)}\left[Y_i - \sum_{x_j \in SV}(a_j - a_j^*)K(x_i, x_j) - \varepsilon\right]\right.$$

$$\left.+\sum_{a_j^* \in (0,C)}\left[Y_i - \sum_{x_j \in SV}(a_j - a_j^*)K(x_i, x_j) + \varepsilon\right]\right\} \quad (2.7)$$

Here, N_1 is the number of support vectors with either $\{a_i \in (0, c), a_i^* = 0\}$ or $\{a_i = 0, a_i^* \in (0, c)\}$. Once the decision function is fully specified by the training dataset, the SVR model can be used as a predicting tool for a new input x.

It is worth emphasizing that the superiority of SVR, or more generally SVM, to other models are that its framework is easily generalized for different problems and it can obtain globally optimal solutions. Its capability of dealing with nonlinear relations by transferring them into high-dimensional linear problem is also impressive for practical applications. Nonetheless, the method is rather time-consuming for large-scale problems [2.8, 2.10]. Recently, immerse efforts has been paid to developing possible ways to optimize its computational efficiency.

c. Statistical regression

Prediction of building energy-consumption relies on a regression analysis to devise a relationship linking an output (i.e. response, Y_i, $i = 1,2 \dots n$) to the contributing inputs (i.e., predictors, $x_{i,j}$, $i = 1,2 \dots n, j = 1,2 \dots m$). In the previous section, we have discussed a regression process based on the SVM theory-SVR. On top of that, there still exist other regression models, e.g., statistical regression, used for predicting building energy consumption. Statistical regression investigates the relationship among different variables in a probabilistic framework, which formulate the output as

$$\text{Multiple:} \quad Y_i = \alpha_i + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_m x_{i,m} + \varepsilon_i \tag{2.8}$$

or

$$\text{Polynomial:} \quad Y_i = \tilde{\alpha}_i + \tilde{\beta}_1 x_{i,1} + \tilde{\beta}_2 x_{i,2}^2 + \cdots + \tilde{\beta}_m x_{i,m}^m + \varepsilon_i \tag{2.9}$$

where ε_i represents a random error assumed to be normally distributed, and α_i, $\tilde{\alpha}_i$, β_j and $\tilde{\beta}_j$ $(j = 1, \dots \dots m)$ are the parameters to be estimated. Note that both Eq. (2.6) and

(2.7) are linear with respect to these parameters whilst they are not necessarily linear with respect to the contributing predictors, as seen as Eq. (2.7). Like other data-driven approach for prediction, the statistical regression equations make use of the finite number of historical data to estimate the involved parameters. For demonstration, we choose the multiple linear regression Eq. (2.6) as an example, in which the estimates of all parameters will derived using the least squares (LS). To be specific, the sum of squared errors (SSE) is first defined

$$SSE = \sum_{i=1}^{n}(y_i - A_i - B_1 x_{i,1} - B_2 x_{i,2} - \cdots - B_m x_{i,m})_i^2 \quad (2.10)$$

In Eq. (2.8) A_i and $B_j(j = 1, \ldots \ldots m)$ are the corresponding LS estimates of α_i, $\beta_j(j = 1, \ldots \ldots m)$ in Eq. (2.9). *SSE* is then minimized which gives rise to $m + 1$ equations. Each of these equations includes one of partial derivatives of SSE with respect to A_i and B_j $(j = 1, \ldots \ldots m)$, to be set zero, respectively. It is these equations that are used to solve A_i and $B_j(j = 1, \ldots \ldots m)$ directly subject to the given historical dataset $\{ x_{i,j}, Y_i, i = 1,2 \ldots n,\ j = 1,2 \ldots m \}$. Finally, the prediction equation with the estimated parameters in multiple linear regression is specified as

$$y_i = A_i + B_1 x_{i,1} + B_2 x_{i,2} + \cdots + B_m x_{i,m} \quad (2.11)$$

In statistical regression, there is another variable introduced to quantify the goodness of fit of the regression line by Eq. (2.9), that is the coefficient of determination R^2,

$$R^2 = 1 - \frac{SSE}{SS_{tot}} \quad (2.12)$$

where $SS_{tot} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, with the mean value $\overline{Y} = \sum_{i=1}^{n} Y_i$. Generally, a regress equation with a larger R^2 indicates it can better fit the original data.

Based on the above discussion, it is seen that statistical regression is an easy-to-use approach for predicting building energy consumption. In particular, it was popular to predict average consumption over a long period in the early studies. However, the regress models require a large number of historical data for training, and the resulting accuracy of a short-term prediction is yet poorer than that of other data-driven approaches, such as ANN or SVM. It is also challenging for statistical regression to select a set of plausible predictors and an appropriate time scale to well fit energy consumption for buildings under a wide range of environment and weather conditions. Worse, the selected predictors in some cases may not be literally independent. The unforeseen correlations among them would result in uncertain inaccuracy in the regression outputs [2.11].

d. Decision Tree

Decision tree (DT) is a technique to partition data into groups using a tree-like flowchart. In this sense, a DT model manifests itself as a graph consisting of a root node and a couple of branch nodes. A DT starts from the root node where the input data are split into different groups based on some predictor variables predefined as splitting criteria. These split data are then disseminated to sub-nodes as branches emanating from the root node. The data on sub-nodes will undergo either further or no splits. The former are the internal nodes where the subsequent data split is conducted to form new subgroups as son-branches emanated graphically at the next level.

Whereas the latter are leaf nodes which treat the corresponding data group at the current level as their final outputs. **Fig. 2.4** illustrates a DT representation used for medium annual source energy consumption per unit floor (kWh/m^2/yr) of a commercial building. In this case, the gross floor area and building use ratio are chosen as predictor variables in the root node and internal node, respectively, and a mixture of data about energy consumption has been purified into a hierarchy of groups.

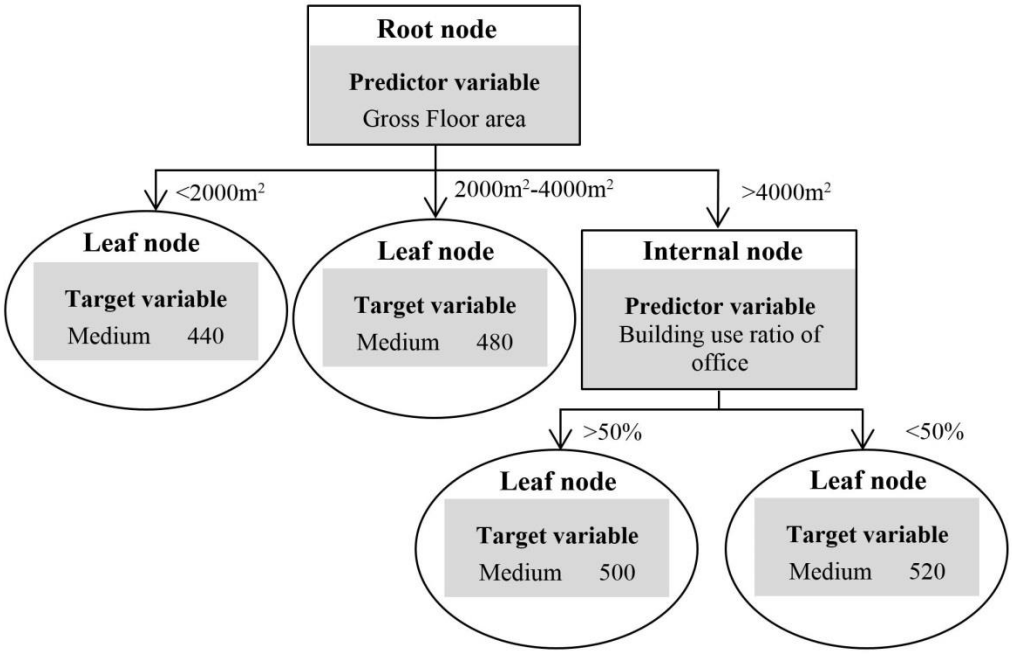


Fig. 2-4: Decision tree illustration of a medium annual source energy consumption per unit floor of a commercial building

Significantly, in a DT analysis the information entropy is an important concept used to quantify data group homogeneity. It is defined by

$$E = \sum_{i=1}^{n} -P_i \, log_2 P_i \tag{2.13}$$

where E is the information entropy. n and P_i are the number of different target values and the probability of a dataset taking the i^{th} target value, respectively. This entropy is used to calculate the information grain or gain ratio, based on which a DT structure linking the top root node to each branch node is specified. Readers can refer to Ref. [3.12] for detailed splitting procedure using the gain ratio or information gain.

In comparison to other data-driven approaches, DT's tree-like structure is easy to understand and its implementation does not involve complex computation knowledge. However, its deficiency is also evident—the targets used in a DT are primarily based on expectations. This usually leads to significant deviations of its predictions from the real results. The DT architecture is also a restriction so as the method is unable to deal well with time-series and nonlinear data.

e. Genetic Algorithms

Genetic algorithms (GAs) are stochastic optimization inspired by natural evolution based on the idea of "survival of the fittest" [3.13]. Many GAs in building energy prediction formulate three kinds of algebraic equations to compute the output (as solution) according to the given inputs:

- Linear: $y = w_1 x_1 + \cdots + w_m x_m,$ (2.14)

- Quadratic: $y = w_1^{(1)}x_1 + \cdots + w_m^{(1)}x_m + w_{1,2}x_1x_2 + \cdots + w_{1,m}x_1x_m + w_{2,3}x_2x_3 + \ \ldots + w_{m-1,m}x_{m-1}x_m + w_1^{(2)}x_1^2 + \cdots + w_m^{(2)}x_m^2,$

(2.15)

- Exponential: $y = w_0 + w_1{x_1}^{\widetilde{w}_1} + w_2{x_2}^{\widetilde{w}_2} + \cdots + w_m{x_m}^{\widetilde{w}_m},$

(2.16)

where $(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_m)$ are m independent inputs contributing to the output, y, and w_i, $w_i^{(1,2)}$ and $\widetilde{w}_i$ are the real-valued weights. In GAs, different sets of weights compose a search space where a point represents a feasible solution to the problem under investigation. The core task of a GA is to model an evolution process to identify the best among all feasible solutions in this space. In implementation, a GA first randomly chooses n sets of weight and encode each weight as a l bit binary string, e.g. $w_i = \overbrace{100\ldots01}^{l}$. In so doing, a set of weight is then represented as a chromosome $X_j = \overbrace{100\ldots01}^{w_1}\overbrace{000\ldots11}^{w_2}\ldots\overbrace{100\ldots10}^{w_m}$, and the n chromosomes form an initial population r. Importantly, every chromosome X_j in the population r is mapped to a fitness $h(X_j)$ (a real value) and assigned a probability P_j. In most cases, these two variables are defined by

$$h(X_j) = \left(y(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_m | X_j) - Y\right) \tag{2.17}$$

and

$$\mathrm{P}_j = \frac{h(X_j)}{\sum_r h(X_k)} \tag{2.18}$$

where Y is the targeted output from historical datasets and the Greek letter "Σ" denotes a sum of the fitness of all chromosomes in the population r. Next, pairs of chromosomes are selected as parents to reproduce the offspring (still chromosomes). Generally, the better fitness the chromosomes have, the more possible they are selected. The chosen parents then proceed crossover and mutation. One simple crossover operation is to randomly choose a crossover point and exchange the alleles up to this point of the two parent chromosomes. As to mutation, a few of bits in the chromosome after crossover, again chosen randomly, are switched between 0 and 1 (e.g. $100\underline{0}1 \rightarrow 100\underline{1}1$). Selection, crossover and mutation will be repeated to generate sufficient new offspring to form a new population, r', at the next level. It should be pointed out that the fitness of all offspring chromosomes in this new generated population will be computed and compared with the user's requirements. Generally, a GA will continue further runs of the above evolution process unless a chromosome (i.e., a set of weights) with satisfactory fitness is reproduced.

The aforementioned introduction of GAs indicates this method is a powerful optimization tool in dealing with complex multi-modal problems [3.14]. The algorithms can obtain suitable solutions based on either the objective functions or subjective judgements when large and sophisticated input data are given. Meanwhile, two major deficiencies in the current GAs are also noted—non-unique results and large computation time. In the literature, attempts to combine a GA with other data-driven approaches (e.g. ANN) have been made to mitigate the negative impacts arisen from the deficiencies.

f. K-means cluster

The K-means clustering algorithm is a classification approach quite popular in building load analysis. Technically, this algorithm partitions a set of data into a number of non-hierarchical groups of similar data points, i.e., clusters. The similarity among data points is quantified by the Euclidean distance, based on which a K-mean clustering procedure includes the following steps. A data set $(x_i, i = 1,2 \dots n)$ is first input with the cluster centers $(\mu_j, j = 1,2 \dots K)$ being specified randomly. The Euclidean distances between each data point and each cluster center are then computed. A datum x_i is set to belong to a cluster C_j if its distance to the cluster center μ_j is shorter than those to any other center. As a consequence, this classification forms K clusters in the input dataset, and the center of each cluster is re-calculated as a mean based on new data grouping. The K mean clustering algorithm will repeat the above distance computation, data classification and center relocate till all the K cluster centers do not move their locations with further iterations [2.4]. In many cases, a squared error function J is introduced to characterize this convergence,

$$J = \sum_{j=1}^{K} \sum_{x_i^{(j)} \in C_j} (x_i^{(j)} - \mu_j)^2 \tag{2.19}$$

where $x_i^{(j)}$ represents a data point belonging to the cluster C_j [3.15]. In the K mean clustering algorithm, *a priori* specifications of the cluster number K and initial positions of the cluster centers are required. This results in the algorithm has to be conducted several times in practice with these parameters with different values. Only the best results after comparison will be deemed as the algorithm's ultimate outcomes.

It is worth mentioning to improve its feasibility, the K-means clustering algorithm has been modified using the fuzzy methods. The modified version, in contrast to the aforementioned discussion, allows soft clustering, i.e., every data point can potentially belong to multiple clusters and a degree of membership is defined to characterize such relationships [3.16]. Ref. [1.13] discusses one widely-used fuzzing cluster approach in building energy projects, i.e., fuzzy C-means (FCM) cluster. Interested readers can refer to it for more details.

g. Self-organizing map

Self-organizing map (SOM) is developed from ANNs which transfers an incoming signal pattern in arbitrary dimensions into a one- or two- or multi-dimensional topographic map [2.4]. The method is trained by an unsupervised learning process and capable of classifying new inputs into clusters with different features in a neurobiological-like manner. **Fig. 2-5** illustrates a frequently-used network architecture of SOM consisting of a one-dimensional input layer and a two-dimensional computational layer. In this computational layer, a number of process units, i.e., neurons $(j = 1,2 \dots m)$, are arranged in rows and columns, each of which connects all input signals $(x_i, i = 1, 2 \dots n)$ with connection weights w_{ij}. The output of the neuron j is sometimes given by $y_j = \sum_{i=1}^{n} w_{ij} x_i$

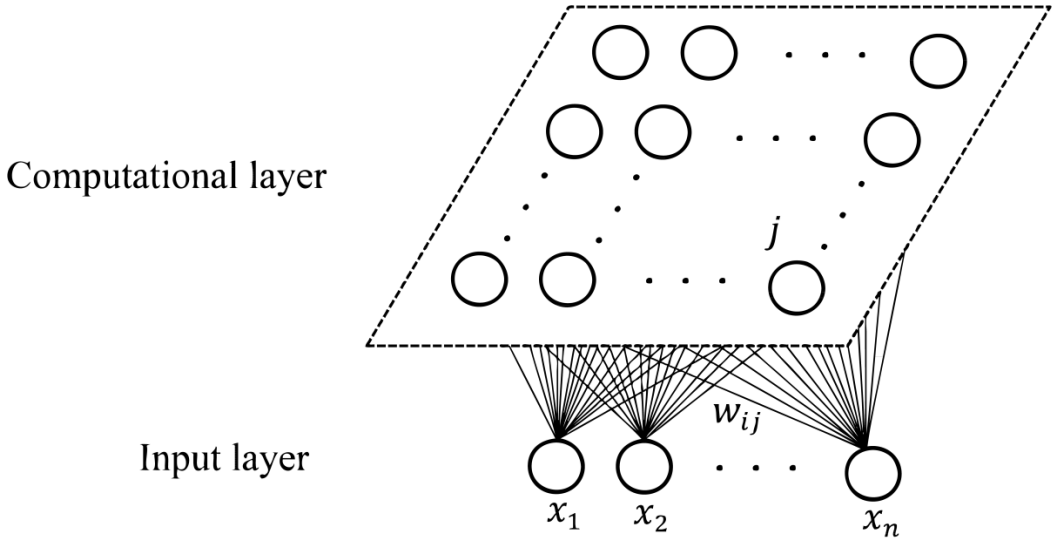


Fig. 2-5: Schematic of SOM

White cycles: process units; Solid lines: connections

In SOM, a squared Euclidean distance between all the input signals and connection weights pertinent to every neuron is computed

$$d_j = \sum_{i=1}^{n}(x_i - w_{ij})^2 \qquad (j = 1,2 \dots m) \tag{2.20}$$

This distance is termed as the discriminant function, and the neuron with the smallest discriminant function is designated as the winner for a given set of input signals. Typically, a SOM iteration starts from initializing all correction weights with small random numbers and choosing a set of input signals from historical database at random to form the input layer. Computation of the discriminant function for each neuron in the computational layer is then performed. Only the neuron with the smallest discriminant function is identified as the winner at this iterative level. Immediate to this, a topological neighborhood centered at the selected winner is

defined, in which the connection weights linking every neuron to the input signals are adjusted subject to

$$w_{ij}(n+1) = w_{ij}(n) + g_j\left[x_i - w_{ij}(n)\right] \tag{2.21}$$

where n represents the current iterative level, g_j is the learning rate depending on n and the distance between the winner and the neighboring neuron j. The next iteration at $n+1$ will be conducted with these adjusted correction weight and the new randomly-chosen input signals. Note that while the SOM iteration proceeds, both the learning rate and the size of the winner's neighborhood will decrease. The whole iteration will terminate once a threshold is met, e.g., $g_j \leq g_{j,min}$ or only the winner itself or none being included in the neighborhood. After training, a particular neuron (i.e., winner) in SOM will be activated the most for a particular type of input signals. This correspondence ensures SOM to be effective mean used for clustering new input signals.

In sum, SOM can effectively reduce the dimensions of a high-dimensional signal pattern to a feature map in which the similarities and differences among input objects are easily discerned. Moreover, its outputs can be directly followed by further classification using other clustering algorithms. This will lead to more mutually exclusive and well-separated groups. On the other hand, it is also noted that SOM clustering suffers from oscillation if a rambling dataset without any pretreatments is used as the input. Importantly, its computational cost will dramatically increase with the increasing dimension of the data. Therefore, a good SOM should be equipped with a well-designed tuning process and a clear parametric analysis on the impacts of

different parameters. These parameters usually include the learning rate, neighborhood function, number of process units, *et al*.

h. Hierarchical clustering

Hierarchical clustering in building energy consumption commonly uses the bottom-up fashion to organize data points into a tree-like hierarchy of clusters [3.13]. Such clustering is known as the agglomerative algorithm starting with n data points $(x_i, i = 1,2, \dots \dots n)$, each of which is treated as a singleton cluster. To characterize the inter-cluster similarity, the distances among different clusters are computed, and form a $n \times n$ matrix

$$H = \begin{bmatrix} 0 & \cdots & D(C_n, C_1) \\ \vdots & \ddots & \vdots \\ D(C_1, C_n) & \cdots & 0 \end{bmatrix}$$

In the above matrix, the distance between two clusters $D(C_i, C_j)$ is defined by

$$D(C_i, C_j) = \min d(x_i, x_j), \qquad \text{with } x_i \in C_i \text{ and } x_j \in C_j, \tag{2.22}$$

where $d(x_i, x_j)$ is the distance (i.e., Euclidean distance) between two data points in these two cluster and $D(C_i, C_j) = 0$ when $i = j$ [3.13]. In the literature, there are the other ways to define the distance between two clusters. Interested readers can refer to Ref. [3.17] for more details. After computing the inter-cluster distances, the next step is to merge two closest clusters having the minimal $D(C_i, C_j)$, and then update the corresponding distance matrix. This merging manipulation will proceed iteratively till all data points have been included in a single cluster.

In hierarchical clustering, merging can be conducted in different ways and terminated at different levels provided the similarity criterion requires. **Fig. 2-6** illustrates an example where two distinct sets of three clusters are obtained in different merging routes based on different merging criteria. In building energy consumption studies, hierarchical clustering has been proven that it can reveal the data internal structure and generate useful knowledge about energy consumption in a building [2.4].

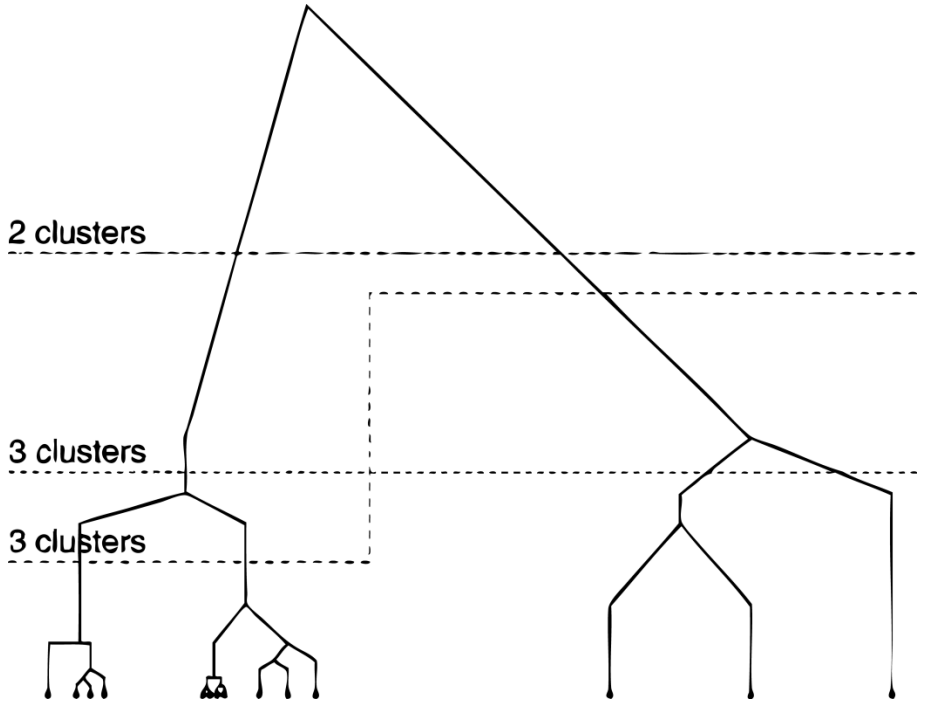


Fig. 2-6: Schematic of hierarchical clustering algorithm

The partitive clusters can be obtains at different levels of similarity [3.17].

2.1.2 Practical application of data-driven approaches

a. Load prediction

All the aforementioned data-driven approaches are widely applied to a large variety of prediction or classification applications of load prediction, energy pattern profile of specific use-cases, regional energy consumption mapping, energy benchmark for building stock, retrofit strategies and guideline making, see a summary in **Table 2-1**. This broad range of applications covers micro-scale and macro-scale studies that provide useful information and instructive suggestions for different stakeholders, including government, investors, engineers and occupants throughout the building life cycle from the early planning/design stage to later operation/retrofit stage.

Table 2-1: Summary of data-driven approach for applications in building energy consumption.

Data-driven approaches	Applications	
	Prediction	**Classification**
ANN	[2.5,2.6,2.20-2.29, 2.45, 2.46, 2.62, 2.64]	[2.57, 2.58]
SVM	[1.18,2.4,2.8,2.10,2.9,2.18, 2.29,2.30]	N/A
Regression	[1.18, 2.32-2.34, 2.44,2.45,2.48-2.50, 2.54, 2.55, 2.58]	N/A
DT	[2.35,2.36]	[2.60]
GA	[2.23,2.37-2.39, 2.63, 2.64]	N/A
K-means cluster	N/A	[1.13, 2.15, 2.40, 2.42, 2.43, 2.51, 2.53, 2.59, 2.61, 2.65]
SOM	N/A	[1.13, 2.15]

Hierarchical cluster	N/A	[2.40, 2.41, 2.61]

Originally, many data-driven approaches were established to predict the energy consumption of building, in particular electricity usage. It is well recognized that estimations of energy usage in the long-, medium- and short-term (i.e., annual, monthly and daily) are of importance for energy market planning and investments. Especially, a very short-term (hours or minutes ahead) estimation of electricity usage can exert a vital influence on the final dispatch for national electricity market [3.18]. Therefore, a precise prediction in these scenarios would lead to more efficient energy management and direct to considerable reduction in operational cost for both energy suppliers and end-users in buildings [3.18-3.20]. At the current stage, ANN and SVM are the two favourable data-driven approaches used for prediction of building energy consumption.

ANNs have been extensively used as a prediction means in diverse areas [2.5]. In building sector, ANNs excels in predicting building energy consumption, electricity demand, heating/cooling loads, important energy parameters and even assessment of software *etc*.. **Table 2-2** has centrally summarized these applications of ANNs in the literature.

In terms of energy consumption，ANNs are the popular candidate for both the short-term and long-term prediction. Kalogirou *et al*. [2.8] used ANNs to predict energy consumption in a holiday passive solar building, where engineers working in the HVAC field were not included. In their study, the RNN model based on the back-

propagation architect was applied for the training process. In so doing, such a model could detect features in the raw data of previous knowledge, e.g., the changing rules of operating conditions along different time epochs. In addition, Sözen and Arcaklioglu [2.21] even derived an ANN model to shed light on causality link behind economic indicators, population and net energy consumption. Their study suggested economic indicators (e.g. gross national product (GNP) and gross domestic product (GDP) *etc.*), rather than conventional energy indicators (e.g. gross generation, installed capacity and years), are playing a more important role for an accurate prediction of energy consumption.

As to electricity demand, the majority of ANN models focus on dynamic and short-term predictions, which require careful selection and pre-treatment of input data. One example is Ref. [2.22] where an on-line chiller electricity prediction model was established through use of both the simulated data and measured data. Their results recommended the sliding-window ANN, which constantly drops the oldest data and adds new measurements during training process, showed better performance than the accumulative ANN based on measured data. Besides, Karatasou *et al*. reported one-day ahead prediction of electricity consumption, called a 24-steps predictor, in Ref. [2.23]. The predictor used previous energy consumption data records with time delays larger than 24 hours as inputs to train the network to perform next day's prediction. Interestingly, An *et al*. [2.24] further developed an (EMD)-based signal filtering which is able to forecast half-hour electricity demand ahead. Such an EMD-based signal filtering can decompose an incoming signal into a series of pure modes and

residues. The results revealed that the EMD-based filter a critically-functioned component in the ANNs prediction model.

In fact, ANNs also play an important role in prediction of heating/cooling loads. In this particular type of applications, the ANN models usually require to input detailed climate information, envelop parameters and occupancy schedules [2.25, 2.26]. Besides reliable input data, algorithm optimization is the other way to promote the prediction accuracy. To minimize the drawback of BPNN (e.g. local optimization of model parameters in training process), a global optimization called "Modal Trimming Method" was proposed by Yokoyama *et al*. [2.27]. This method was composed of two steps, shown as **Fig. 2-7**: (1) search for local optimal solution of input variables in an objective function ($x_o{}^{fs} \rightarrow x_1{}^{lo}$), Normally, the objective function was defined as calculation error between predicted and measured values; (2) search for another feasible solution of the same objective function value with previous local optimization ($x_1{}^{lo} \rightarrow x_1{}^{fs}$). These two steps were repeated ($x_1{}^{fs} \rightarrow x_2{}^{lo} \rightarrow x_2{}^{fs}$) until tentative global optimal one $x_3{}^{lo}$ is found. They validated this method and concluded that significant error of predicted cooling demand from measured data was reduced compared to traditional local optimization method.

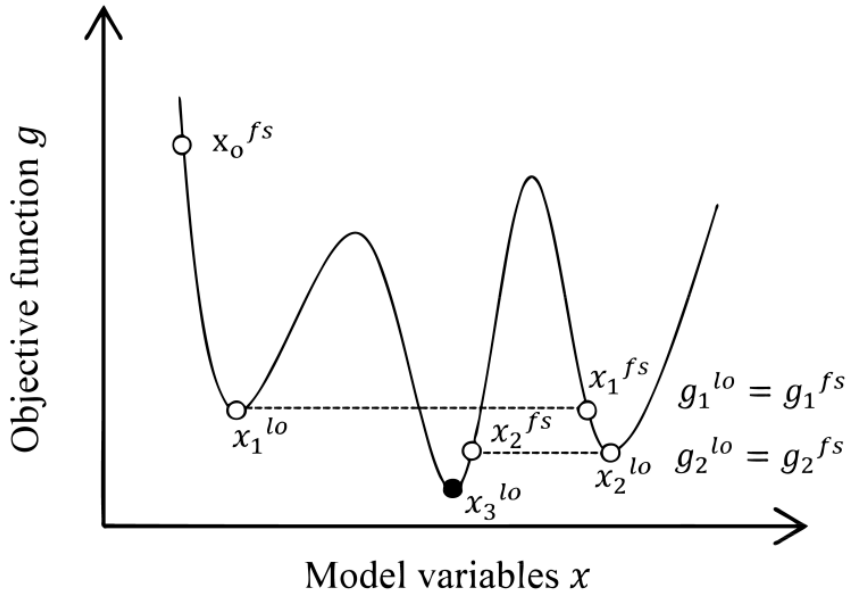


Fig. 2-7: Concept of modal trimming method [2.27]

On top of the above applications, ANNs' application is also extended to predicting the key parameters of energy performance of building. For instance, Olofsson *et al.* [2.28] proposed use of the BPNN model to estimate the total heat loss coefficient (HLC) and domestic energy gain factor of inhabited single-family buildings. Here, the total HLC characterizes heat loss resulted from transmission and air-flow while the domestic energy gain factor focuses on the gain of heating or cooling from inside sources. In this kind of ANN model, flag parameter of each measured case was introduced to distinguish non-linear dependences among various predictors, instead of average dependency from previous experience.

It is worth mentioning that ANNs are sometimes used as tools to assess simulation software for building performance. Neto *et al.* [2.27] compared the BPNN against EnergyPlus by using both to predict building energy consumption. The latter is recognized a mainstream simulator in building sector which can deliver much more

accurate results than Energy_10, Green Building Studio web tool, and eQuest [2.25]. Interestingly, Neto et al. found that when building and climate data were just briefly described, the used BPNN model works much better in daily energy demands prediction than EnergyPlus does. Importantly, especially for hourly prediction, all these simulation tools in current market give rather poor results in comparison to ANNs.This finding equips ANNs a new function as a benchmark to test accuracy of commercial software for estimating building energy performance.

Type of house	Scale	Inputs	Output	Data source	Measure length	Algorithm
Holiday passive house [2.6]	Single	Season, insulation function, wall thickness, heat transfer coefficient, time of day	Energy consumption	Measured data: ZigBee Input Device (ZID)	Two seasons	RNN combined with BPNN
Multiple [2.21]	National	Economic indicators (GNP and GDP), population	Net energy consumption	World Energy Council	37 years (1968-2005)	BPNN
Office building [2.22]	Single	Outdoor dry-bulb temperature, outdoor humility, water temperature of chiller, compressor status etc.	Dynamic chiller electric demand	Simulated data (DOE 2.1E) and measured data	1 year	Sliding window ANN and accumulative ANN
Office building [2.23]	Single	Previous load, temperatures of previous day, occupancy condition, sin and cosine of the hour	One day ahead electric power consumption	Great Building Energy Predictor Shootout I and measured data	1 year and a half	BPNN
Multiple [2.24]	Reginal	Previous electricity consumption	Half-hour ahead electricity demand	Australian Energy Market Operator	9 weeks	Multi-output BPNN

Residential [2.26]	7 buildings	18 building envelope parameters, heating degree day, cooling degree day	Heating and cooling energy consumption	Simulated data (DeST)	1 year	BPNN
Commercial [2.27]	Single	Previous cooling demand, air temperature and relative humidity	Cooling demand	Measured data	45 weekdays	BPNN
Residential [2.28]	7 single family-building	Supplied heating demand, electricity domestic demand, flag parameter	Indoor-outdoor temperature difference	Measured data	2 years	BPNN
Solar house [2.25]	Single	Outdoor temperature, relative humility, set point temperature, occupancy schedule	Heating/cooling consumption	Measured data	2 days	BPNN
Office building [2.20]	Single	Outdoor dry-bulb temperature, day type (working day or weekend)	Daily total consumption	Measured data: energy demand measurement system	54 days	BPNN

Table 2-2: Summary of ANNs in predicting building energy consumption

Prediction is also a primary function of SVM use in building energy simulation. **Table 2-3** lists the up-to-data studies on SVM-prediction applications. Generally speaking, SVM works in high accuracy in the medium-term [2.9] and short-term [2.18] prediction. Significantly, the method only requires a few model-parameters to implement its calculation. On the other hand, however, computing speed of SVM is slower than that of other approaches, such as linear regression and the ANNs. Currently, how to optimize SVM algorithm is regarded as the core task for its future development.

Many efforts have been actually made on SVM optimization in recent years. To save the computer memory and expedite the time-consuming training process, Zhao and Magoulès [2.8] proposed targeted solutions for dual optimization process (see Eq. (5)) and Kernel function calculation. The main idea was to divide the entire dual optimization problem into sub-problems and calculate them in parallel. Then, the Kernel function matrix would be updated for each sub-problem calculation. This parallelized training process could be stopped until convergence. The modified SVM gains a capability of dealing with a large amount of data to predict energy consumption of multiple buildings. Another possible optimization solution is to develop a hybrid SVM. For example, Li *et al*. [2.29] presented a hybrid approach combining SVM and FCM clustering algorithm to forecast building cooling loads. In this research, FCM was first employed to extract valid data records from the pool of raw data, and then the SVM followed with a training procedure based on the extracted valid data records. Clearly, such a pretreatment of data records effectively reduce the noise of inputs for SVM calculation. It should be pointed out the SVM is compatible

with diverse input information. Besides the conventionally-used energy loads and climate conditions, Paudel *et al*. [1.18] also used energy load characteristics and hidden inertial effects of building as their SVM inputs. The energy load characteristics were described as operation level of HVAC system and the occupancy profile, the hidden inertial effects were provided as fluctuation of internal temperature. The block diagram of such a model for predicting building energy consumption is shown in **Fig. 9**. As we can see, partial selection of input data called dynamic time warping (DTW) was adopted during prediction process, which measures on the outdoor temperature difference between training days and prediction days. The minimal difference between two time series was chosen as optimal path for solution. Similarly, the previous energy load database was also partially selected by DTW as inputs to consider the most recent data rather than whole data. The result showed that the designed training leads to higher accuracy and better computational efficiency in comparison to that based on the whole input data.

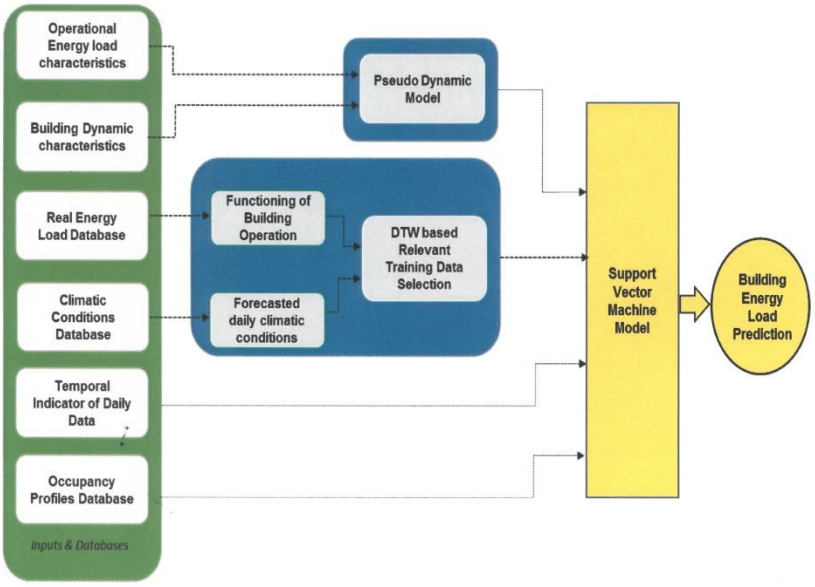


Fig. 2-8: Block diagram of SVM in prediction of energy demand using pseudo dynamic approach [1.18]

Statistical regression is well treated as a simple tool for prediction for a long time [1.18]. However, this approach suffers from low-accuracy in its prediction results, and such a deficiency has greatly limited its applications in building energy consumption analysis. This motivates a great deal of modification and optimization in statistical regression, which are briefly illustrated in **Table 2-4**. Among various modifications are multiple linear regression (MLR) proposed by Li and Huang [2.31] for short-term prediction. This model utilized not only climate data, room temperature set point, but also the cooling loads of previous four hours as its inputs. The obtained prediction results achieved very impressive accuracy higher than that of conventional ANN models. Moreover, autoregressive, integrated and moving average (ARIMA) model under the statistical regression framework was designed to correlate time-series data. Amjady's study [2.32] has well examined the exactitude of ARIMA model for

predicting daily peak and hourly load based on national power net. He further extended ARIMA model with use of the estimated electricity load as an extra input. The accuracy of his model reach a higher level even compared to original ARIMA and ANNs.

In most cases, statistical regression models were adopted to estimate important parameters characterizing energy performance. For instance, Mejri's *et al*. [2.33] investigated statistical regression modelling for predicting indoor air temperature. In their study, they analyzed the similarity in dynamic behaviours among different thermal zones for HVAC system design. Another example goes to Wauman's *et al*. [2.34], who used statistical regression to explore correlation between heat balance ratio and heat gain factor of some school buildings exemplified in their research. These obtained correlations are regarded of crucial significance for designing, tracing and analyzing building thermal behaviours. They are also important supportive materials for drafting heating control strategy for energy saving.

In the large family of data-driven approaches for building energy consumption prediction, DT is a relatively new member, but involves much simple techniques. Tso and Yau [2.35] compared statistic regression method, BPNN and DT by predicting the electricity consumption in summer and winter periods. Results show that DT used in their study performed as well as BPNN, both of which deliver accurate results than statistical regression did. Yu *et al*. [2.36] also applied the DT approach to predict energy use intensity (EUI) of residential buildings. They designed ten predictor variables concerning indoor temperature, building envelop, appliance types and occupant number in the DT framework. Their result clearly demonstrates that DT is

able to well predict building energy consumption level as high/medium/low. The significances of these predictor variables were ranked in terms of degree of closeness to the outdoor temperature (predictor variable of root node), which is the most important determinant of EUI. The results show that several building parameters, e.g. heat loss coefficient and equivalent leak area, deserve more attention at early design stage and benefit energy conservation in retrofit.

GA has been regarded as a powerful prediction approach in building energy consumption. As shown in **Table 2-4**, most applications of GA models are national analysis. One typical example is prediction model of energy consumption for residential-commercial building section in Ref. [2.23]. Three different scenarios were proposed in order to find out the best fit solution. The result showed that GA model, which considers residential housing production, house appliances of washing machine, television, vacuum cleaner and refrigerator as the input parameters, can obtain the most accurate quadratic prediction model of energy consumption. Sadeghi et al. [2.37] developed prediction model of electricity consumption using GA on national level. It was found that exponential equation had the more accurate results compared to linear and quadratic forms.

Hybrid methods of GA and ANNs are widely used in electricity prediction application [2.38]. K. Li and H. Su [2.39] predicted the daily air-conditioning consumption by using the genetic algorithm-hierarchical adaptive network-based fuzzy inference system (GA-HANFIS). Before developing prediction model, clustering algorithm was applied to identify the nature groups and qualities of a large data set, and GA was used to optimize the unknown cluster-parameters through minimizing the error of

predicting result. **Fig. 2-9** shows the architecture of GA-HANFIS, in which the outdoor temperature of predicted day $T\ (k)$, the air-conditioning consumption of past two days $y\ (k-1)$ and $y\ (k-2)$ were identified as more significant inputs of network layer 1. These less-significant variables, i.e., $T\ (k-1)$, $y\ (k-3)$, $T\ (k-2)$ and $T\ (k-3)$, were selected as inputs of network layer 2 and layer 3. Output $y\ (k)$ was air-conditioning consumption of predicted day. The rule base of each layer contained two if-then rules; readers can refer to Ref. [2.40] for more details. Moreover, the calculation rules were different according to different clusters. This hybrid method outperformed regular BPNN in prediction accuracy.

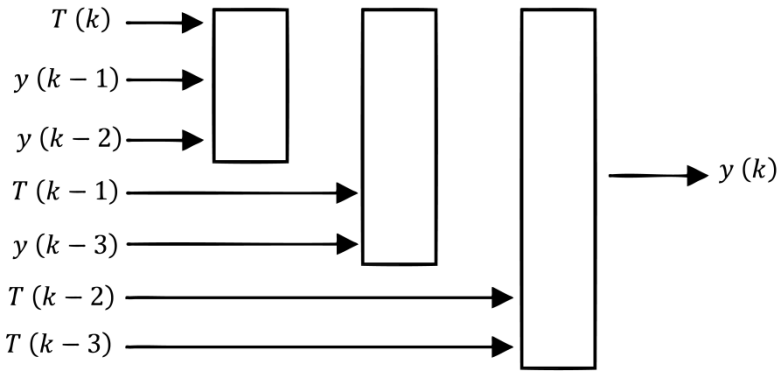


Fig. 2-9: Architecture of GA-HANFIS model with 3 layers [2.39]

Table 2-3: Summary of SVM in predicting building energy consumption

Type of house	Scale	Inputs	output	Data source	Measure length	Algorithm
Commercial [2.9]	4 single buildings	• Outdoor temperature, relative humility, global solar radiation, previous electricity consumption	• Building energy consumption per month	• Survey: monthly utility bill, National Environment Agency	3 years	SVM
Multiple [2.18]	Regional	• Historical electricity consumption data	• 5-minute ahead electricity load	• Australian electricity operator	3 years	SVM, statistical regression, and BPNN
Office buildings[2.8]	100 buildings	• Heating consumption, electrical consumption	• Heating demand, electrical load	• Simulated data (EnergyPlus)	5 months	Parallel SVM
Campus building [2.29]	Single	• Cooling load	• Cooling load	• Measured data	4 months	Fuzzy SVM combined FCM clustering
Office buildings [1.18]	Single	• Previous energy load, building dynamic characteristics, outdoor temperature, occupancy schedule	• Building energy demand	• Measured data: data acquisition system	7 months	SVM with pseudo dynamic approach
Office buildings [2.30]	Single	• Previous cooling load, air temperature, relative humidity, solar radiation intensity	• hourly cooling load	• Simulated data (DeST)	Half year	SVM, BPNN

Table 2-4: Summary of statistic regression, DT and GA in predicting building energy consumption

Type of house	Scale	Inputs	output	Data source	Measure length	Algorithm
Office building [2.31]	Single	• Dry bulb outdoor air temperature, solar horizontal radiation, and room temperature set point, cooling load of previous 4h	• Cooling load	• Simulated data from TRNSYS	• 60 measured case	MLR, ANN, grey-box approach
Multiple [2.32]	National	• Previous load, estimated current load, temperature	• Hourly electricity load and daily peak	• National dispatching center	• 1 year	ARIMA
Office building [2.33]	Single	• Indoor temperature of four rooms	• Room temperature	• Measured data	• 2 months	Statistical regression
Multiple [2.35]	groups	• Power rating of appliance, consumption time	• average weekly electricity consumption	• Survey	• Two seasons	BPNN, Least-squares regression, DT
Residential [2.36]	80 buildings	• Outdoor temperature, building characteristics, appliance energy source and usage (10 inputs)	• energy use intensity	• Survey and research committee	• 3 years	DT
residential - commercial [2.23]	National	• GDP, population, import, export, house production, basic house appliance consumption figures	• Future energy demand (2003-2030)	• World Energy Council and State Statistics Institute	• 8 years	GA

Residential [2.37]	National	• GDP, real price of electricity and natural gas in residential sector	• Future per-capita consumption of electricity (2009-2025)	• Iran Statistics Center, Central Bank of Iran	• 39 years	GA
Hotel [2.39]	Single	• Outdoor temperature of past 2 days, air conditioning consumption of past 3days	• daily air conditioning consumption	• Measured	• 7 months	GA-HANFIS

b. Energy pattern profile

The energy consumption profile in building is to quantify the total consumption contribution to sub-components, or further distinguish the usage characteristics. Regarding the positive influence for end-users, the capability of profiling the energy use as the feedback can educate the occupants on how to consume and change the consumption behaviours to certain extent. As for utility companies, DSM measures are implied after extracting load profiles in order to reach a proper load-shape objective, i.e. "peak clipping", "valley filling", "strategic conservation", "flexible load shape", "load building" and "load shifting". The commonly-used methods for energy and electricity profiling are clustering based method, which is detailed in **Table 2-5.**

Table 2-5: Summary of data-driven approaches in building energy consumption profiling.

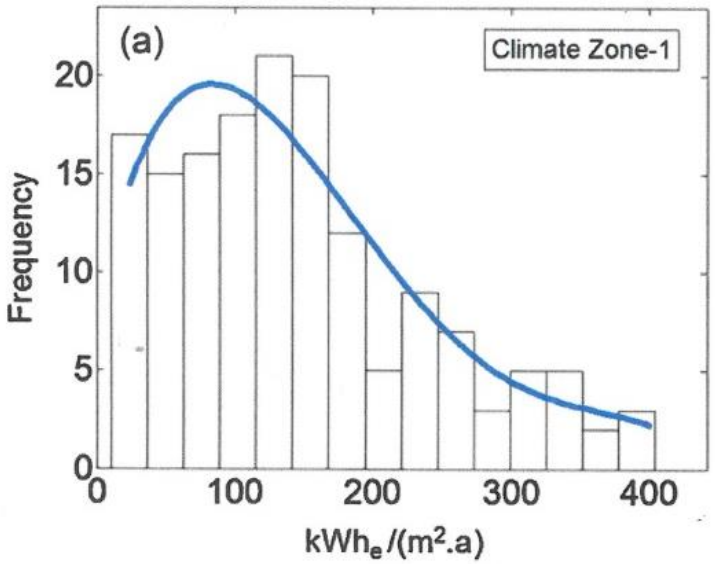
Type of house	Scale	Inputs	Data source	Measure length	Algorithm
Multiple [2.40]	94 buildings	Daily electricity consumption	Meters	10 months	K-means cluster, fuzzy K-means cluster, seven hierarchical cluster
University building [2.15]	27 buildings	Daily electricity consumption	Meters	2 years	SOM combined with K means ++
National office buildings [2.41]	24 provinces	Annual electricity consumption	Survey	1 year	Hierarchical cluster
High performance	134 buildings	Energy end use	Simulated data	3 year	K-means cluster

buildings [2.42]					
Campus buildings [2.43]	85 buildings	Heating demand	School manage service	5 years	K-means cluster combined with MLR
Residential buildings [2.44]	791 customers	Weather data and energy consumption of residential appliances	Survey	2 years	Statistical regression
Residential buildings [2.45]	8767 customers	Weather data and energy consumption of residential appliances	Survey	1 years	Statistical regression and ANN
Residential buildings [2.46]	8767 customers	Appliance, lighting, cooling loads, space heating, domestic heat water	Survey	1 year	PBNN

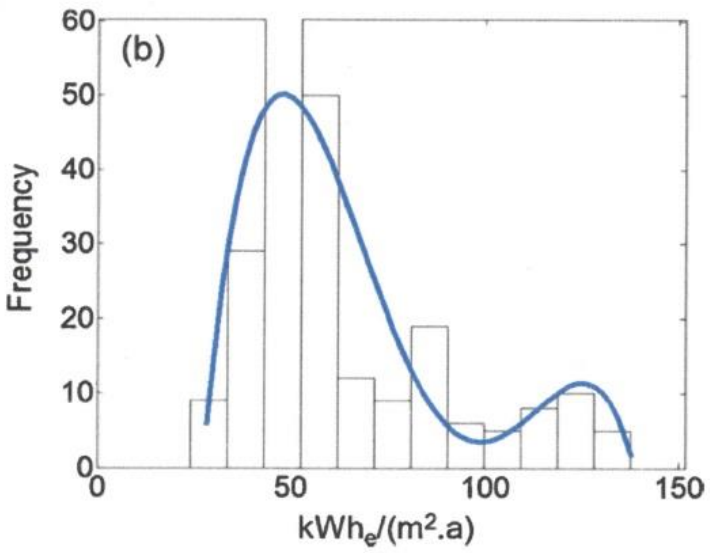
As one application of cluster method, analyzing electricity behaviour through pattern recognition and load curve classification has been investigated by massive researches. Tsekouras *et al*. [2.40] developed a two-stage pattern recognition for customer's classification. The first stage was to pattern load curves of each customer; the second stage was to cluster the customers according to pattern features. In their research, K-means cluster was proven by adequacy measures as the most appropriate approach compared to other methods. The function of adequacy measures is to evaluate the within-group similarity and between-group dissimilarity, in order to obtain a well-separated classification. Panapakidis *et al*. [2.15] incorporated K-means ++ cluster within SOM to reduce the number of centers and increase the accuracy. The data

records including vast of load curves were aggregated from various buildings, SOM was thus an appropriate approach to map high-dimensional database into low-dimensional patterns. As the improvement of the basic K-means clustering algorithm, K-means ++ algorithm tries to initialize the centroids that far from each other rather than random selection. The combination of SOM and K-means++ resulted in small errors in all cases.

Recently, cluster method becomes prevailing to profile EUI of buildings on large-scale. Xiao *et al*. [2.41] conducted a study on EUI (excluding district heating) of business office buildings in China. Each data point was defined as (x_{1i}, x_{2i}) in which x_{1i} and x_{2i} refer to EUI and gross floor area of corresponding building. Eventually, two clusters were formed by using hierarchical cluster and the frequency distribution of EUI is illustrated in **Fig. 2-10**. The cluster results revealed the unique "dual section distribution" pattern which is different from developed countries. Heidarinejad *et al*. [2.42] used K-means cluster algorithm to classify the EUI of 134 U.S. high-performance buildings (HPBs) by the squared Euclidean distance. These HPBs were well separated into three clusters, as high/medium/low EUI. Studies showed that unregulated loads which include various equipment and uncategorized loads, accounted for 30%-40% total energy consumption that should be reduced specifically through effective programs and modification. It can be found out that studies mentioned above that analyze building energy issues on large scale, are greatly dependent on the clustering methodology.



(a) US climate zone-1



(b) Certain city of China (excluding district heating)

Fig. 2-10: Frequency distribution and polynomial fitting plot of EIU in office buildings [2.41]

Clustering technology can be also applied for heating/cooling demand classification. K-means cluster analysis combined with MLR were proposed by Arambula *et al*. [2.41] to analyze the heating demand of 85 high schools. In their model, MLR analysis was firstly conducted to select 6 significant building thermal indicators according to R^2 value. Three clusters were developed by K-means cluster analysis based on these 6 indicators, while later R^2 was calculated for each cluster. The regression analysis showed that cluster 3 need to be further divided by clustering analysis since its low within-group similarity ($R^2 < 0.5$). Finally, more reasonable classification results could be obtained after such twice MLR analysis and twice clustering analysis when comparing to the sole clustering.

One regression method specialized for profiling energy consumption of residential buildings is conditional demand analysis (CDA). The basic idea of the CDA model is that total household consumption is the sum of various end-use consumptions.

CDA is frequently used to profile building energy consumption at national level [2.44]. Aydinalp-Koksal and Ugursal [2.45] used CDA to profile residential end-use energy consumption at national level, large-scale database including the surveys from occupants, weather conditions as well as historical energy bills were used. Their CDA model adopts 6 electricity end-uses including main and supplementary space heating, domestic heating water, space cooling, lighting, major and minor appliances. Meanwhile, they also developed neutral network model for comparison purpose [2.46]. In their research, BPNN outperformed CDA model in evaluating the effects of socio-economic factors, such as income, dwelling ownership and area sizes of residence. Because these socio-economic factors were considered as input variables in

BPNN while CDA cannot not include comprehensive variables due to the limitation of statistical regression.

c. Regional energy consumption mapping

Energy mapping methods, usually based on the Geographic Information System (GIS) city building database, consider using data-driven technology for pre-and post-progressive operation [2.47]. Thanks to the capabilities of GIS, immediate updating of energy evaluation and visual representation via maps are both permitted in a user-friendly model, to provide energy consumption distribution within the city. Among the massive technologies for energy mapping, statistical regression (MLR) and clustering algorithm are the mostly utilized data-driven methods, as displayed in **Table 2-6**.

Table 2-6: Summary of data-driven approaches in energy mapping

Type of house	Scale	Energy consumption Inputs	Data source	Measure length	Algorithm
Multiple [2.48]	45 cities	Annual electricity, population density, age of inhabitants, weather condition, living expenditure	Organizations involved in city affairs	1 year	MLR
Multiple [2.50]	City	Annual electricity consumption, natural gas, steam, and fuel oil consumption	Utility company, RECS, CBECS, geo-rectified database	1 year	MLR
Residential buildings	City	Yearly consumption of natural gas and	GIS database, Royal Netherlands	Nearly 50	MLR

[2.49]		electricity	Meteorological Institute, metering	years	
Residential buildings [2.51]	City	energy consumption and carbon dioxide emission	GIS database, electricity map, surveys,	18 months	Clustering algorithm
Multiple [2.53]	City	Electricity load, thermal loads,	GIS database, simulated and measured data, building standards	1 year	K-means cluster

MLR is a traditional used approach in energy mapping of building section at zip-code level [2.48]. Mastrucci *et al*. [2.49] applied MLR model to map the energy consumption of dwellings in a city of Dutch. The contributing inputs included floor area, number of occupants and type of house defined for each combination of type of dwelling and period of construction during 50 years. The predicted natural gas consumption was apportioned into space heating, domestic hot water and cooking. The results indicate that space heating is the biggest contribution (average 50%) in energy consumption. After 50 years tracking from 1965, they found the percentage of energy reduction is nearly zero for dwellings after 2005. Besides single building function of residential family, energy mapping model has been expended to profile the building energy consumption of multiple functions. For example, Howard *et al*. [2.50] calculated the annual EUI in New York City through MLR analysis, both the tax lot designations and building area categories were used to place the buildings into *n* building functions (e.g. residential family, office, warehouse, education and *et al.*).

The MLR analysis is explained by using Eq. (2.16), y_i is the energy consumption of i^{th} zip-code, x_{in} is the total building area of each building function in i^{th} zip-code. β_i the coefficient need to be determined in MLR. On top of that, it was found extra contributing inputs are needed in some regions to distinct the unique characteristics of energy consumption. However, the research excluded energy consumption for cooking, electrical heating and other end-uses, which inevitably causes errors in energy mapping.

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} \tag{2.23}$$

Clustering algorithm is typically used as subsidiary approach for mapping the energy consumption at urban scale. In Jones et al. [2.51] research, cluster analysis technique was adopted to classify 55,000 dwellings with similar energy consumption and carbon dioxide emission in a Local Authority of UK. The energy rating results and carbon dioxide emission results were profiled on the regional map for further retrofit purpose. Clustering algorithm is not limited to classification of energy consumption, also utilized to develop geographical clusters. For instance, Yamaguchi *et al.* [2.52] proposed a district clustering model for commercial buildings in Osaka city. Firstly, clustering of district were presented by small grid cells, each of them was classified into certain representative building-type category. Then, EUI was used as evaluation for the district typology. Fonseca *et al.* [2.53] proposed a model for mapping the spatiotemporal building energy consumption in a city district of Switzerland. The model involved K-means cluster for spatial grouping in the band of 50-200m, where spatial association of every variable of interesting buildings was strongly persistent (e.g. infrastructure types and temperature requirements). Two significant variables

were used to measure the intensity of spatial clusters and similarity of groups. GIS framework gathered overall results and enabled 4D visualization that provides understandable display. The peak space heating demand of buildings in four zones at 10-11am (April 1st, 2010) is presented as **Fig. 2-11**. The height and color code of buildings represent the demand level in relation to their associated zones.

Fig. 2-11: Spatio-temporal energy map of space heating demand of a city district in Switzerland[2.53]

d. Energy benchmark for building stock

Different from individual building energy analysis, benchmarking was used to address large-scale building energy related issues. Two fundamental issues in benchmarking are: 1) ascertaining the current energy performance of certain building (good, average or poor) compared to same types of building stock; 2) identifying the previous/current energy performance for energy saving potential and retrofit changes [2.54].

Regression-based model, ANNs, cluster algorithms and DT are the typical data-driven techniques for building energy benchmarking. **Table 2-7** shows the benchmarking pilots that usually adopt EUI as the single benchmarking index.

Table 2-7: Summary of data-driven approach in building energy benchmarking

Type of house	Scale	Benchmarking variables	Data source	Measure length	Algorithm
Supermarkets [2.55]	30 buildings	• Building age, occupancy condition, indoor temperature, energy system type and *et al.*	Survey	45 year	MLR
mix-used buildings [2.57]	60 buildings	• plug load, lighting, HVAC	Questionnaire	1 year	BPNN
Commercial buildings [2.58]	National	• building-operation hours, age category, building-area, cooling category, lighting category, CDD, number of floors category	CBECS	1 year	ANNs, MLR
School buildings [2.59]	340 buildings	• Heating demand, electricity demand, total energy consumption	Energy bills	3 years	Fuzzy cluster
office buildings [1.13]	30,000 buildings	• Heating load, cooling load, thermal comfort	Simulated data: VBD	1 year	K-means cluster, SOM, FCM cluster
Residential buildings	324 buildin	• EUI, CDD efficiency, HDD efficiency, bath room oriented, total	Panel dataset	3 years	TOPSIS, PCA, K-

[2.61]	gs	room oriented efficiency and etc.			means cluster
Commercial buildings [81]	1072 buildings	• EUI, gross floor area, building use ratio	Official building register, Korea Appraisal Board	3 years	DT

Regression technique is one popular method in building energy performance benchmarking. Chung *et al*. [2.55] benchmarked the EUI of 30 supermarkets in Hong Kong. MLR model was established to calculate EUI based on nine significant variables. By using bootstrapping function [2.56] for the empirical sample $\{EUI_{(1)}, EUI_{(2)} \dots EUI_{(30)}\}$, they obtained the estimation of EUI cumulative distribution as percentiles $\{EUI_{10}, EUI_{20} \dots EUI_{90}\}$. Although conducting on small-scale samples, they formed a benchmarking table through the percentiles. The results show that average value of energy consumption is greater than UK energy benchmarking. They also raised the suggestions that only unmanageable factors (e.g. building thermal characteristics) should be considered during benchmarking process while all manageable variables (e.g. occupancy behaviour) were set into average values, in order to present clearer improvement suggestions for government.

ANN method in energy benchmarking was initially presented by Yalcintas [2.56]. He developed three sub-models to predict EUIs as output for the plug load, lighting and HVAC components over 60 mix-used buildings. The information from questionnaire includes lighting types, floor area, equipment types and hours were used as inputs.

The elaborated ANNs model could be used to identify the EUI if new data was entered. The most outstanding advantage of ANNs benchmarking method is to renew the algorithm itself rather than manual update. Yalcintas [2.58] also developed a national energy benchmarking model for commercial buildings based on ANNs. Different from abovementioned ANN model which included continuous value of inputs/output, both input variables and output EUI were standardized into categorical forms for classification purpose in this model. In order to avoid inappropriate benchmarking results, database was firstly divided into 9 geographic regions. The results show that ANN model yields more accurate EUI estimation and reasonable benchmarking result than MLR model in all cases except one.

Fuzzy cluster algorithm is a frequently-used methodology for energy benchmarking for buildings. Santamouris [2.59] *et al.* proposed an energy rating system for 340 schools based on fuzzy clustering technology. Five classes of the total and thermal energy consumption had been defined. Compared to frequency distribution rating system, fuzzy clustering rating system is more reasonable to avoid unbalanced classification, such as either too small or too large range. Apart from building energy consumption benchmarking, thermal comfort rating system was also proposed by Nikolaou *et al.* [1.13] based on FCM cluster. The predicted mean vote index, which represents mean response about thermal comfort from a larger group of people, was used as thermal comfort indicator. In their study, the thermal comfort of each climate zone was classified as three clusters, respectively. The majority commercial buildings in Greek were belong to class 2, while “best practice office buildings” were belong to class 1.

Another energy benchmarking method for improving energy efficiency of office building is DT. Park et al. [2.60] developed DT model to benchmark the energy consumption of 1072 office buildings in South Korea. Gross floor area and building use ratio were identified as two significant predictor variables by correlation analysis, source EUI was defined as target variable. As result, six rating groups of EUI were developed for each type of building use. After establishment of benchmarking model, analysis of variance was utilized to test the difference among groups. DT model was believed to improve the conventional baseline benchmarking system via a more reasonable and fair classification.

Although most benchmarking projects are developed based on single EUI indicator, there is much effort for the development of multi-criteria benchmarking indicators. Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) based energy efficiency benchmarking approach using seven indicators was developed by Wang *et al*. [2.61]. The illustration of TOPSIS of two indicators (energy use per occupant and EUI) is showed as **Fig. 2-12**. When building A and building B have the same deviations to the most energy efficient condition I_P, deviations to the least energy efficient condition I_N were used to evaluate A and B. Because MLR cannot easily produce reliable weights among highly correlated indicator, principle component analysis (PCA) was adopted to weight the importance of seven energy indicators. PCA can transform a high-dimensional dataset consisting of possibly correlated variables into a less number of their linear combinations. Finally, K-means cluster was adopted to classify the TOPSIS space into six categories as benchmarking

table. Without a doubt, the benefits were obvious compared to single-criteria benchmarking which is observed with collision during evaluation process.

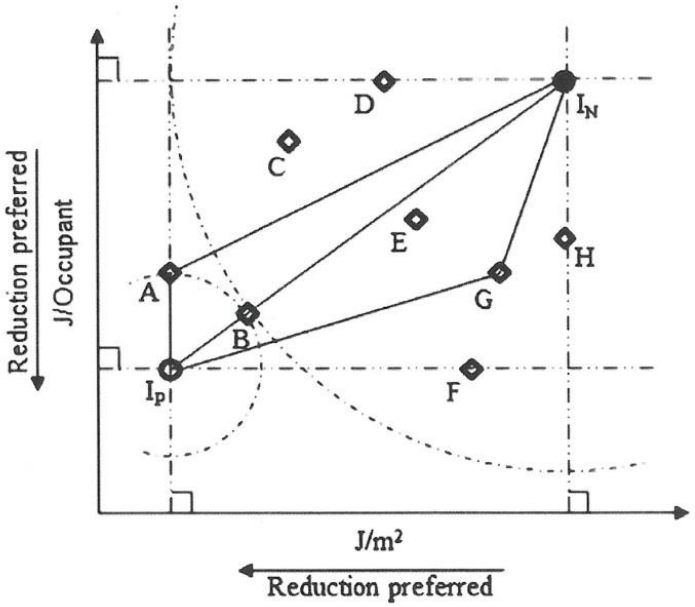


Fig. 2-12: Illustrative example of TOPSIS for building energy benchmarking [2.61]

e. Retrofit strategies and guideline making

Retrofit is based on the knowledge of energy profiling and benchmarking on existing buildings, presenting the largest potential of incorporation of renewable energy technology and energy conservation after efficiency retrofit measures. ANNs and GA are the main data-driven approaches in building retrofit projects, within a brief introduction in **Table 2-8.**

Table 2-8: Summary of data-driven approach in building retrofit

Type of	Retrofit measures	Algorithm

house		
Hotels [2.62]	Install energy management systems and Variable Frequency Drives (VFDs) on the air-handling units. New cooling towers and VFDs on motor fans.	BPNN
Residential buildings [2.63]	More than twenty retrofit measures under six main criterion, including safety, usage, convenient, comfortable, utility and health.	GA
School [2.64]	External wall insulation materials, roof insulation materials, the windows type, solar collector type, the HVAC systems.	ANNs and GA
Residential buildings [2.65]	Lower the carbon dioxide emission of grid, renewable resources, improve energy efficiency, and change occupancy behaviour.	K-means cluster algorithm

ANNs are usually applied to predict energy-saving potential for single retrofit project. Yalcintas et al. [2.62] developed BPNN model for two hotel equipment-retrofit projects. Energy usage data, weather data and occupancy data of post-retrofit period were used to train the neutral network model. It then estimated the energy consumption of pre-retrofit equipment as output. The difference between recorded and predicted energy consumption was regarded as the energy saving.

As a powerful optimization algorithm, GA has been frequently adopted as the evaluation tool in building retrofit project. Juan et al. [2.63] presented a GA-based on-line decision support system to offer residents a series of optimal refurbishment actions considering two objectives, cost and quality. In GA, each chromosome represented a set of retrofit solutions, the distance between chromosome and trade-off curve of cost and quality was used as fitness function to select the parents for

generation. With the process of evolution, the trade-off curve would gradually converge to the best retrofit solutions with higher quality and acceptable cost.

Developed based on two-objective optimization, multi-objective optimization model was conducted by Asadi *et al*. [2.64]. They adopted GA associated ANNs to study the interaction among three main conflicting target variables, including energy consumption (EC), retrofit cost (EC), thermal discomfort hours (TDH) and assess their trade-offs in school retrofit project. First, the database was created in simulation tool for training and validating ANN model. BPNN model adopted in this study was composed of input layer representing different retrofit measures, one hidden layers and one output layer of energy consumption and thermal discomfort indicator. Then, the GA tool was used for minimize these three target variables (as Eq. 2.24) and provide optimal combinations of retrofit measures.

$$\mathrm{Min}\, y_1 = EC(X) \tag{2.24a}$$

$$\mathrm{Min}\, \mathrm{y}_2 = \mathrm{RC(X)} \tag{2.24b}$$

$$\mathrm{Min}\, \mathrm{y}_3 = \mathrm{TDH(X)} \tag{2.24c}$$

$$\mathrm{X} = \{\mathrm{xWALL}, \mathrm{xROOF}, \mathrm{xWINDOW}, \mathrm{xCOLLECTOR}, \mathrm{xHVAC}\}$$

Where x represent different materials/types of alternative retrofit choices. The treat-off curves of multi-objective optimization could be available on 3D visualization. The proposed approach presented variety of recommendations with high computation efficiency. However, simultaneous optimizations of conflicting variables gave large

diversity of retrofit choices, which are difficult to understand the impact of each retrofit action at whole level.

Cluster algorithm is usually adopted to make a distinction of retrofit measures among different buildings on large scale. Lannon *et al*. [2.65] developed model of 55,000 houses over 50-year performance via cluster analysis, aiming to investigate the retrofit pathways to UK government's ambitious target of 80% reduction greenhouse gases emission by 2050. 100 clusters were developed to identify the dwelling with similar energy consumption and built age. Different combinations of retrofit measures were proposed and analyzed in the simulation tool. Overall, challenges and barriers in aggregate are still difficulties for individual family house.

2.1.3 Analysis of the review works

Data-driven approaches for predicting and classifying building energy consumption typically focus on total energy consumption, electricity demand, heating/cooling load and important energy parameters. The scopes of these researches are from sub-system level to single building level or even to national level.

Substantial up-to-date mythologies are proposed in order to enhance the accuracy and reliability of data-driven models, such as algorithm optimization and data pretreatment. As for algorithm optimization, micro-scale researches based on individual buildings are proposed with considerations to develop variants of basic algorithms and hybrids of several approaches. The improvements of macro-scale analysis of building energy performance are invested to increase calculation efficiency when the raw data is large and chaos. In addition to algorithm optimization,

data pretreatment is another focus for many researches. Appropriate pretreatment layered on the top of data-driven approach is the premise of accurate results and high computation efficiency. In short, high similarity between training and testing dataset is important for establishing a good model.

Meanwhile, substantial studies applied the simulated database to test model performance rather than the measured data. The analysis results of these models cannot be regarded persuasive enough as simulated data records are less fluctuant than real situation. In these scenarios, the question arises for reliability of simulated data again with no clear answer.

So far, the researches on residential buildings are not elaborated as researches of commercial buildings. The main reasons are including (1) lack of energy-use database from family-houses; (2) more freedom of occupancy behaviour in residential buildings. Hence, most researches on residential buildings are at low granularity, such as roughly profile energy consumption on regional level.

CHAPTER 3 LOAD PROFILING MODEL

3.1 Overall review of cluster analysis for occupant-behaviour

The identification of characteristic energy load patterns could make occupants aware of their energy intensive behaviours, and provide the social-technologica basis for leveraging the economic benefits and enhance the competitiveness of utility companies. Panapakidis et al. [3.1] concluded that load profiling could facilitate: (a) the comparison of similar buildings in terms of energy consumption, (b) the establishment of benchmarking procedures, boundaries and classification schemes and (c) the proposal of possible energy-efficient and environmentally-conscious improvements. Identifying the energy load patterns of occupants and classifying them based on their load profile characteristics can be useful to stakeholders aiming to improve the energy-efficiency of buildings effectively.

For this purpose, the application of clustering algorithms to analyze and classify the energy consumption behaviour of a building was proposed [3.2]. Owing to their effectiveness, the most common clustering methods in load profiling are the K-means, the self-organizing map (SOM), the minimum variance criterion (MVM), and the fuzzy C-means (FCM) [3.1-3.3]. Other less used approaches include the Hopfield neural network [3.4], the ISODATA algorithm [3.5] and the Support Vector Clustering (SVC) [3.6]. In addition, different combinations of these algorithms are also found in past studies [2.40].

Because load profiling in residential buildings are generally hindered by privacy concerns, most research has only focused on office and public buildings—despite the

fact that residential occupant behaviour is more complex and characterized by randomness [3.7]. Due to the recent surge of interest in residential energy consumption, the amount of energy data harvested through the growing installation of smart meters has also increased. Consequently, it is now necessary to develop valid and beneficial methods for presenting such data in meaningful ways to both the occupants and other stakeholders. This information is of great importance to developing countries, such as China, where residential occupant behaviour is more diverse and also contributes to considerable carbon emissions.

After reviewing peer research, a few key points are identified:

- Most research has only focused on office and public buildings—despite the fact that residential occupant behaviour is more complex and characterized by randomness [3.7]. Consequently, it is now necessary to develop valid and beneficial methods for presenting such data in meaningful ways to both the occupants and other stakeholders.

This research aims to fill the research gap of load profiling in residential buildings by exploring the situation in a developing country utilizing long-term electricity. This research presents the complex electricity behaviour of two residential communities in Shanghai in a meaningful way and provides recommendations for different stakeholders based on the findings.

3.2 Methodology for load profiling models

In this research, K-means clustering has been adopted as the cluster analysis tool to classify the occupant-behaivor based on residential electricity consumption. The classical K-means clustering method groups a dataset of N input vectors to C clusters using an iterative procedure. The specific process of K-means clustering method is shown in section 2.1.1.

Pre-processing, the first step, is aimed at restructuring the dataset and eliminating invalid data. First, the original data collected at 15-minute intervals was processed in MATLAB and structured into hourly electricity consumption from 0:00 till 23:00 for each household. Then, the datasets were analysed with MATLAB and 34 curves were discarded as corrupted data (due to noise and network failures). At the end, 36 318 days of data from 138 households were deemed suitable for further analysis.

The second step encompassed the K-means cluster analysis. Deciding the number of clusters (the K value) is of great importance, as too few clusters would affect the accuracy and too many clusters would reduce the calculation efficiency. Consequently, we tested several K values and selected 10 as the most reasonable for our purpose. Root mean-square deviation (RMSD) was used to test the difference between the measured data and the centroid data, whereas the coefficient of variation (CV) was used to compare several datasets with different sample sizes.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}\left(X_{1,t}-X_{2}\right)^{2}}{n}} \times 100, \tag{3.1}$$

$$\mathrm{CV} = \frac{\mathrm{RMSD}}{\bar{X}} \times 100, \tag{3.2}$$

The initial centroids of K clusters were randomly chosen by MATLAB from the electricity consumption dataset. In Eq. (3.1), $X_{1,t}$ is the measured data and X_2 is the cluster centroid data and $\bar{X}$ is the mean value of within-cluster data. K-means clustering used RMSD to minimize the distance between measured data with their cluster centroids. It is worth noting from **Fig. 3.1** that the rate of decrease in the case of CV (RMSD) is gradual, while the calculation time doubles as the cluster number exceeded 15. As a result, we decided to adopt a 10-cluster approach, as this configuration performed well both in terms of efficiency and accuracy.

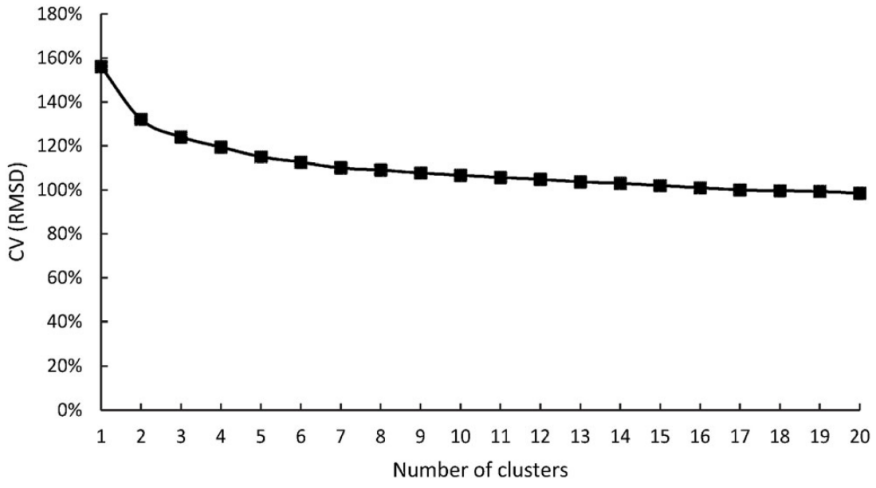


Fig. 3-1: CV value at different number of clusters

3.3 Data sets

The meteorological data for the period May 2013 to December 2015 was obtained from two public housing communities in Shanghai:

- Community A was built in 2012 and is located in Yangpu District. The 2 air-conditioning units (671 W), the washing machine, and the refrigerator (0.49

kWh per day) were pre-installed. There are 40 metered households in this community with 2–3 residents and 70 m^2 floor area on average.

- Community B was built in 2013 and is located in Putuo District. Here too, the basic domestic appliances were pre-installed and the households were equipped the same appliances as Community A. There are 132 metered households in this community with two apartment configurations of 45 m^2 and 60 m^2 size and with an average of 2–3 residents per household.

According to the Köppene-Geiger climate classification, Shanghai has a humid subtropical climate with four distinct seasons. Hence, Shanghai's climate corresponds well to therules specified in the Energy Efficiency Design Standards for Residential Buildings in the Hot Summer and Cold Winter Zone. The set-point heating temperature is about 18 °C and the heating season is from December to February. The set-point cooling temperature is around 26 °C and the cooling season is from June to August. The average outdoor temperature in July and January is 27.8 °C and 3.7 °C, respectively. Neither of the observed districts are supplied with centralized heating systems.

Considering the complexity of the energy performance of an actual building, it is difficult to separate the influence of the occupant behaviour from other factors, such as the climate, the physical characteristics of the building, and the type of the installed appliances. However, since our data is obtained from the same city, the effect of climate is expected to be weak. Likewise, since both communities were built about the same time with similar technologies, their building characteristics (such as house

type, floor area, equivalent leakage areas and heat loss coefficient) can be assumed to be similar as well. Additionally, as appliances were uniformly pre-installed, their impact on the electricity loads due to their technological differences is expected to be also marginal. Consequently, our collected dataset enables us to examine the influence of end-use behaviours on energy use patterns.

3.4 Analysis of daily consumption

Figure 3-2 presents the centroids of hourly electricity consumption patterns over a day for each cluster. The 10 clusters, derived from the 138-household data, represent 10 different electricity consumption behaviours. The breakdown distribution of the ten clusters is illustrated in **Fig. 3-3**. C2 represents 51% of residents, followed by C9 and C8 with 23% and 6%, respectively.

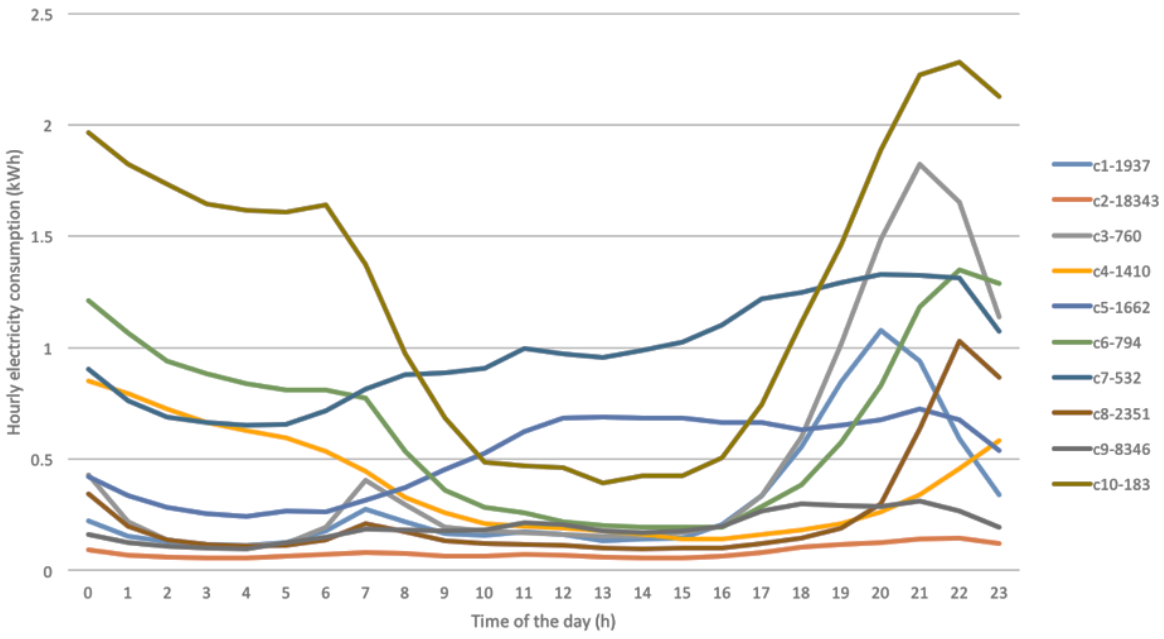


Fig. 3-2: Centroids of 10 clusters

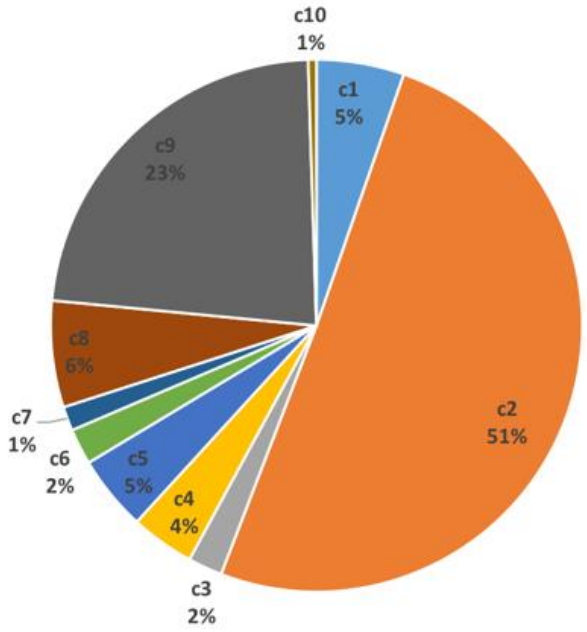
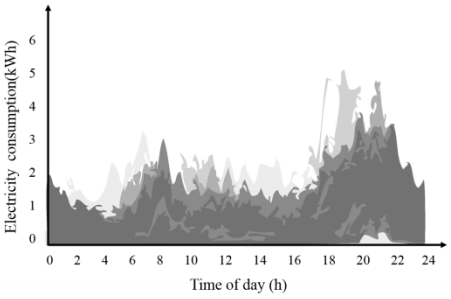


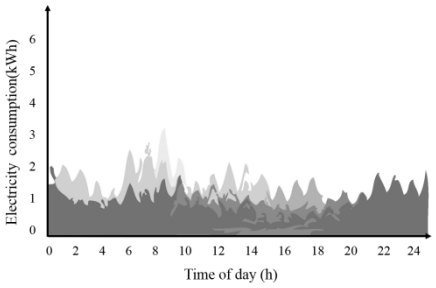
Fig. 3-3: Percent of sample size of 10 clusters

Daily electricity load profiles are shown in **Fig. 3-4**. A double peak, with low morning and high evening consumption levels, characterizes the electricity use profiles of the C1, C3 and C8 clusters—categorized as the clusters of mostly white-collar workers. They regularly use lighting and other appliances in kitchens, bathrooms and living rooms in the morning and evening, before and after office hours. In contrast, the households of C2, C5, C7 and C9— believed to be predominantly poor or older families—exhibit pronounced mid-day energy demand that extends well into the evening. These occupants demonstrate energy-conscious behaviour owing to their culture. Particularly, the load profiles of C2 and C9 (poor and/or elderly occupants) are distinguished by their relatively evenly distributed low electricity consumption levels. In addition, load curves from C4, C6 and C10 clusters—exemplifying rich and/or young families—have extremely high nighttime electricity consumption levels

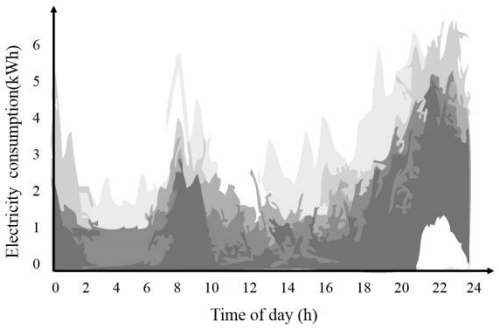
due to air conditioning. Differences between the 10 identified clusters are large owing to the diversity of occupant behaviours. Consequently, considerable energy-saving potential can be identified among these groups. Unlike to residential communities in developed countries—with comparable social status, climate and time of construction, which are characterized by fully automatized control systems and full-time, full-space mode of operation—occupant behaviour is extremely random and fluctuates over a wide range in emerging economies, such as China.



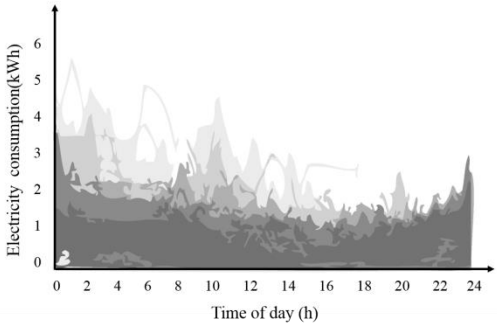
Cluster 1



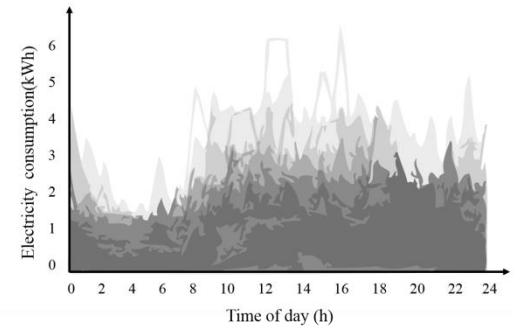
Cluster 2



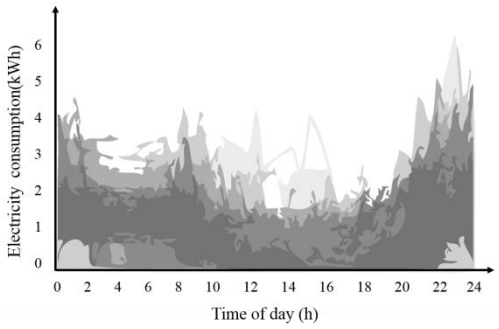
Cluster 3



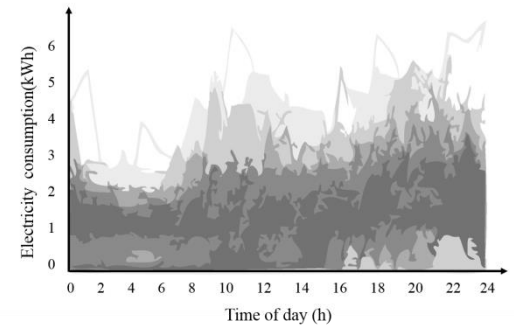
Cluster 4



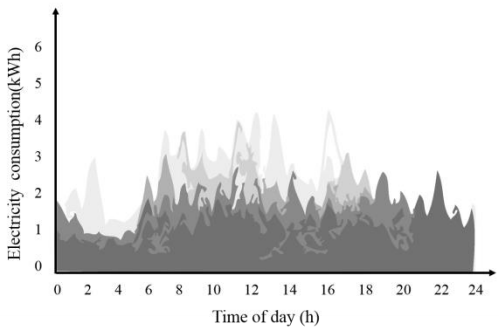
Cluster 5



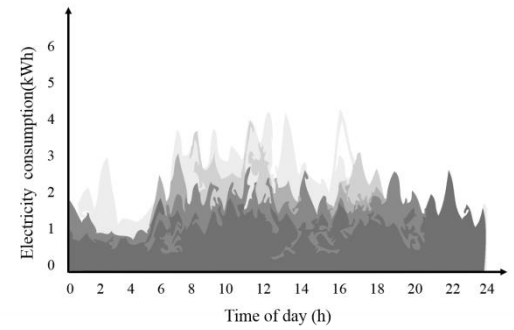
Cluster6



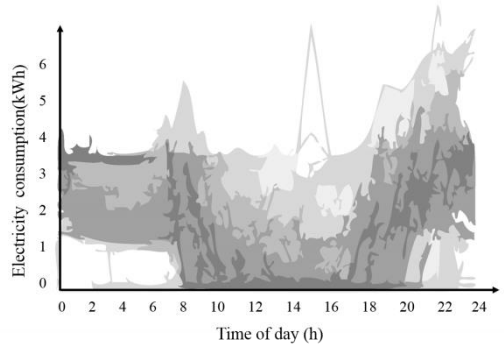
Cluster 7



Cluster 8



Cluster 9



Cluster 10

Fig. 3-4: Results of 10-cluster K-mean clustering analysis

As part of demand-side management (DSM) measures, a few load shaping strategies could be recommended on the basis of each clusters' diurnal load profile. Some profiles have low load values (according to **Fig. 3-2**), in which case load shifting, valley filling, conservation and peak clipping are suitable measures to smooth the profile shape, and hence to improve the building energy efficiency. The prolonged high daytime consumption pattern charactering C5 and C7 could be addressed both by peak clipping or conservation methods. Compared to compulsive peak clipping measure that generally impair the quality of life by reducing the thermal comfort of residents, visible smart metering could encourage occupants to improve their energy efficiency. Besides behaviour change, replacing old plug-in equipment with high efficiency appliances is also an effective DSM approach. However, when it comes to necessary loads during morning and evening periods that cannot be reduced further—as in the case of in C1, C3, C6 and C10—load shifting and valley filling might be appropriate measures. In these cases, dynamic Time-Of-Use (TOU) pricing plays a critical role in DSM, as it enables residents to optimize their energy use by selecting appropriate periods/tariffs for operating their appliances.

3.5 Analysis of seasonal consumption

Table 3-1 shows the correlation analysis of different months and clusters. In order to minimize the influence of different amount of days and thus data in each month and to perform a more direct evaluation of seasonal impacts, the collected data was normalized to percentage values. As presented in **Table 3-1**, seasonal changes in electricity consumption are indicated by increased heating and cooling demands. Based on their seasonal energy usage, the consumption patterns of the 10 clusters can

be categorized as: dominated by the heating period (C1, C3, C6, C7 and C10), dominated by the cooling period (C4, C5 and C8), no distinguished features (C2) and dominated by the transitional-seasons (C9). According to the clustering results, the overall electricity consumption in the two communities was slightly higher in winter than in summer.

Table 3-1: Adjusted percentage of correlation analysis between cluster and month

Month\Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	19%	5%	27%	17%	10%	31%	24%	16%	6%	35%
2	15%	7%	24%	11%	12%	21%	20%	13%	5%	33%
3	9%	8%	8%	7%	5%	8%	8%	11%	9%	6%
4	3%	10%	3%	5%	1%	0	1%	2%	11%	0
5	1%	11%	0%	1%	1%	0	0	1%	9%	0
6	2%	11%	0	1%	2%	0	0	1%	9%	0
7	13%	4%	2%	24%	30%	9%	15%	19%	7%	0
8	9%	7%	1%	15%	20%	3%	4%	12%	9%	0
9	4%	10%	1%	3%	3%	0	0	3%	10%	0
10	1%	12%	0%	0%	0%	0	0	0%	8%	0
11	4%	10%	1%	2%	3%	0%	2%	4%	11%	0
12	21%	4%	32%	13%	13%	27%	27%	17%	6%	27%
	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Characteristics	W&S	\	W	S&W	S&W	W&S	W&S	W=S	\	W

The lifestyle and behaviour of southern Chinese residents are unique. According to a relevant survey [3.8], it is common to use air conditioning combined with partially heating. In residential buildings, the percentage of effective heating (when the air-conditioning system is turned on) to occupancy period is around 20% in winter. In other words, residents rarely keep their AC running all the time for heating. Furthermore, nearly 85% of occupants prefer to open their windows for fresh air. The research by Jian et al. [3.9] showed that summertime electricity consumption from AC is strongly influenced by occupancy patterns. The seasonal electricity consumption within the 10 clusters ranges from less than 0.1 kWh/m^2 to 7.4 kWh/m^2. **Figure 3-5** depicts each cluster's electricity load in the form of boxplots based on the seasonal energy use levels. The centroid values of the 10 clusters are regarded as reference values, while minimum and maximum consumptions are given in average values. The average seasonal electricity load in the identified clusters varies significantly, from nearly 0.01 kWh to 3.5 kWh. These diverse energy consumption levels indicate great potentials for electricity saving by improved occupant behaviour, especially for clusters belonging to the white-collar workers and rich and/or young family category.

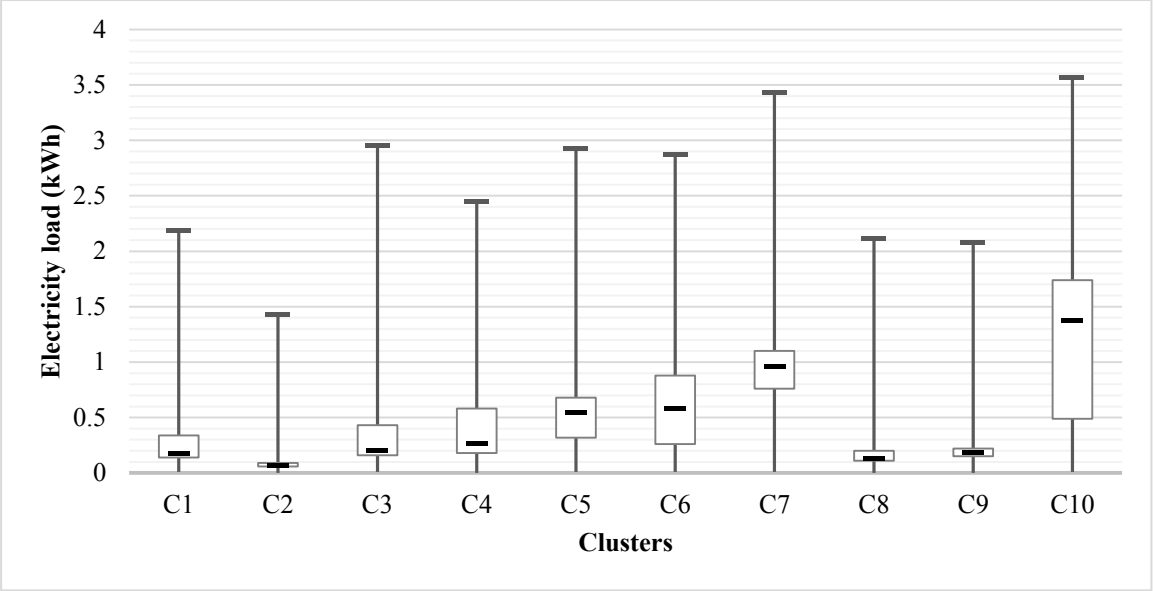


Fig. 3-5: Boxplot of electricity load of each cluster

3.6 Analysis of weekly consumption

Table 3-2 shows the correlation analysis between the energy consumption of each cluster and different periods of the week. Because weekend days comprise 28.57% (2/7) of a week and weekdays comprise 71.4% (5/7), C5, C7 and C9 indicate significant load shifting to weekend days. A comprehensive summary of the electricity consumption characteristics is shown in **Table 3-3**. The combination of K-means clustering and statistical analysis presents a viable approach to identifying and characterizing different electricity load patterns. The cooling season in Shanghai lasts from July to August. However, according to our results, typical cooling season characteristics are not reflected in C2. This can be partly explained by the fact that community B is located near a campus and most of its residents are lecturers who are

generally away for holiday in August. Another possible explanation is the occupants' strong energy conservation awareness. These observations indicate that even a limited amount of information can affect the assumed electricity load pattern and hence the accuracy of the predicted energy demand. Therefore, future studies should place more emphasis on gathering additional information about the occupants.

Table 3-2: Correlation analysis between cluster and weekday/weekend

Day/ Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Weekday	1449	13394	573	1006	947	579	306	1788	5256	128	25426
Percentage	75%	73%	75%	71%	57%	73%	58%	76%	63%	70%	
Weekend	488	4949	187	404	715	215	226	563	3090	55	10892
Percentage	25%	27%	25%	29%	**43%**	27%	**43%**	24%	**37%**	30%	
Total	1937	18343	760	1410	1662	794	532	2351	8346	183	36318
Percentage	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
Characteristics					Weekend		Weekend		Weekend		

Table 3-3: Electricity consumption characteristics of each cluster

Cluster	**Night period**	**Morning peak**	**During daytime**	**Night peak**	**characteristics**
1		3kWh@7:00		4kWh@22:00	Normal weekday

2					No significant consumption
3		4kWh@7:00		5kWh@21:00	Normal weekday, higher than c1
4	Constant				Winter/summer, weekday
5			Peak at different time		Weekend
6	Constant	3kWh@7:00		5kWh@23:00	Winter/summer, weekday
7	Constant		Peak at different time		Winter/summer, weekend
8		2.5kWh@7:00		3kWh@23:00	Normal weekday, later peak
9			Peak at different time		Weekend, lower than c5
10	Constant, and high			5kWh@23:00	Winter/summer, extreme weather

In order to balance the effects of the seasonal variability of electricity intensity distribution owing to varying heating and cooling loads (see **Table 3-1** and **Table 3-3**), electricity suppliers should improve the reliability of the national grid and avoid compulsory peak clipping measures. In general, energy use associated with heating and cooling in residential buildings has the greatest potential for conservation, since different set-point temperatures resulting from distinct occupant behaviours and preferences can cause markedly different energy consumption patterns. Hence, energy

awareness education and outreach programs aiming to increasing building energy efficiency without negatively affecting the thermal comfort of residents should be widely implemented to reduce peak demands.

3.7 Chapter summary

This paper presented a systematic approach to characterizing the electricity load patterns of two residential communities in Shanghai on the basis of occupant behaviours using the standard K-means clustering method. Daily/weekly/seasonal electricity consumption patterns have been profiled and analyzed.

Occupants were categorized as white-collar workers, poor or older families and rich or young families owing to their load patterns. In our study, the group of poor or older families constituted the largest group, accounting for nearly 80% of the total sample. In our case, the observed occupant behaviours were much more random and fluctuated over a wide range. The majority of metered households are characterized by continuous low consumption levels. Only a small proportion of households displayed the dual peak pattern with increased morning and evening consumption levels. The weekly analysis found significant load shifting towards weekend days in the case of the poor or old family group. Based on the seasonal electricity loads patterns, the clusters could be classified as dominated by heating period, dominated by cooling period, no distinguished features and dominated by energy use during the transitional seasons. The seasonal electricity consumption ranged from less than 0.1 kWh/m^2 to 7.4 kWh/m^2. The overall electricity consumption of the observed communities was slightly higher in winter than in summer. Great electricity-saving

potential was observed within the group of white-collar workers and among the rich or young families as the individual loads varied a lot.

Based on the load profiling results, our recommendations to stakeholders for smoothing the load shape and improving building energy efficiency include approaches such as load shifting, valley filling, conservation, peak clipping and TOU.

CHAPTER 4 WINDOW BEHAVIOUR MODEL

4.1 Overall review of window behavior modelling approaches

In recent years, research about building energy gradually focuses on the occupancy behaviours, because indoor occupancy is an important factor directly influencing building energy use[4.1, 4.2]. The need to integrate occupancy behaviour into building energy use has brought more awareness to the window operation behaviour. Specifically, various stochastic models of window operation have been proposed aiming to capture the occupants' interaction with windows based on several influencing factors.

Logistical regression is the most popular stochastic method used for change of window state [4.3, 4.4]. For example, logistical regression method was used to infer the probability of opening and closing a window based on 15 dwellings in Denmark [4.5]. Indoor CO_2 concentration and the outdoor temperature were identified as the most important influencing variables in determining the probability of opening and closing windows, respectively. More elaborated models were conducted by D. Calì et al. [4.6] based on logistical regression to identify the drives of opening and closing windows according to different room typologies (kitchen, bathroom, living room and others) and the window types (balcony doors and typical windows). The time of day and average outdoor temperature remained the most frequent drivers for opening and closing, respectively, across all room typologies and window types. They also found that window operation behaviour was usually driven by the activity in the home. As

for office building, F. Naspi et al. [4.7] conducted the model to predict the window opening at arrival and opening/closing during intermediate period. According to F. Stazi et al. [4.8], CO_2 concentration was found to have no statistical meaning to window operation in school classrooms. Besides change of window state, the state of window was also widely analyzed [4.9]. According to V. Fabi et al. [4.10], the CO_2 concentration, illumination and sun hours were able to be recognized as the common drivers for both transition model and state model. Moreover, Wei et al. [4.11] monitored the end-of-day window positions in an office building. Besides environmental parameters, non-environmental factors including seasons, daylight, occupant absence, window orientation, floor level and gender were considered in determining the end-of-day window position, and this merits further investigation.

Yun and Steemers [4.12] found that the current window status is also strongly influenced by previous status of the window. Therefore, Markov chain process, as another stochastic model, was adopted in their case study. Markov chain process not only could estimate the window status on next time step, but also perform occupant's window operation behaviour (i.e. transition probabilities of window). Based on measurements of four office rooms in winter, a Markov chain model was proposed by Fritch et al. [4.13] to predict the window angel with ambient temperature as the driving variable. The result indicated that calculated probabilities and measured probabilities are very approximate. However, one reason of the satisfying result is that the proportion of closed status is more than 92%; even up to 99% in some rooms. Different from above model, Yun and Steemers [4.14] considered that indoor conditions was the main influencing factor on window status. One-hour time step

Markovian model was built with the inputs of indoor temperature at different time of day and previous window state. More comprehensive model was conducted by Haldi et al. [4.15] based on Markov process to simulate occupants' interactions with window, more relevant environmental parameters were considered and three sub-models were developed based on different events, such as arrival, during occupancy and departure. Indoor and outdoor temperature were still considered as two main factors and involved in the probability calculation. In addition, the weather condition (e.g. raining), location (e.g. ground floor) and the room occupancy on next time were considered as the drivers of the windowing behaviour. As for residential buildings, Calì et al. [4.16] developed a stochastic window state profile in python, based on Markov chain technique. Their research focused on the model in three different combinations of influencing factors to calculate the probability of window state change. The transition probability matrix was depended on daily ambient temperature (DAT) and time of the day, DAT and the day of week, DAT range of the actual day and preceding day, respectively. During the coldest period, the first model outperformed than the other two models. Moreover, during the heating period, the model involved preceding day's temperature performed best.

Besides Logistic Regression and Markov chain technique, machine learning (ML) algorithms are also adopted to analysis window-opening behaviour, including Gaussian distribution model, deep learning, Bayesian Network, cluster analysis and association rules mining. ML models are now widely used to predict building energy consumption, but rarely used to predict window status. S. Pan et al. presented the Gauss distribution model to predict window behaviour in an office building. When

modelling, three types of input variables, i.e., indoor temperature, outdoor temperature and their combination were used. The results showed that Gauss distribution models could provide 9.5% higher prediction accuracy than Logistic regression model [4.17]. Romana Markovic et al. [4.18] built a window-opening model based on an office building in Germany with deep learning methods. This model was trained as a fully connected feed-forward neural network with twenty-five as in the input layer and five hidden layers to identify and classify the window status. Totally 21 driving factors including indoor and outdoor physical parameters were involved in the model. For Bayesian Network, Verena M. et al. [4.19] investigated the relationship between influencing factors and window status in residential buildings. Bayesian Network represented the probabilistic dependencies between a window status and a set of variables that potentially affect the status. This model with Bayesian Network algorithm could flexibly represent different typologies and handle a mix of various data types. They also mentioned that choosing appropriate influencing factors provides assurance of model's accuracy. In order to identify patterns of window opening/closing behaviour, cluster analysis and association rules mining were adopted by Simona D'Oca et al. [4.20] as the pattern recognition technique to analyze corresponding motivation, opening duration, interactivity and degree of window opening position.

Except for stochastic models above, non-stochastic models was proposed by Farhang Tahmasebi et al. [4.21] to investigate the potential in predicting occupants' interaction with windows and the effectiveness to enhance the reliability of simulation results. In the non-stochastic models, the window was opened if the indoor temperature achieved

the set temperature range. On the contrary, the window was closed when the indoor temperature exceeded a certain temperature range. Although the non-stochastic models provided closer estimations of annual heating demand and peak heating demand than stochastic models in the research, the neglecting of window opening in heating season hinder the non-stochastic models from accurate simulation, in which the occupants' control plays an important role over natural ventilation.

Recently, the $PM_{2.5}$concentration has become a highly concerned factor on occupants' interaction with windows in China. Dayi Lai et al. found that the $PM_{2.5}$concentration is the highest-percentage reason for not using windows for ventilation, which means the $PM_{2.5}$concentration do influence occupant window-opening behaviour. According to S. Pan et al., occupants' window-opening behaviour is strongly correlated to environmental factors including outdoor $PM_{2.5}$concentration in office buildings [4.22]. Mingyao Yao et al. [4.23, 4.44] regarded CO_2concentration, $PM_{2.5}$concentration, wind speed, indoor and outdoor humidity, indoor and outdoor temperature as the main influencing factors on window status. According to the result of logistic regression, the probability of window opening decreased dramatically when $PM_{2.5}$concentration was around $150\mu g/m^3$. In addition, most of observed windows tented to be closed when the concentration was higher than $150\mu g/m^3$. Besides, $PM_{2.5}$concentration was also considered in another study conducted in the hospital wards [4.45], however no significant correlations were found between the window opening probability and outdoor $PM_{2.5}$ concentration. They also pointed out that $PM_{2.5}$ concentration does influence the window opening behaviour in residential buildings in the same city.

A summary of the various model types, building use, locations and reported accuracies of window states is presentable in **Table 4.1**.

Table 4-1: Overview of referenced studies of window opening/closing models

Source	Model	Building use	Location	Most important influencing factors	Seasons	Accuracy
[4.3]	Logistical regression	Residential buildings	UK	indoor and outdoor air temperature and wind speed	1 year	NA
[4.4]	Logistical regression	Residential buildings	Seoul	indoor and outdoor air temperature	winter and spring	
[4.5]	Logistical regression	Residential buildings	Denmark	indoor CO_2 concentration and the outdoor temperature	winter, spring, and summer	NA
[4.6]	Logistical regression	Residential buildings	German	time of day and average outdoor temperature	4 years	NA
[4.7]	Logistical regression	office buildings	Italy	Time of day, indoor and outdoor temperatures.	spring, summer and autumn	AUC =0.54 -0.9
[4.8]	Linear and logistical regression	School classrooms	Italy	indoor and outdoor air temperature	Spring (1 month)	AUC =0.719
[4.9]	Logistical regression	office buildings	UK	Outdoor temperature	Winter and spring	r^2=0.831

[4.10]	Logistical regression	Residential buildings	Japan, Switzerland and Denmark	CO_2 concentration, illumination and sun hours	3-8 months	2%-91%
[4.11]	Logistical regression	office buildings	UK	Outdoor temperature and some non-environmental factors	winter, spring, and autumn	NA
[4.13]	Markov process	office buildings	Switzerland	Outdoor temperature	winter	NA
[4.14]	Markov process	office	UK	Indoor temperature	Warm summer(5 days)	NA
[4.15]	Markov process	office buildings	Switzerland	Relative humidity, wind speed, indoor and outdoor temperature	95 months	ACC =0.664
[4.16]	Markov process	Residential buildings	German	Time of day, day of week, average ambient temperature	3 years	MAE =0.51 %
[4.17]	Gauss distribution	office buildings	China	indoor temperature, outdoor temperature and a combination of them	Autumn(1.5 months), spring(3 months)	ACC =0.74
[4.18]	Deep learning	office buildings	German	Indoor climate and weather data (21 features totally)	16 months	ACC =0.89 F1=0.63
[4.19]	Bayesian Network	Residential buildings	Denmark	Solar radiation, CO_2 concentration, relative humidity, time of day, indoor and outdoor temperature		ACC =0.93
[4.20]	Cluster analysis and	office buildings	German	Arrival time, occupant presence, time of day, indoor	Winter and summer	NA

	association rules mining			and outdoor temperature		
[4.21]	Non-stochastic	office buildings	Austria	Comfort temperature, indoor and outdoor temperature	1 year	Fraction of correct states =0.95
[4.22]	logistic regression and Pearson correlation	office buildings	China	Indoor and outdoor air temperatures, wind speed, relative humidity, outdoor $PM_{2.5}$ concentrations, solar radiation, sunshine hours	9.5 months	NA
[4.23]	Logistical regression	Residential buildings	China	CO_2 concentration, $PM_{2.5}$ concentration, wind speed, indoor and outdoor humidity, indoor and outdoor temperature	spring	NA
[4.24]	Logistical regression	Hospital wards	China	Relative humidity, wind speed, wind direction, rain, indoor and outdoor temperature, CO_2 concentration	1 year	$AUC_{cooling}$= 0.836 $AUC_{transition}$ =0.875 $AUC_{heating}$= 0.823
[4.25]	Logistical regression	Residential buildings	China	CO_2 concentration, $PM_{2.5}$ concentration, wind speed, indoor and outdoor humidity, indoor and outdoor temperature	4 seasons	Percentage correct=0.72

The development of window behaviour model is based on Logistic Regression, Markov process, Gauss distribution and artificial neural network. In this section, these main approaches applied in window behaviour would be introduced.

After reviewing the peer researches, a few key points are identified:

- Most methods employed to simulate window operation include logit regression and Markov chain technique. The application of machine learning to occupant window behaviour is preliminarily under investigation and could be potentially highly effective. Especially, ANN based approaches possess a number of advantages, such as fault tolerance, robustness and noise immunity. Hence, ANN based approaches have achieved great success in solving non-linear problems. Furthermore, the ANN also allows for consideration of a variety of explanatory variables and multiple target variables in a network structure. Although ANN based approach like deep learning has been applied to investigate window operation, the application of the algorithm still needs further exploration. As one of data-driven models, the reliability and robust of ANN rely on large size of the database, but in China, there are neither enough relevant researchers, nor tens of thousands available datasets like Romana Markovic et al. did. Hence, the application of ANN models with fewer data samples requires further investigation. In this paper, the application and optimization of ANN model with BP algorithm under thousands available training samples would be investigated.

- Researchers adopted different indices to evaluate the performance of their models, where lacks a horizontal comparison among these models. Therefore, it is significant to deliver a more comprehensive evaluation of different models developed based on multiple algorithms under the same indices and the same datasets.
- Most published studies referring to occupant window behaviour have been carried out within European countries, where the influence of outdoor air quality is rarely taken into account. Until now, the investigation of occupants' window behaviour in China is shallow. The in-depth research of window behaviour in China area is necessary because this country has very different conditions in atmospheric environment and living habits. People tent to prevent the leakage from outdoor when the air pollution was serious. Therefore, investigating the influence of $PM_{2.5}$concentration on window status in China is very necessary.

This research aims to fully investigate occupant window opening models during transition season based on logistic regression, Markov process, Gauss distribution and ANN algorithm, which is proposed to explore the application and optimization of ANN algorithm under less samples condition in the paper. Moreover, $PM_{2.5}$concentration is considered as an influencing factor to build window opening model of office building in China area.

4.2 Methodologies for window behaviour model

4.2.1 Logistic Regression

The first approach used to infer the window state (0, 1) from the selected environmental parameters is logistic regression. As we know, linear regression and logistic regression models are two effective and prevailing statistical methods for determining the main influencing factors on the dependent variables. The application of linear regression requires the analyzed data to be linear, normal, or homoscedastic, which does not work when the datasets are dependent or have binary outcomes. Thus, logistic regression was proposed to solve the limitations mentioned above. The logistic regression model transforms the simplified x to the form $g(x) = \frac{1}{1+e^{-x}}$ based on a sigmoid function, normalizing the parameters into a binary result as 0 or 1 [4.46]. The definition of the logistic regression is:

$$\ln \frac{P_i}{1-P_i} = \alpha + \sum_{k=1}^{k} \beta_k x_{ki} \tag{4.1}$$

In (4.1):

P is the probability of the window opening state;

i is the sample number;

k is the number of independent variables;

α is the intercept; and

β are coefficients.

The coefficients were estimated by maximum likelihood estimation. The magnitude of the coefficients reveals the degree of each factor's influence on the window state. Therefore, logistic regression is the basic procedure to select effective parameters.

Indoor and outdoor temperatures were considered as highly correlated to the window state in the existing studies. In this study, four more related factors were found to have significant influence on the window status through 10-fold validations, as presented in **Table 4-2**. In this study, the numbers of influencing factors were determined as four (group 1), five (group 2), and six (group 3).

Table 4-2: Coefficients and intercept of the window state models based on multiple parameter regression

Group	Model input	Coefficients β						Intercept α
		Indoor temperature (θ_{it})	outdoor temperature (θ_{ot})	outdoor humidity (θ_{oh})	wind speed (θ_{ws})	sunshine hours (θ_{sh})	$PM_{2.5}$ concentration (θ_{PM})	
1	$\theta_{it}, \theta_{ot}, \theta_{ws}, \theta_{sh}$	0.157 ±0.006	0.040 ±0.003	-	0.329 ±0.015	-0.047 ±0.002	-	-4.497 ±0.163
2	$\theta_{it}, \theta_{ot}, \theta_{ws}, \theta_{sh}, \theta_{oh}$	0.156 ±0.017	0.044 ±0.001	0.006 ±0.001	0.349 ±0.050	-0.052 ±0.005	-	-4.741 ±0.418
3	$\theta_{it}, \theta_{ot}, \theta_{ws}, \theta_{sh}, \theta_{oh}, \theta_{PM}$	0.153 ±0.020	0.044 ±0.004	0.006 ±0.007	0.391 ±0.036	-0.046 ±0.004	-0.002 ±0.001	-4.427 ±0.427

4.2.2 Discrete-time Markov processes

Although the method of logistic regression is able to select factors that have significant effects on the window state, it does not consider the transition in the state of the window ($0 \rightarrow 1$ and $1 \rightarrow 0$). Therefore, we described discrete-time Markov processes, which more accurately simulate the transition between the states of window, rather than predicting the window status. The general scheme of this model is shown in **Fig 4.1**. The first step was to estimate the probability of the initial window state (for any selected period), based on relevant factors. The second step was to construct a Markov transition matrix that describes the transition probability of window states [4.47]. The calculation is developed based on the coefficients shown in **Table 4.3**. The third step was to use the Markov process' definition to predict the window state at next time step, based on the current state. The Markov process' definition is described as in formula (4.2):

$$(X_0^{T+1}, X_1^{T+1}) = (X_0^T, X_1^T) \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix} \quad (4.2)$$

In formula (4.2):

X_0^{T+1}, X_1^{T+1} is the probability of window closed/open state on the next period;

X_0^T, X_1^T is the probability of window closed/open state on the current period;

P_{00}, P_{11} is the probability that the window remains closed/open;

P_{01}, P_{10} is the probability that the window is to be opened/closed; and

$P_{00}+P_{01}=1$; $P_{10} + P_{11} = 1$.

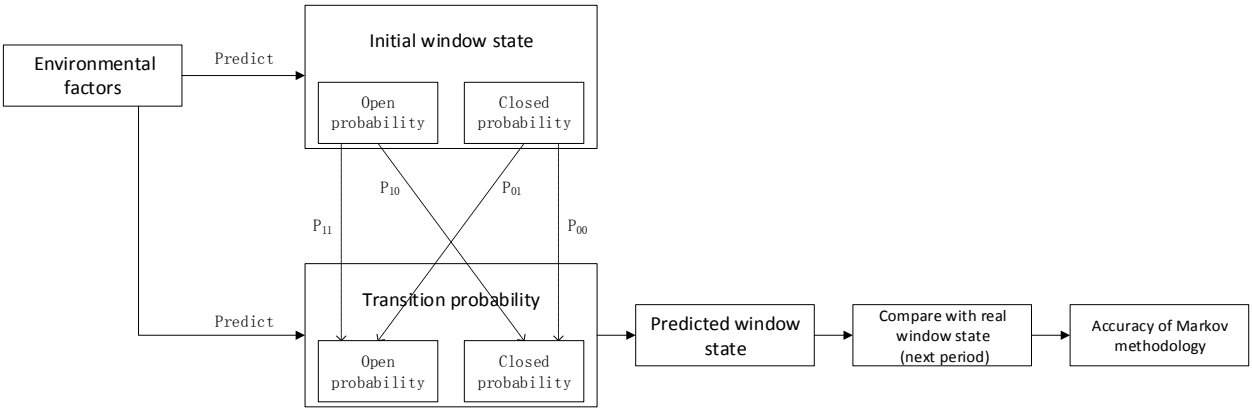


Fig. 4-1: General scheme of Markov process

Table 4-3: Coefficients and intercept of Markov transition matrix's element based on multiple parameter regression

Model input	P_{01} Coefficients β			Intercept α
	solar radiation(θ_{sr})	Indoor temperature (θ_{it})	$PM_{2.5}$ concentration (θ_{PM})	
θ_{sr}	0.001	-	-	-4.363
θ_{it}	-	0.133±0.008	-	-3.997±0.236
θ_{it}, θ_{PM}		0.130±0.001	-0.002	-3.642±0.088
θ_{sr}, θ_{it}	0.001	0.153	-	-4.610
Model input	P_{11} Coefficients β			Intercept α
	outdoor temperature (θ_{ot})	Indoor temperature (θ_{it})	wind speed (θ_{ws})	
θ_{it}	-	0.205	-	-4.650
θ_{it}, θ_{ws}	-	0.257	-0.188	-5.804
$\theta_{ot}, \theta_{it}, \theta_{ws}$	0.025±0.003	0.199±0.006	-0.218±0.044	-4.847±0.163

4.2.3 Artificial neural network

Sometimes, the prediction performance of the traditional probability approaches is not satisfactory, because they cannot adequately reflect the nonlinear and stochastic relationships among parameters in the round. Hence, Back Propagation (BP) neural network has been adopted as a machine learning approach to define the nonlinear

relationships between input parameters and output parameters. The structure of BP network is consisted of input layer, output layer and hidden layers, as shown in section 2.2.1.

The modelling idea of window status prediction based on BP algorithm is shown in **Fig 4-2**. Firstly, appropriate influencing factors were selected based on the analysis of window opening behaviour; then BP network was established with suitable structure and model parameters. According to the result from section 4.2.1, the window states are potentially influenced by four, five and six factors, respectively. In order to investigate the performance difference among logistic regression model, Markov model and ANN model, the influencing factors of these three prediction models are same. The measured data was randomly divided into training group and testing group by 4:1 in order to test whether the network has good learning ability and generalization. After that, BP network was trained for classification of window states until the output errors fall within the required threshold. Finally, the testing group was used to validate the model performance in prediction accuracy of window opening behaviour.

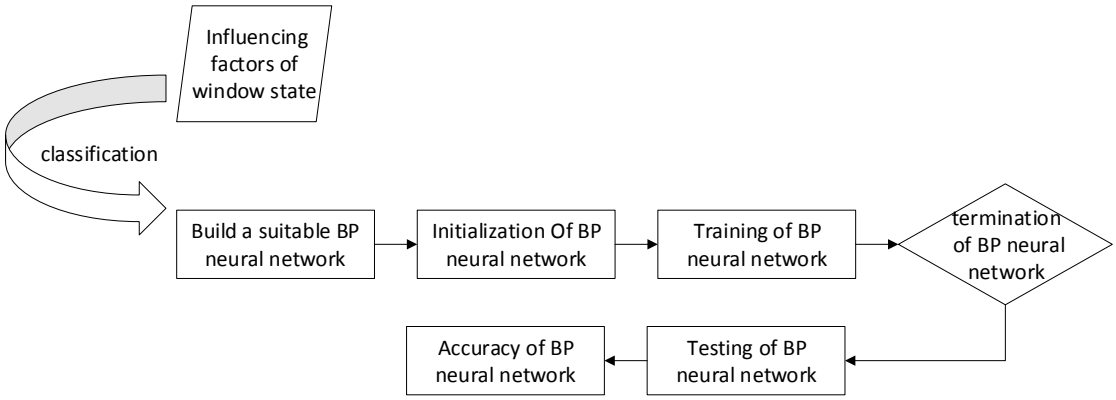


Fig. 4-2: modelling of window-opening behaviour based on BP

4.2.4 Gauss Distribution model

Window behaviour is a very complex phenomenon with a high degree of nonlinearity and randomness and may be driven by many factors. In existing studies, environmental factors, such as indoor temperature and outdoor temperature, have been identified as main driving factors and were usually used as inputs of predicting models for window behaviour. Hence, multivariate Gauss distribution model has been adopted in this study because more than one environmental factor were considered, expressed as a vector superposition of Gaussian distribution. The specific calculation of window opening probability is shown in eq. (4.3):

$$P(X) = \sum_{i=1}^{t} m_i F\left(x_i \mid \mu_i, \sigma_i^2\right) \tag{4.3}$$

where $P(X)$ represents probability of window opening; X is a vector representing t kinds of environmental factors (x_i , $i = 1, \dots, t$); m_i represents the weight coefficient of the i^{th} influential factor; $F(x_i \mid \mu_i, \sigma_i^2)$ represents the cumulative distribution function of the Gaussian distribution of the i^{th} influential factor; x_i represents the specific value of the i^{th} influential factor, and μ_i and σ_i^2 represent the corresponding mean value and variance, respectively.

As seen from eq (4.3), to calculate the probability of window opening, $P(X)$, the cumulative distribution function of Gaussian distribution $F(x \mid \mu, \sigma^2)$ and the weight

coefficients for all input variables need to be determined, following the specific steps described below:

1. Calculating the cumulative distribution function of Gaussian distribution $F(x \mid \mu, \sigma^2)$

The cumulative distribution function of the Gaussian distribution $F(x \mid \mu, \sigma^2)$ refers to the probability that random variable *X* is less than or equal to *X*, expressed as a density function and shown in eq. (4.4):

$$F(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{4.4}$$

Before calculating $F(x \mid \mu, \sigma^2)$ for each input variable, corresponding values of μ and σ^2 need to be determined. In this model, the least-square method has been adopted to find the corresponding μ and σ^2 for the training dataset, following the steps below:

a. For each environmental parameter that has a significant impact on window behaviour, the corresponding probability of window opening needs to be determined. Since all environmental parameters are time-continuous, a suitable interval length should be selected;

b. There is a set of datasets of environmental parameters for each interval selected in step a, the probability of window opening based on each dataset is calculated and the median is adopted to present the probability of window opening at that interval;

c. The least-squares method is used to fit these datasets and determine the mean and variance of each environment variable.

2. Calculating weight coefficient m

In reality, there is more than one environmental factor that has a significant influence on occupant window behaviour. Therefore, to predict window behaviour more accurately, it is necessary to describe the influence degree, named weight, for each factor, as shown in Formula (4.3). Then, in view of the normalization requirement, the influential factors selected as model inputs are considered to impact window behaviour, leading to eq. (4.5):

$$\sum_{i=1}^{t} m_i = 1 \tag{4.5}$$

In this study, the following assumptions have been used: the window was "closed" when the window opening probability $P(X) < 0.5$, and it was "open" when the window opening probability $P(X) \geq 0.5$, as shown in eq. (4.6):

$$\begin{cases} \sum_{i=1}^{t} m_i \sum_{l=1}^{k} F\left(x_{il} \mid \mu_{il}, \sigma_{il}^2\right) < 0.5, window\ is\ closed \\ \sum_{i=1}^{t} m_i \sum_{l=k+1}^{p} F\left(x_{il} \mid \mu_{il}, \sigma_{il}^2\right) \geq 0.5, window\ is\ open \end{cases} \tag{4.6}$$

where k is the number of closed window states, and p is the total number of window states for the training dataset.

As mentioned above, the cumulative distribution function of the Gaussian distribution $F(x \mid \mu, \sigma^2)$ could be determined in accordance with Steps a–c. Afterwards, the weight coefficient m could be determined by eq. (4.5) and (4.6). In addition, the

Monte Carlo method is used to determine the final window state (opened or closed) in this model. A random number in the range of 0 to 1 is generated and compared with calculated probability of window behaviour. If the random number is less than the calculated probability of window behaviour, the output is "opened", otherwise it is "closed".

The frame structure of the Gauss distribution model is shown in **Fig. 4-3**.

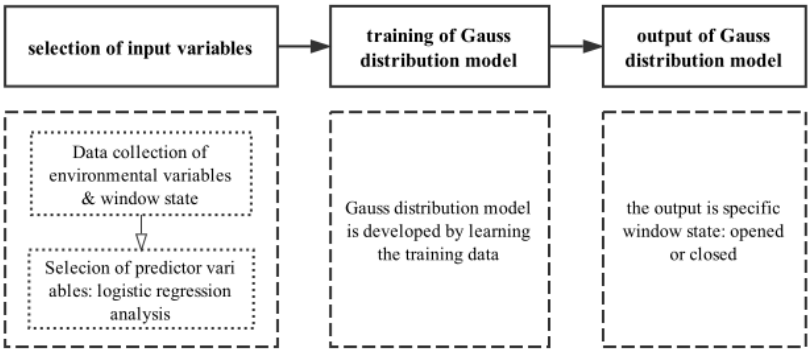


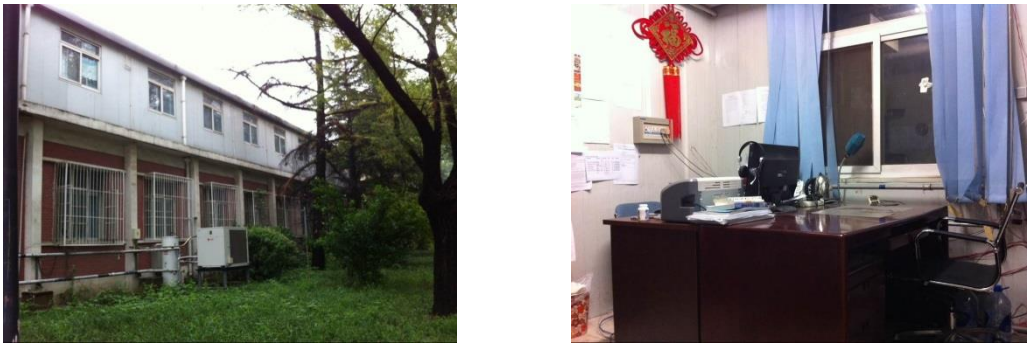
Fig. 4-3: Frame structure of the Gauss distribution model

The establishment of these four window behavior models was based on the MATLAB program.

4.3 Data sets

An office building constructed of reinforced concrete and brick, located at a university in Beijing, was selected for the field study. As shown in **Fig 4-4(a)**, the building has two stories with laboratories located on the ground floor and nine offices on the second floor with the same size of $10m^2$. There are no tall buildings or trees blocking solar gains. The internal arrangement of the typical office is shown in **Fig 4-4(b)**, there was one south-facing sliding window which could be controlled by the occupants in each office. Five out of nine offices were selected for the experiment and

all of them were single occupied by the same and non-smoking teachers during the entire measurement period. The data was collected during transition season (03/15/2015-05/16/2015), when natural ventilation is the main strategy to adjust indoor thermal environment and air quality. The influence of noise on window status has been included in the onsite questionnaires. The result showed that surrounding is usually quiet, hence the noise factor would not be considered as an driver for window-opening behaviour in this case study.



(a) (b)

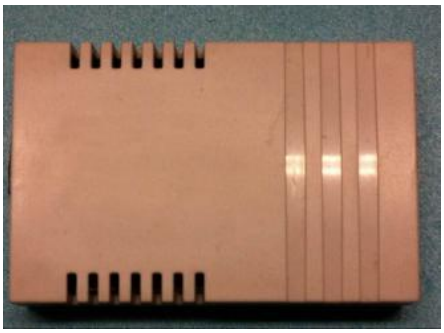
Fig. 4-4: Case study building (a) and a typical office (b)

During the measurement, an indoor air temperature sensor and portable outdoor meteorological temperature sensor were used to measure indoor and outdoor air temperature respectively. Indoor air temperature sensor was placed inside each room to avoid direct sunshine and local heating resources. Extra outdoor parameters included air humidity, $PM_{2.5}$ concentration, solar radiation, sunshine hours, wind speed and direction, were measured by outdoor meteorological station on the roof of the office building. In addition, intelligent human body inductor was used to record occupancy of the monitored office. Window displacement tester was applied to detect

and record the state of windows. The window displacement tester recorded the window state by means of the magnetic induction of two dry spring pipes positioned on the window. It identified the window as opening when the window was opened more than three centimetres and the opening period exceeds three seconds. The detailed information of all instruments are shown as **Table 4-4** and **Fig.4-5**. Total 4630 recorded datasets based on 5 rooms with the time interval of 10 min were used for model establishment.

Table 4-4: Measurement range and accuracy

	Monitoring instruments/parameters	Recording interval	Sensitivity	Accuracy
Indoor monitoring instruments	Infrared instrument	10 min	5 m	
	Window displacement tester	10 min	3 cm	
	Indoor temperature sensor	10 min		±0.5°C
Outdoor meteorological station	outdoor temperature	1min		±1.0°C
	outdoor humidity	1min		±5%
	solar radiation	1min		$\pm 10 W/m^3$
	sunshine hours	1min		±0.5h
	wind speed	1min		±1m/s
	wind direction	1min		±10°
	$PM_{2.5}$ concentration	1h		$\pm 10 \mu g/m^3$



(a)

(b)

(c)

(d)

(e)

Fig. 4-5: Measuring device

Indoor temperature measuring device (a); Intelligent human body inductor (b);The window displacement tester (c) ; Outdoor temperature measuring device (d) and outdoor meteorological station (e)

4.4 Logistic regression model

Accuracy (ACC) and Mean Squared Error (MSE) has been adopted to evaluate the performance of the window behaviour models. Accuracy reflects the correct rate of the algorithm for the overall prediction of all samples, the definition of Accuracy was proposed as Eq. (4.7). In general, the higher the accuracy value, the better the performance of the algorithm. The MSE is the average difference of the square between the estimated value and true value, shown as Eq. (4.8). The smaller the value of MSE, the better performance of the BP network model describing the experimental data. Finally, the best structure of ANN was 4-25-25-1, 5-25-25-1 and 6-25-25-1, for different number of influencing factors.

$$ACC = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{4.7}$$

Where:

TP is true positive, accurately predict state 1

TN is true negative, accurately predict state 0

FP is false positive, predicted status is 1 while actual status is 0

FN is false negative, predicted status is 0 while actual status is 1

$$MSE(\mathrm{y}, \hat{\mathrm{y}}) = \frac{1}{n_{sample}} \sum_{i=1}^{n_{sample}} (\mathrm{y} - \hat{\mathrm{y}})^2 \quad (4.8)$$

Where:

y is true value of samples

$\hat{\mathrm{y}}$ is estimated value

n_{sample} is number of samples

Window state models in **Table 4-2** could be divided into three groups with different combination of related factors. Specially, the window states in model 1, 3, 5, 6 and 9 are dominated by four factors ($\theta_{it}, \theta_{ot}, \theta_{ws}$ and θ_{sh}); the window states in model 2 and 10 are dominated by five factors ($\theta_{it}, \theta_{ot}, \theta_{ws}, \theta_{sh}$ and θ_{oh}); in model 4, 7 and 8, six factors are employed as the important drivers of window state ($\theta_{it}, \theta_{ot}, \theta_{ws}, \theta_{sh}, \theta_{oh}$ and θ_{PM}). In addition, the positive influence and negative influence of the related factors on window states are shown in **Table 4-2**. It can be observed based on the sign of the coefficient, e.g., a negative coefficient shows a negative correlation to the probability of state. For the presented logistic regression model, the probability of window open state grows with increasing indoor temperature, outdoor temperature, wind speed and outdoor humidity. The remarkable factors that negatively influenced the window opening state are the sunshine hours and $PM_{2.5}$ concentration.

The rank of the influencing related factors for the window opening state is illustrated in **Table 4-5**. The results highlight that indoor temperature, outdoor temperature, wind speed and sunshine hour are the most important factors with window opening state. On the contrary, the outdoor humidity and $PM_{2.5}$ concentration with the p-value higher than 0.01were not selected as necessary contributors in logistic regression model 1, 3, 5, 6 and 9.

Table 4-5: factors rank for the window opening state

Window opening status		
Rank	factor	Sig.
1	Indoor temperature	3.8×10^{-9}
2	Outdoor temperature	3.95×10^{-9}
3	Wind speed	4.51×10^{-8}
4	Sunshine hours	6.63×10^{-7}
5	Outdoor humidity	0.033936
6	$PM_{2.5}$concentration	0.148733

4.5 Markov model

As mentioned before, logistic regression models neglect the real dynamic processes of occupants' actions on the window. The related factors were used to infer the window states, rather than the actual window operation. Specifically, these models do not describe the actual probability of opening or closing the window, but rather the probability for the window to be identified as open, from the given physical

parameters. In the Markov model, a dynamic description of window operation was considered after logistic regression analysis.

Table 4-3 provides the coefficients and intercept of the Markov transition matrix. The matrix is used to calculate the probability of window opening action P_{01} and the probability of window closing action P_{10}. The window opening probability is related to solar radiation, indoor temperature, and $PM_{2.5}$ concentration, while the window closing probability is related to outdoor temperature, indoor temperature, and wind speed. Specifically, the window opening models could be divided into four groups with different combinations of related factors (θ_{sr}, θ_{it}, $\theta_{it} + \theta_{PM}$, and θ_{sr}+θ_{it}). In addition, the window closing models could be divided into three groups with different combinations of related factors (θ_{it}, $\theta_{it} + \theta_{ws}$, and $\theta_{ot} + \theta_{it} + \theta_{ws}$).The probability of a window opening action grows with an increase in solar radiation and/or indoor temperature, and declines with an increase in $PM_{2.5}$ concentration. This means occupants tend to open the office window for more sunlight, and cool and fresh air. Furthermore, an increase in wind speed and decrease in outdoor and indoor temperatures corresponded to a window closing action, to some extent.

Table 4-6 shows the rank of the strongest influencing factors for the window opening and closing actions. For the window opening action, the most important factor in this case study is the indoor temperature. The solar radiation and $PM_{2.5}$concentration are regarded as relatively mild factors for a window opening action. As for the window closing action, the wind speed and indoor and outdoor temperatures are identified as the strongest influencing factors.

The ACC results of logistic regression models and Markov models are shown in **Table 4-7**. The ACC range of the logistic regression models is between 50.50% and 54.22%, whereas the Markov models could reach an ACC between 54.00% and 60.44%. As for the logistic regression models, the results show that there is strong causality between temperature (both indoor and outdoor) and window state. One possible reason is that natural ventilation is the only method for improving an occupant's thermal comfort during the inter-seasonal period. In parallel, an increasing trend of probability of a window-open state is noticed with an increase in wind speed. This is probably because indoor occupants prefer more fresh air from the outdoors to improve the indoor air quality. Another important and negative factor for a window state is the sunshine hours, meaning that an occupant is used to opening the window during morning times. The proportion of the window-open state is decreased with the increase in sunshine hours. Although the $PM_{2.5}$ concentration and outdoor humidity have been eliminated by several logistic regression models, the ACC values of the logistic regression model with these two parameters are slightly higher. This means that models with 6 inputs > models with 5 inputs > models with 4 inputs >, where > indicates 'performs better than'. With regard to the Markov model, the ACC of the Markov model 4 is highest as compared with the other nine models. In model 4, the initial window state is estimated based on four factors ($\theta_{it}, \theta_{ot}, \theta_{ws}$, and θ_{sh}), whereas the probability of the window opening action is calculated based on the indoor temperature and $PM_{2.5}$ concentration. In addition, the impact of the $PM_{2.5}$ concentration on the calculation of the Markov model is validated by comparison of model 5 and model 6. The ACC of the Markov model increases when the $PM_{2.5}$

concentration is added as a factor. Hence, it is concluded that $PM_{2.5}$ concentration is an important contribution for window operation in this case study. Furthermore, the ACC of model 8 (with outdoor temperature) is higher than that of model 7 (without outdoor temperature), implying that the outdoor temperature plays an important role in a window close action.

Looking more closely into the ACC of these models, the boxplot of **Table 4-7** is plotted in **Fig. 6**, according to the different numbers of input parameters and prediction models. It is clear from **Fig. 6** that the Markov models that consider window operation are able to obtain a higher prediction accuracy of window states. However, the performance of the Markov models is more unstable than that of the logistic regression models. The reason for the instability of the Markov models might be the small size of the training database, which amplifies the model errors. The range of training data is incomprehensive because of the small size of the database, e.g., the ratio of windows closing is higher than the frequency of opening in the overall data, and is one of the reasons for the instability of the model accuracy. Moreover, because few abnormal data were deleted during preliminary data processing, the treated database is not strictly listed in chronological order, and with a time resolution of 10 min. This might lead to the instability of the model, because the calculation of the Markov algorithm largely depends on the data's chronological order. Moreover, the model uses two prediction probabilities (window open and window close), which can lead to an offset/ amplification of an error. Hence, the window states predicted by the Markov models are less stable as compared to those from logistic regression.

Table 4-6: Factor rank for the window opening and closing actions

Rank	Window opening action		Window closing action	
	factor	Sig.	factor	Sig.
1	Indoor temperature	7.13×10^{-5}	Wind speed	6.09×10^{-9}
2	Solar radiation	0.02564	Outdoor temperature	0.01777
3	$PM_{2.5}$concentration	0.115337	Indoor temperature	0.03323

Table 4-7: Prediction ACC of logistic regression models and Markov models

Model	Model inputs			Acc (%)		
	window state	window *open* action	window *close* action	Logistic Regression model	Markov chain model	Difference
1		θ_{sr}	$\theta_{ot}, \theta_{it}, \theta_{ws}$	52.33	54.44	2.11
2	$\theta_{it}, \theta_{ot}, \theta_{ws},$	θ_{it}	$\theta_{ot}, \theta_{it}, \theta_{ws}$	50.50	56.89	6.39
3	θ_{sh}	θ_{sr}, θ_{it}	$\theta_{ot}, \theta_{it}, \theta_{ws}$	52.67	56.44	3.77
4		θ_{it}, θ_{PM}	θ_{it}	53.99	60.44	7.11
5	$\theta_{it}, \theta_{ot}, \theta_{ws},$	θ_{it}	$\theta_{ot}, \theta_{it}, \theta_{ws}$	52.78	58.89	6.11
6	θ_{sh}, θ_{oh}	θ_{it}, θ_{PM}	$\theta_{ot}, \theta_{it}, \theta_{ws}$	52.89	56.89	4.00
7	$\theta_{it}, \theta_{ot}, \theta_{ws},$	θ_{it}	θ_{it}, θ_{ws}	54.22	54.00	-0.22
8	$\theta_{sh}, \theta_{oh}, \theta_{PM}$	θ_{it}	$\theta_{ot}, \theta_{it}, \theta_{ws}$	52.79	58.82	6.03

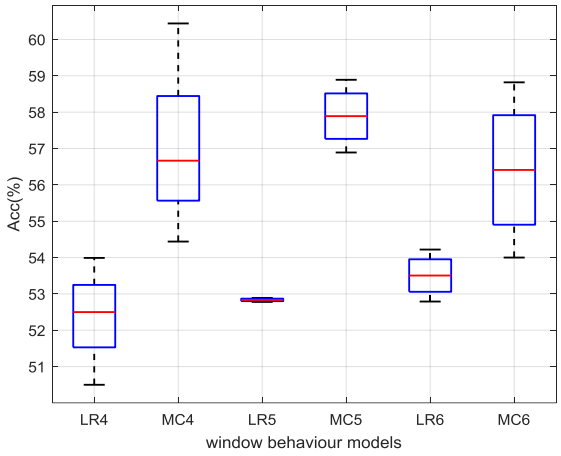


Fig. 4-6: Comparison of ACC values of logistic regression models and Markov models

4.6 ANN model

In this case study, the size of input layer was four, five or six neurons while the size of output was one neuron. An optimal number of hidden layers and neurons for each hidden layer are critical, because it will affect the speed of convergence and generalization. According to the analysis of the influencing factors of window opening behaviour (taking the size of input was six as an example), the BP neural network is designed as 6-Y-1 or 6-Y-Z-1. That is, the input layer has 6 nodes, hidden layer has Y nodes, the output layer has 1 node, or the input layer has 6 nodes, and the first hidden layer has Y nodes, and the second hidden layer has Z nodes, and the output layer has 1 node. In case of this study, single hidden-layer model was trained and validated with the number of hidden layer nodes was 5, 10, 15, 20 and 25, respectively. Accordingly, double hidden-layer model was also trained and validated with the number of nodes in each hidden layer to be 5, 10, 15, 20 and 25, respectively.

Figure 4-7 and **Fig. 4-8** show the value of MSE and ACC in testing datasets for all structures of neural networks. Although the ACC of BP network 6-15-25-1 is highest, the overall performance is not stable as 6-25-25-1. It is shown that MSE is the lowest when the neurons of first hidden-layer and second hidden layer are both 25, which is regarded as the best structure of the model. Besides, we found that 3600 datasets are the smallest amount of training data for the ANN model of window states. The model would become unstable and fail to predict the windowing behaviour if the amount of training data further reduced.

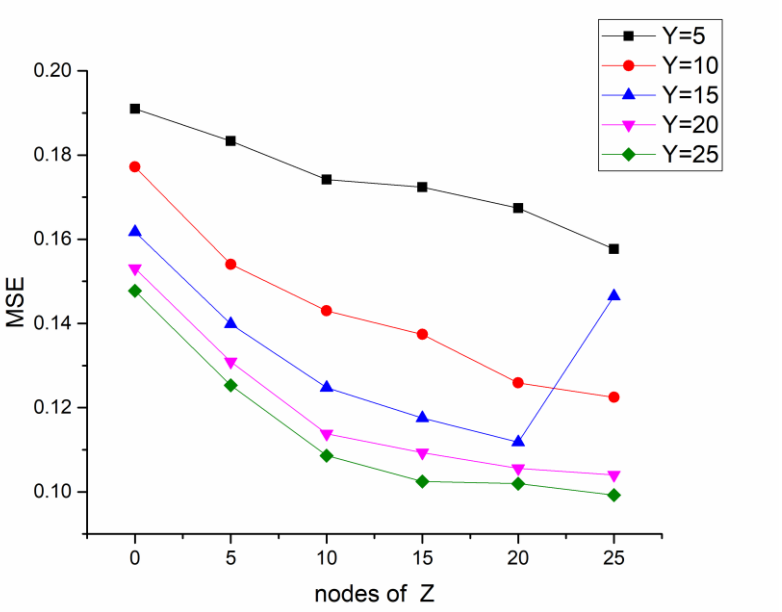


Fig. 4-7: MSE value of BP network with different number of neurons in the hidden layer

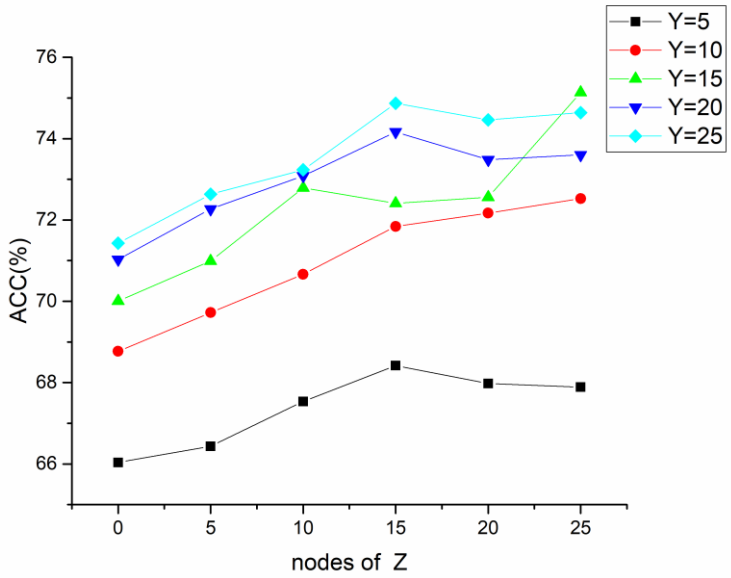


Fig. 4-8: ACC value of BP network with different number of neurons in the hidden layer

4.7 Gauss distribution model

Basic information about Gauss distribution models has been presented in Section 4.3.4. When modeling occupant window behaviour, another issue to consider is the selection of input parameters, as inter-correlated input parameters both decrease the modeling accuracy and increase the computational time. To handle this issue and remove insignificant variables, the stepwise regression method was adopted to determine input parameters for the Gauss distribution model.

The stepwise regression method has been used in many fields, and it introduces variables one by one and test the imported variables simultaneously. If the introduction of a new variable makes any existing variables less significant, that

existing variable(s) is eliminated to ensure that only significant variables are included in the regression model. This process is repeated until no more input variables are to be added or eliminated from the regression model. In this study, SPSS22, a professional statistical package, was used to for the model development.

The distribution of the datasets is shown in **Fig. 4-9** and **Fig. 4-10**. It can be seen that the datasets from two transitional seasons cover a similar indoor/outdoor temperature range. In addition, p-p diagram generated by SPSS22 was used to validate that the datasets obey normal distribution if majority of them are approximately located near the diagonal line. The results from **Fig. 4-11** show that the both indoor and outdoor temperature datasets approximately follow a normal distribution. On the other hand, it also reflects rationality of data which was selected in random.

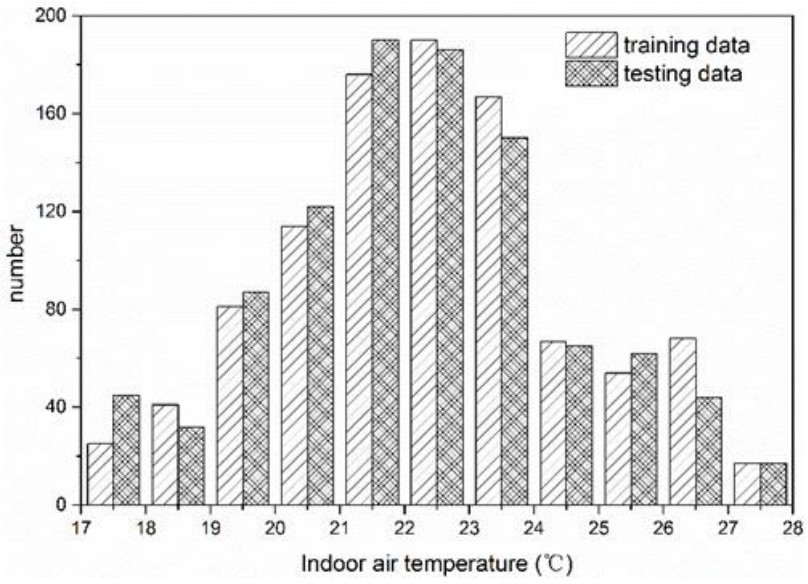


Fig. 4-9: Distribution of indoor temperature for two datasets

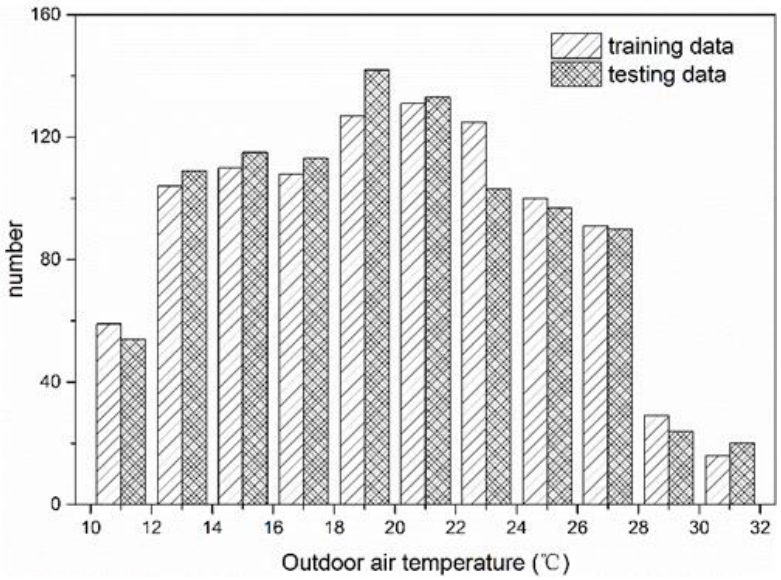
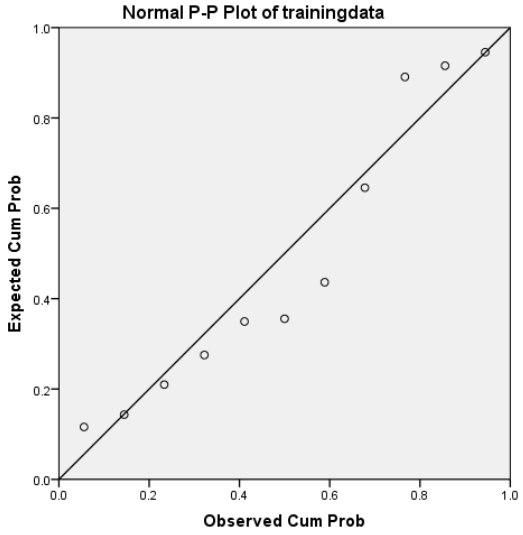
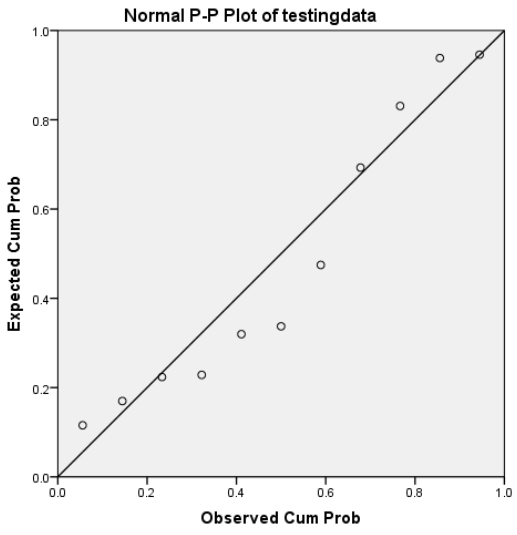
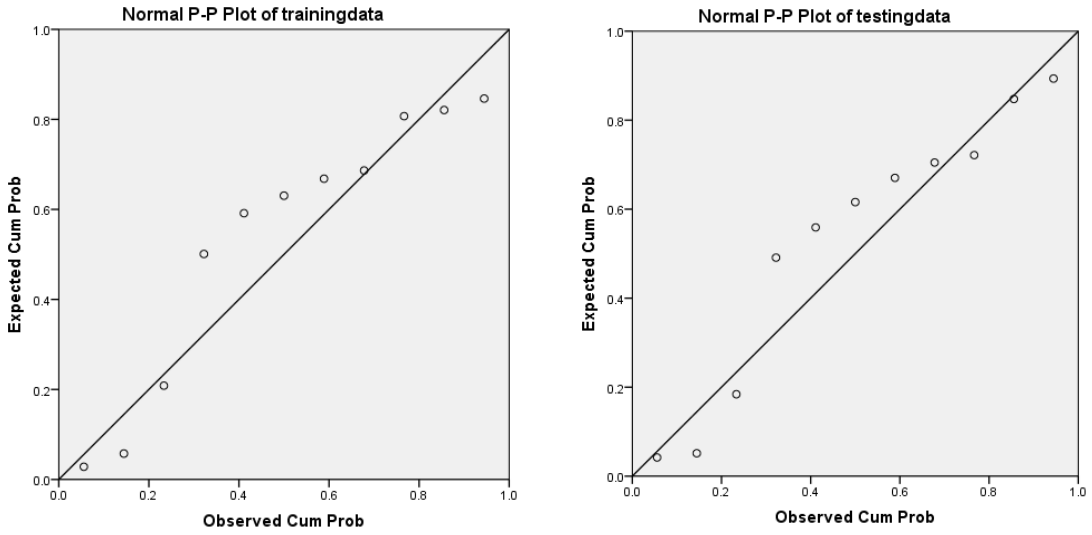


Fig. 4-10: Distribution of outdoor temperature for two datasets





(a) *(b)*



(c) *(d)*

Fig. 4-11: normal distribution test of datasets for indoor temperature (a/b) and outdoor temperature (c/d)

Except accuracy (ACC), true positive rate (TPR) and true negative rate (TNR) were also adopted as testing criteria to assess the performance of the model, shown as Eq. (4.9)-(4.10):

$$TPR = c/m \tag{4.9}$$

$$TNR = d/n \tag{4.10}$$

where c is the number of correctly predicted open states; d is the number of correctly predicted closed states; m is the total number of open states; and n is the total number of closed states.

The training dataset introduced in Section 4.1 was used to develop the Gauss distribution model. The approach introduced in Section 3.1 was used to select input parameters for the model using SPSS22, with results shown in **Table 4-8**.

Table 4-8: Results of stepwise regression

Variables	**Beta**	**T**	**Sig.**	**Partial correlation**	**Tolerance**
indoor temperature(℃)	0.006	0.105	0.916	0.005	0.881
outdoor air humidity(%)	-0.046	-0.937	0.349	-0.048	0.996
outdoor $PM_{2.5}$ concentration($\mu g/m^3$)	-0.155	-2.833	0.105	-0.143	0.800
solar radiation(W/m^3)	0.034	0.680	0.497	0.035	0.984
sunshine hours(h)	0.302	6.167	0.015	0.299	0.927
wind speed(m/s)	-0.027	-0.553	0.580	-0.028	1.000
wind direction(°)	-0.058	-1.158	0.248	-0.059	0.968
	B	T	Sig.	S.E	
Outdoor temperature	0.023	4.738	<0.001	0.005	

where B was regression coefficient; Beta referred to standardized regression coefficient and the magnitude of its absolute value directly reflected the influence of independent variable on dependent variable; t was the result of hypothesis testing on B/Beta; the Sig. value was the probability value corresponding to t and when the Sig.

value was less than 0.05, independent variable was considered as significant; The partial correlation referred to the correlation between one variable that excludes the influence of other independent variables and Y that excludes the part that other independent variables can explain; Tolerance was used to test multicollinearity between variables, and when it was less than 0.1, there was multicollinearity between variables; S.E represented standard error and reflected the degree of dispersion between sample means.

It can be seen from **Table 4-8** that except for outdoor temperature, the Sig. values of all other variables were greater than 0.05, meaning that the paramount environmental factor influencing occupant window behaviour was outdoor temperature for the selected dataset. This could be explained by the use of natural ventilation, which is highly dependent on the outdoor conditions. However, in many existing studies, indoor temperature has been identified as another factor influencing occupant window behaviour in buildings. Therefore, indoor temperature has been used to train the model as well. When using different independent variables, the main parameters of the trained Gauss distribution models are listed in **Table 4-9**. And the test results from using different independent variables as input parameters for Gauss distribution models to predict window state are as shown in **Table 4-10**.

Table 4-9: Corresponding mean value and variance of Gauss distribution models

Model	μ	σ^2
Gauss dist. (with θ_{in})	28	8

Gauss dist. (with θ_{out})	32	10
Gauss dist. (with θ_{in} and θ_{out})	28/32	8/10

Table 4-10: Predicting results of Gauss distribution model based on validation datasets

Model	TPR(%)	TNR(%)	ACC(%)
Exact	100.0	100.0	100.0
Gauss dist.(with θ_{in})	32.1	66.6	58.3
Gauss dist.(with θ_{out})	14.2	93.0	74.1
Gauss dist.(with θ_{in} and θ_{out})	15.8	87.8	70.5

where θ_{in} refers to indoor temperature and θ_{out} refers to outdoor temperature.

The validation of the models is based on the validation datasets. **Table 4-9** shows the prediction performance of the Gauss distribution models discussed above when using different input parameters. Outdoor temperature is still shown to be the most significant input parameter, giving the highest prediction accuracy, i.e. ACC equals to 74.1%, comparing to 58.3% when using indoor temperature only and 70.5% when using both indoor and outdoor temperatures.

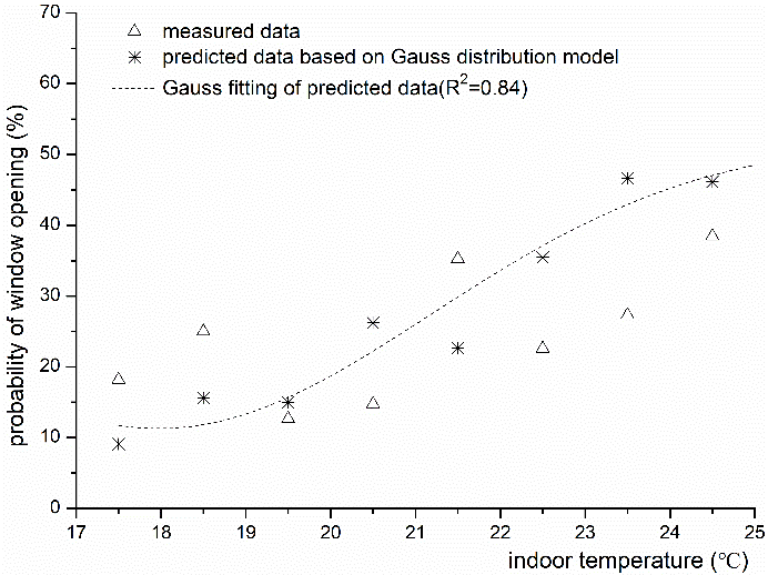


Fig. 4-12: Results of Gauss distribution model using θ_{in} as input

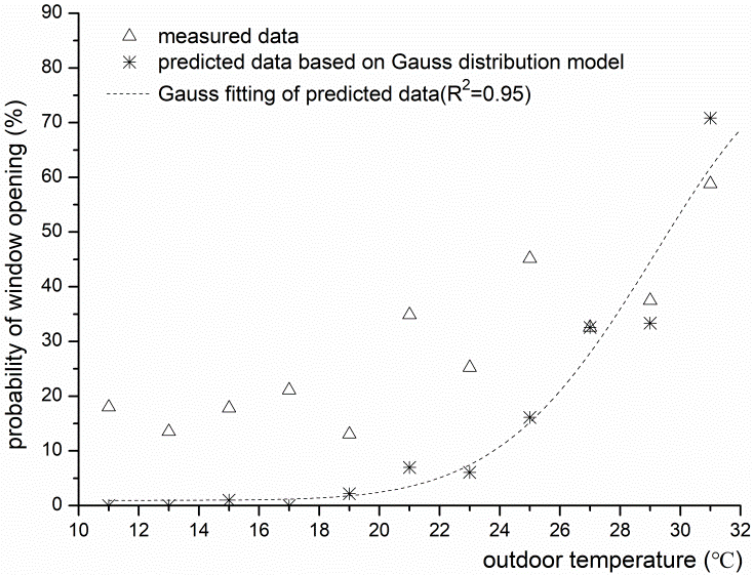
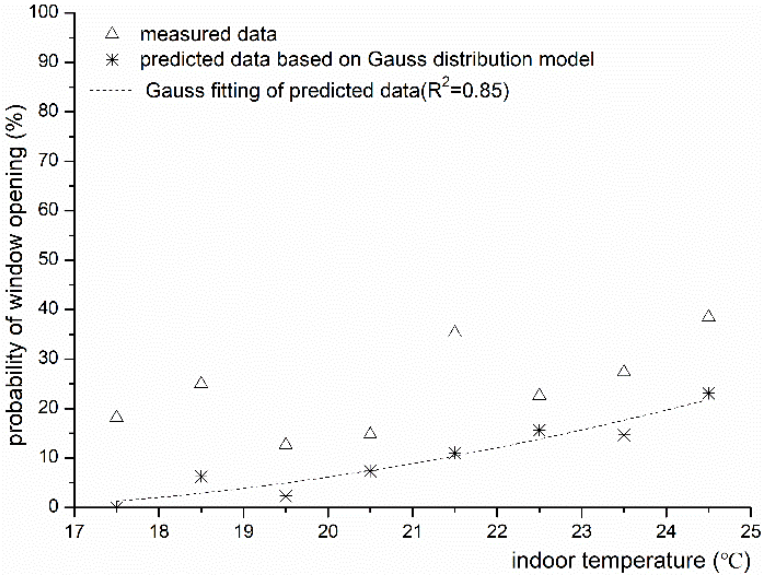
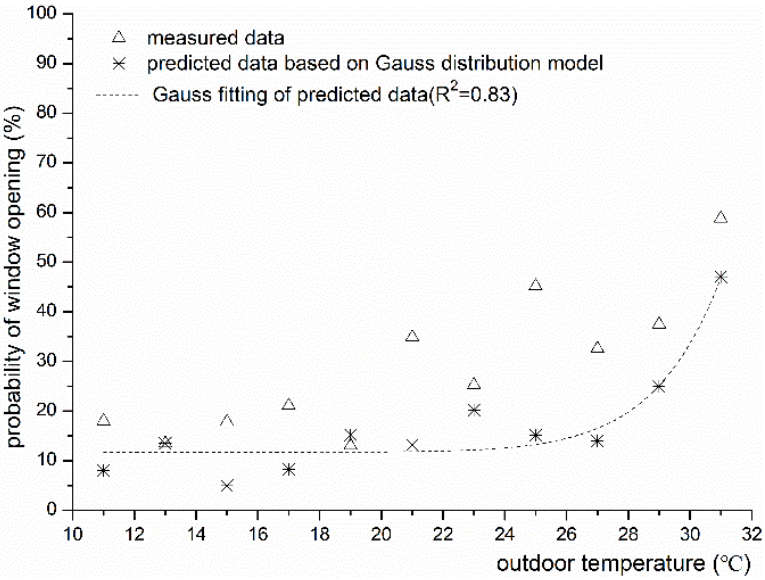


Fig. 4-13: Results of Gauss distribution model using θ_{out} as input



(a)



(b)

Fig. 4-14: Results of Gauss distribution model using both θ_{in} (a) and θ_{out} (b) as inputs

Figure 4-12 and **Fig. 4-13** show the results of Gauss distribution models whose input variables are indoor temperature and outdoor temperature, respectively. When the inputs are two variables, the probability of window opening for each variable is calculated and exhibited in **Fig. 14 (a)** and **(b)**. It can be clearly seen that aiming at this measurement of window behaviour during transitional seasons, the predicted probability of window opening shows a similarly increasing tendency as the outdoor or indoor temperature, which is similar with measured probability. Compared with the results of **Fig. 4-13** and **Fig. 4-14**, the results of **Fig. 4-12** seem to show better fitting. One possible reason for this is that Gauss distribution model whose input variable is indoor temperature has a highest true positive rate than the other two models. In addition, **Fig. 4-14(a)** and **Fig. 4-14(b)** show that though there is the same tendency of probability of window opening, increasing with outdoor/indoor temperature, the predicted probability are generally lower than measured values. One possible reason is that the model predicting window state based on two variables (indoor and outdoor temperatures), hence the statistical relationship between predicting result and each single variable is weak. However, we can conclude that the predicted probability of window opening can track the measured data fairly well. Gauss distribution model has been validated as a qualified model for window state prediction.

For comparison, the same input parameters have been applied to the training dataset and corresponding Logistic regression models were developed, with key values listed in **Table 4-11**.

Table 4-11: Coefficients of Logistic regression models

Model	a_1 for θ_{in}	a_2 for θ_{out}	b
Logit dist. (with θ_{in})	0.037	0	-1.913
Logit dist. (with θ_{out})	0	0.0317	-1.688
Logit dist. (with θ_{in} and θ_{out})	-0.009	0.33	-1.535

where a_1 and a_2 refer to the coefficients for indoor temperature and outdoor temperature, respectively, and b is a constant.

Table 4-12: Predicting results of logistic regression model based on validation datasets

Model	TPR(%)	TNR(%)	ACC(%)
Exact	100.0	100.0	100.0
Logit dist. (with θ_{in})	21.7	73.8	61.3
Logit dist. (with θ_{out})	28.3	76.4	64.6
Logit dist. (with θ_{in} and θ_{out})	27.1	75.4	64.1

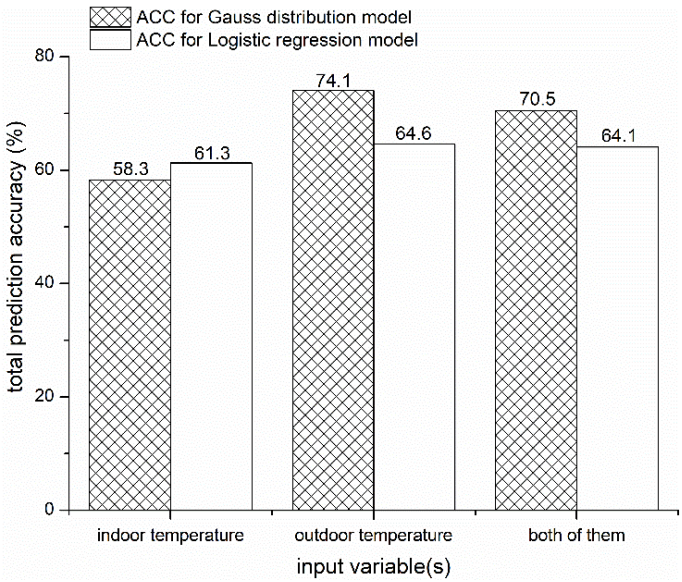


Fig. 4-15: Comparison of results between Gauss Distribution models and Logistic regression models when using different input parameters

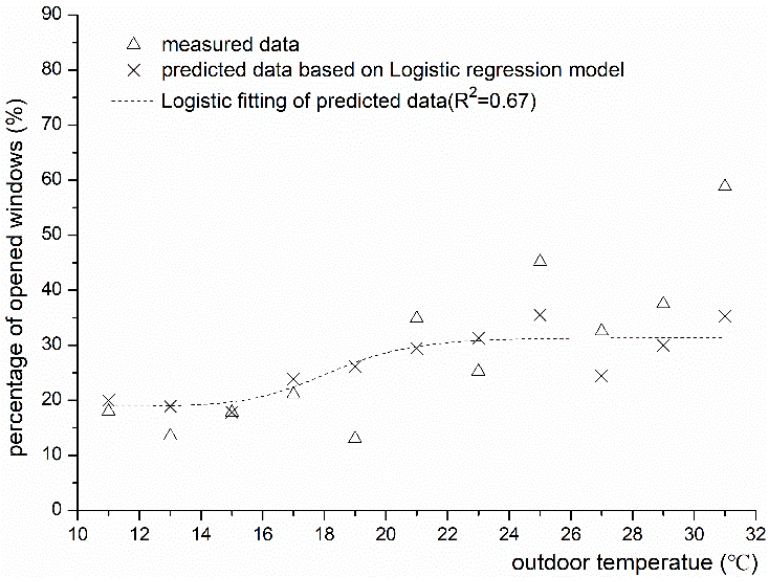


Fig. 4-16: Results of Logistic regression model using θ_{out} as input

Table 4-12 shows the prediction results when using the Logistic regression models developed for comparison. Same as Gauss distribution model, the ACC is highest for

Logistic regression models when using only one input parameter (outdoor temperature), which can be easily seen in **Fig. 4-15**. **Fig. 4-16** shows the results of Logistic regression model using optimum input (outdoor temperature). It can be seen that the performance of Logistic regression models is satisfying when outdoor temperature is low. However, the predicted accuracy is decreased with the increase of outdoor temperature. In addition, when comparing the ACC listed in **Table 4-10** and **Table 4-12**, it could be seen that when the influential factor was appropriately selected, the Gauss distribution model gave a more accurate prediction (74.1% of ACC) for window states, 9.5% higher than Logistic regression model (64.6% of ACC) when using outdoor temperature as input. However, Gauss distribution models seem to be more sensitive to input variables than Logistic regression models because the predicted accuracy of Gauss distribution models is varied from 58.3% to 74.1% while Logistic regression models give smaller variance of predicted accuracy of 61.3% to 64.6%. Therefore, the selection of input variables is significant to Gauss distribution models. It is observed that when suitable input parameters are selected, Gauss distribution models provide more accurate prediction of occupants' window behaviour than Logistic regression models.

4.8 Comparison of models

A validation procedure is involved to compare the ability of three models in predicting the window states. In **Fig. 4-17**, the obtained ACC results have been classified in four groups, including true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The logistic regression models give the much more number of FN (22%-27%) than Markov models (11%-13%) and ANN models

(10%-12%). It means several open states were identified as closed states in logistic regression models. On the contrary, the Markov models give the much more number of FP (25%-32%) than logistic models (20%-25%) and ANN models (10%-17%). It means several closed states were identified as open states in Markov models. We also see that more than 70% of true estimations can be obtained by ANN models. This comparison clearly illustrates that ANN model is a reliable method to predict the window states. In addition, the true positive rate (TPR) and true negative rate (TNR) are also adopted as criteria to evaluate the performance of the model. TPR is the number of correctly predicted open-state intervals divided by the total number of open-state intervals, and TNR is the number of correctly predicted closed-state intervals divided by the total number of closed-state intervals. The horizontal comparison of these studies is provided in **Table 4-13**. In their studies the frequency of "window closing" is much higher than the frequency of "window opening". Hence, the TNR is extremely higher. In summary, our models produce more accurate predictions in terms of TPR.

Table 4-13: Horizontal comparison of accuracy among existed model

Sources	Model	True positive rate (TPR)	True negative rate (TNR)
This study	Logistic regression model	52%	55%
	Markov model	63%	57%
	ANN model	83%	80%
Existed studies	Logistical regression model	3%	98%
	Markov model	31%	87%
	Gaussian distribution model	14%	93%
	Deep learning algorithm	37%	96%

Figure 4-18 shows the comparison of average ACC of different models, it could be found that adding outdoor humidity and $PM_{2.5}$ concentration as related factors can

increase the true estimation (TP and TN) of logistic regression models and ANN models, except Markov models. Furthermore, ANN models have extreme higher average ACC than Markov models and logistic regression models. This result demonstrates that proposed ANN approach yields a prediction model of office window state with higher accuracy and better interpretability of highly correlated factors, compared to logistic regression models and Markov models. Although some physical variables, such as outdoor humidity and $PM_{2.5}$, are identified as unnecessary contributors to occupant window behaviour by logistic regression, mild effects they explain are helpful to increase the prediction accuracy of ANN models.

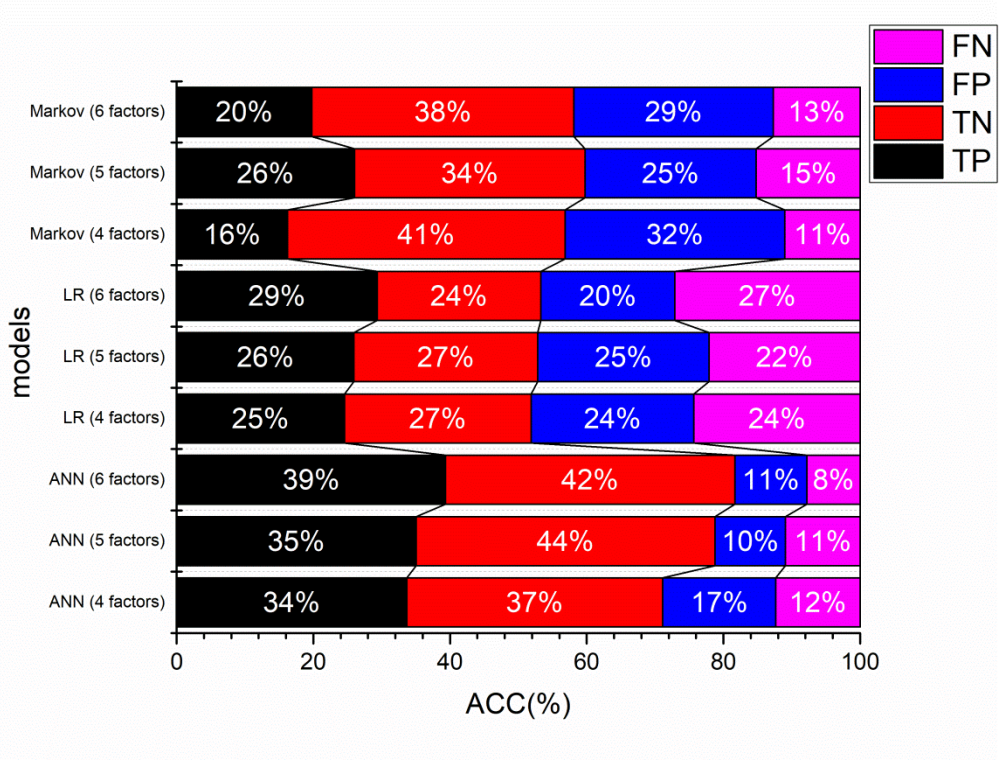


Fig. 4-17: Classification of ACC results of different models

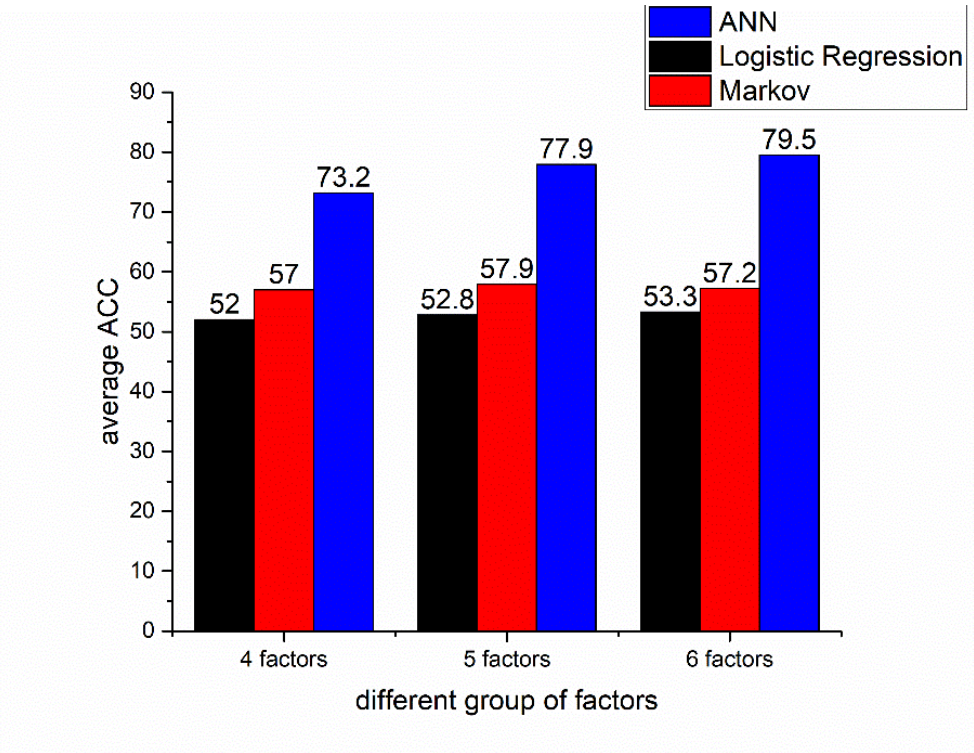


Fig. 4-18: Comparison of average ACC of different models

4.9 Chapter summary

In this research, we present logistic regression model, Markov model, Gauss distribution and ANN model as the methodologies to model window opening behaviours in office buildings. The case study is conducted based on the measured data in an office building located in Beijing, China. Specifically, 10-fold validation is conducted in logistic regression to identify the most influencing factors on window state. In discrete-time Markov model, the probabilities of window opening and closing actions have been taken into account to estimate the window state on next time step. In addition this study provided an exploration of using Gauss distribution for modelling occupant window behaviour and compared it with the more conventional method. Furthermore, the potential benefits of using ANN for modelling

the window states with consideration of comprehensive datasets were addressed through the case study.

From this work, it is generally concluded that $PM_{2.5}$ concentration and outdoor humidity should be taken into consideration in the modelling of occupant window behaviour in China area, although they are identified as unnecessary contributors by statistical analysis. Mild effects they explain on window states are benefit to increase the prediction accuracy. In addition, most methods employed to simulate window operation include logit regression and Markov chain technique. The present work is the first step in addressing the application of machine learning to model occupant's window behaviour. Especially, we can see that more true estimations can be obtained by ANN models than logistic regression model and Markov model. This comparison clearly illustrates ANN model is a reliable method to predict the window states. The proposed approaches provide a new detailed way for engineers and building operators to better understand the occupant window behaviours and their impact on energy use in office buildings.

CHAPTER 5 OCCUPANCY ESTIMATION MODEL

5.1 Overall review of occupancy estimation models

The number of occupants is not easy to count and should be measured by various sensors. The camera video and pattern recognition technique cannot be widely adopted because they are intrusive and expensive [5.1, 5.2]. According to Wang et al. [5.3], the CO_2 concentration sensor was necessary to revise the results of video camera data during the low illumination condition. Hence, it is difficult for a single video camera to obtain complete occupancy profile. The non-intrusive types of sensor, such as pyroelectric infrared (PIR) sensors and ultrasonic sensors can only be used to determine whether the room is occupied rather than the actual number of occupants [5.4, 5.5 and 5.6]. Some terminated-based methods are only effective when the occupant are using certain devises or on the seats, such as smart meters [5.7], Wi-Fi power signal [5.8], radio frequency identification tag [5.9]or chair sensors [5.10]. These sensors cannot be applied to detect the situation when people not using the devices or standing at times. Apart from terminated-based method, acoustic sensors are also used [5.11, 5.12]. However, false detection is usual because sound signals from outside can easily fool the sensor [5.13].

Besides occupancy detected sensors mentioned above, several models have been proposed to estimate the number of occupants which require sufficient training data (people-counting data). For example, researchers proposed many probabilistic approaches to estimate occupant level based on first-order Markov chain model [5.14, 5.15], inhomogeneous Markov chain model [5.16-5.18], semi-Markov model [5.19],

which generate a time series of the state of occupant presence. Additionally, non-probabilistic models were also conducted to generate occupancy profile, such as stochastic modelling [5.21, 5.21].The results illustrated that stochastic presence provides a more accurate representation of occupants' presence rather than distribution and peak values. Recently, a logistic regression model was proposed by Jie Shi et al. [5.22] to forecasting indoor occupant state. The proposed model had been validated outperform the Markov chain algorithm. However, limited to the implementation cost and ease of constancy, people-counting data used in the above studies was difficult to be collected during a long observational survey.

Occupancy estimation algorithm based on environmental measurements has been already investigated [5.23]. Mumma et al. [5.24] developed an equation to calculate the number of occupants by using CO_2 concentration. However, slow response with a time delay and difficult identification of physical parameters (e.g., door/window open) were the main drawbacks of this equation. In order to reduce the error caused by the time delay, Nishi [5.25] suggested that to reduce the number of samples when calculating the moving average value during data pre-smoothing. The authors in [5.26, 5.27] utilized sensing by proxy methodology to develop an occupancy estimation algorithm. The dispersion rate and convection effect of the CO_2 concentration and indoor air were considered in their algorithm. The results showed that the proposed algorithm outperforms a range of machine learning algorithms.

Statistical methods were also developed aiming to further understand the relationship between carbon dioxide concentration and the number of occupants. A CO_2 detection sensor network with 19 indoor sensors and 1 outdoor sensor was installed by Dong et

al. [5.28]. Support Vector Machines (SVM), Neural Networks (NN) and Hidden Markov Models (HMM) were introduced for studying the occupancy profile. Nine feature parameters of CO_2 were used to calculate the informative gain, and then most informative combinations of features were selected as inputs to the occupancy estimation models [5.29]. According to the results, HMM more realistically described the occupancy profile and levels. Different from the conventional HMM which applies time invariant transition probability matrix and mixture of Gaussian for emission probability in terms of continuous observations, inhomogeneous HMM model with multinomial logistic regression (MLR) was used by Chen et al. [5.30] to capture the temporal dependency among occupancy and environmental parameters. As for the probabilistic models, simulation in ref. [5.31] was conducted to generate a grey-box model based on CO_2 concentration measurement to estimate the occupant number for a multi-room case. In their model, the physical parameters and occupant number were estimated by Maximum Likelihood and regularized deconvolution approach [5.32]. When compared with SVM model and NN model, their estimator had the best performance indexes. However, all these statistical methods mentioned above request both people-counting data and environmental parameters as training data sets, meaning that comprehensive data has to be measured and collected.

Besides continuous number of occupants, subranges of occupancy level were determined by classification algorithms from CO_2 concentration [5.33, 5.34]. For example, Decision Tree (DT) had been used by Hailemariam et al. [3.35] to detect room status (binary detection). The other classification model, such as Random Forest (RF), Gradient Boosting Machines (GBM), Linear Discriminant Analysis (LDA) and

Classification and Regression Trees (CART) were evaluated by [5.36]. The LDA model corresponded to the best accuracies in the test sets when only two predicators were input, because too many predictors which were highly correlated would cause a decrease in the accuracy. Yang [3.37] et al. developed occupancy models using SVM, NN and DT for both single-occupancy and multi-occupancy office. Different from [5.38], it was observed that the overall accuracy generally increased as the number of predictors increased. In addition, Local Receptive Fields (LRF) with random weights was adopted as feature learning by [5.39, 5.40], then the Extreme Learning Machine (ELM) classifier was trained for division of occupancy level.

A summary of the various algorithms, model types, sensors and reported accuracies of occupancy models developed based on CO_2 concentration data is presentable in **Table 5.1**.

Table 5-1: Algorithms, model types, sensors and reported accuracies of occupancy models developed based on CO_2 concentration data

Source	Model employed	Model type	sensors	Calculation accuracy
[5.24]	massive conservation equation	continuous number	CO_2 concentration	NA
[5.26, 5.27]	Sensing by proxy	continuous number	CO_2 concentration	RMSE of 0.6311
[5.28]	SVM, NN, HMM	continuous number	A CO_2 detection sensor network(19 inside, 1 outside)	61%-75%
[5.30]	Inhomogeneous HMM	multi-class estimation	CO_2 concentration, relative humidity, temperature, air	75%-78%

			pressure	
[5.29]	FFNN	continuous number	CO_2 concentration, sound, case temperature, PIR sensors	67%-69%
[5.38]	STD	continuous number	CO_2 concentration	94.68%
[5.35]	DT	binary detection	CO_2 concentration, computer current, light, PIR, sound	94.68% for only CO_2
[5.36]	GBM, LDA, CART	binary detection	Temperature, humidity, light, CO_2 concentration, and humidity ratio.	32.68%-99.32%
[5.34]	RBF, SVM, RF, NB	binary detection	Temperature, humidity, CO2, sound, pressure and illumination (light) sensors	96%-99%
[5.37]	SVM, NN, DT	binary detection & multi-class estimation	CO_2 concentration, reed sensor for door, relative humidity, temperature, light, sound, PIR	88.9%-98.2% for DT 66.36%-89.86% for only CO_2
[5.40]	ELM	multi-class estimation	CO_2 concentration, relative humidity, temperature, air pressure	74.5%
[5.39]	ELM	continuous number	CO_2 concentration	50%
[5.32]	LTI	continuous number	CO_2 concentration	Above 82%

After reviewing the peer researches, a few key points are identified:

- Various applications of abovementioned approaches need careful training process where both people-counting data and environmental parameters are necessary. Collection of comprehensive data might be expensive or not feasible, especially for small time intervals.
- The occupancy estimation algorithm developed based on analytical method is not as accurate as other models, because the identification of physical

parameters (both measured parameters and unmeasured parameters) is not evaluated during calculation.

- The models for binary detection for occupancy profile do not exploit additional capability for estimating occupant numbers. In other words, the simple occupancy profile (e.g. occupied/unoccupied) gives limited information for building energy efficiency.

In this thesis, the occupancy model is proposed with the aim to providing a non-intrusive and accurate algorithm to estimate the number of occupants in office room. Different from most of works which depended on training sets, this model could be used to blindly compute the number of occupants. The measured true occupancy is used for model validation. Consequently, the calculation accuracy of the occupancy model would be identified for comparison purpose.

5.2 Methodologies for occupancy estimation

In this section, we address the methodologies to identify the relationship between number of occupants and indoor CO_2 concentrations. Parameter estimation is also called blind system identification (BSI), the word blind means that the system's inputs are not available to the system [5.41]. It is a technology aimed at retrieving a system's unknown information from its outputs only. The implementation of BSI is to estimate any unknown input signals and noise variable based on the likelihood function of given output data. It has been applied in seismic community, medical community and etc.

In this case, the occupant number is estimated from parameter estimation algorithm based on the distribution of CO_2 concentration data. The calculation is completed with the help of mass-conservation equation of CO_2 concentration. Normally, the occupant number can be roughly estimated from mass-conservation equation based on analytical method. The difference is that the measure error term and prior experience of parameters are considered in our model. The calculation process is regarded as the optimization problem solved iteratively until accurate solution is found [5.42]. The superiority of parameter estimation algorithm is that the calculations are completed automatic without parameter tuning or training process. In addition, the estimation result is more reliable than analytical method because measure error term and prior experience of parameters are considered during calculation.

Among different estimation approaches, frequentist ML and Bayesian estimation are adopted to occupancy estimation problem. The frequentist ML method involves calculating unknown parameters in a distribution (e.g. mean value, variance value, etc.) that are most likely to yield the observed data sets. For small sample sizes, the frequentist ML does not always provide satisfied results [5.43] because the parameter estimation is absolutely dependent on the observed data sets. Bayesian estimation is another parameter estimation algorithm which is designed to deal with this situation. In Bayesian estimation, prior distribution for the parameters, to specify the initial state of knowledge about them, before the observed data are used for calculation. For example, different factors that influences on indoor CO_2 concentration are initially assigned with the significance weights based on prior knowledge. The aim is to obtain an unprejudiced estimation result from limited amount of data sets [5.44]. Thus, the

initialization of unknown parameter is fundamental in the Bayesian estimation algorithm.

5.2.1 Frequentist Maximum Likelihood (ML) approach

The physical-based model adopted in this research is based on the following assumptions:

- Air is well-mixed [5.45], and therefore the actual CO_2 concentration of the considered room is denoted as $\bar{C}(t), \bar{C}(t) \in R$.
- Outdoor air CO_2 concentration is assumed constant as C and equals to 420ppm [5.46, 5.47].
- The time domain of the signals is discrete and finite.
- The fresh air system keeps a continuous fresh air flow in the room.
- CO_2 generation rate per person, g, is assumed to be 0.005L/S as the adult initially [5.48,5.49].

In this study, different assumptive values of outdoor CO_2 concentration are tested because of the measurement lack. The outdoor CO_2 concentration is assumed as 420ppm where the calculation error is least. With these assumptions, the variation of indoor CO_2 concentration is, based on mass-conservation law, mathematically derived as follows:

$$\frac{d\bar{C}(t)}{dt} = \frac{\dot{Q}^{vent,sup}(t)+\dot{Q}^{leak,in}(t)}{V} C - \frac{\dot{Q}^{vent,exh}(t)+\dot{Q}^{leak,out}(t)}{V} \bar{C}(t) + \frac{g}{V} O(t) \quad (5.1)$$

Where the V is the volume of considered room; $\dot{Q}^{vent,sup}(t)$ and $\dot{Q}^{vent,exh}(t)$ are the supply fresh air rate and exhaust mechanical ventilation rate; $\dot{Q}^{leak,in}(t)$ and $\dot{Q}^{leak,out}(t)$ are the inflow and outflow of leakages from open windows and doors; $O(t)$ is the number of occupants at time t and $O(t) \in N_+$.

According to ASHRAE standard [5.49], the number of occupants could be roughly estimated by equation (1), the unknown variables such as $\dot{Q}^{leak,in}(t)$, $\dot{Q}^{leak,out}(t)$ and g are fixed or looked up from appendix [5.49]. However, the calculation accuracy of analytical method is not satisfactory as other methods mentioned in the literature.

In this model, we also assume the balanced ventilation of the considered room where $\dot{Q}^{vent,sup}(t) \approx \dot{Q}^{vent,exh}(t) \approx \dot{Q}^{vent}(t)$ and $\dot{Q}^{leak,in}(t) \approx \dot{Q}^{leak,out}(t) \approx \dot{Q}^{leak}(t)$. The measured CO_2 concentration suffers from serious spikes occasionally, which are caused by measurement noise, random air movement or occupants' breath. Hence, the measured CO_2 concentration, denoted by $y(t)$, is expressed as $y(t) = \bar{C}(t) + e(t)$, where the measurement error $e(t) \sim \mathcal{N}(0, \sigma^2)$ is the Gaussian white noise. Specifically, unknown variance σ^2 is introduced to describe the difference between measured CO_2 concentration data and actual CO_2 concentration data. The structure form of dynamics (5.1) can be simplified as:

$$\frac{dy(t)}{dt} = \frac{\dot{Q}^{vent}(t)+\dot{Q}^{leak}(t)}{V}\left(C - \bar{C}(t)\right) + \frac{g}{V}O(t) + e(t) \tag{5.2}$$

Then, we discretize the continuous-time model into discrete-time model by standard backward Euler discretization, we can rewrite the dynamics (5.2) as follows:

$$\frac{y(k)-y(k-1)}{T} = \frac{\dot{Q}^{vent}(k-1)+\dot{Q}^{leak}(k-1)}{V}\left(C - y(k-1)\right) + \frac{g}{V}O(k-1) + e(k-1) \quad (5.3)$$

Where T is the sampling time and k is the discrete time domain, $k = 1,2 \ldots n$.

Next, define the signals as:

$$\mathbf{y}: = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(n) \end{bmatrix}, \mathbf{O}: = \begin{bmatrix} O(1) \\ O(2) \\ \vdots \\ O(n) \end{bmatrix}, \mathbf{e}: = \begin{bmatrix} e(1) \\ e(2) \\ \vdots \\ e(n) \end{bmatrix} \quad (5.4)$$

Equation (5.3) shall be used to calculate the real-time number of occupants in this model. The dynamic (5.3) can by rewritten as following:

$$(I - a\Delta)\mathbf{y} = (1-a)C + b_o\Delta\mathbf{O} + \mathbf{e} \quad (5.5)$$

Where $a = 1 - \frac{\dot{Q}^{vent}(k)+\dot{Q}^{leak}(k)}{V}T$， $b_o = \frac{g}{V}T$, I is the n-dimensional identify matrix, Δ is the n-1-dimentional identify matrix.

$$\Delta: \ = \begin{bmatrix} 0 & \ldots & 0 \\ & & \vdots \\ I_{N-1} & & 0 \end{bmatrix} \in \mathbb{R}^{n \times n} \quad (5.6)$$

Hence, the indoor measured CO_2 concentration can be expressed as:

$$\mathbf{y} = (I - a\Delta)^{-1}(1-a)C + (I - a\Delta)^{-1}b_o\Delta\mathbf{O} + (I - a\Delta)^{-1}\mathbf{e} \quad (5.7)$$

In Eq. (5.7), $\mathbf{y}$ and C are known parameters, $a, b_o, \mathbf{O}$ and $\mathbf{e}$ are the unknown parameters which need to be estimated in the next step.

What should be remarked is that Eq. (5.7) is a simplified model which does not account for the changes on parameters a and b_o due to human activities with time

variations. In this study, the simplification is reasonable as the human activity of all office workers does not vary largely. However, in other types of buildings, these non-ideal issues would be considered.

In this model, let the unknown parameters which reflect the physical property of the considered room and occupant number, represented as θ, as

$$\theta \coloneqq [a\ b_o\ \sigma^2\ O(1)\ O(2)\ \dots O(n)]^T \tag{5.8}$$

Since Eq. (5.7) follows the form as $\mathbf{y} = \mathbf{A} + B\mathbf{e}$ and measurement error $\mathbf{e}$ is assumed as $\mathcal{N}(0, \sigma^2)$, measured CO_2 concentration $\mathbf{y}$ follows the distribution function $\mathcal{N}(\mu, \sigma^2)$, which is expressed as:

$$p(\mathbf{y}; \theta) = \mathcal{N}(m_{\mathbf{y}}, cov_{\mathbf{y}}) \tag{5.9}$$

Where m_y is the expected value and cov_y is the variance value, both of them are expressed as:

$$\begin{cases} m_y = (I - a\Delta)^{-1}(1 - a)C + (I - a\Delta)^{-1}b_o\Delta\mathbf{O} \\ cov_{\mathbf{y}} = \sigma^2(I - a\Delta)^{-1}(I - a\Delta)^{-T} \end{cases} \tag{5.10}$$

When using normal distribution to describe a given independent vector, the likelihood function is adopted to estimate the unknown parameters, showed as:

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \bar{x}_i)^2}{2\sigma^2}} \tag{5.11}$$

Introduce the log-likelihood function as:

$$\log L(\theta) = -\frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum(x_i - \bar{x}_i)^2 \tag{5.12}$$

For this model, the log-likelihood function of equation (5.12) can be expressed as:

$$\log L(\theta) = \log\left(\mathrm{cov}_{\mathbf{y}}\right) + \frac{1}{2\sigma^2}\mathbf{e}^{\mathrm{T}}\mathbf{e} \tag{5.13}$$

Thus, the realization of ML estimation can be regarded as optimization of equation (3.19). The execution of the operation can be carried out by iterating of following two steps in MATLAB: (1) assign the initial values for parameter $[a, b_o, \sigma^2]$, calculate the occupancy $\mathbf{O}$ by ML estimation; (2) fix the occupancy $\mathbf{O}$ and recalculate the parameter $[a, b_o, \sigma^2]$; stop calculation when required minimum cost function is reached.

Partial code programming could be found in the appendix.

5.2.2 Bayesian estimation approach

In the Bayesian estimation, we consider the occupancy estimation problem as a multiple-input and single-output (MISO) linear time-invariant discrete-time dynamic system as shown in **Fig. 5.1** and equation (5.14).

$$y_{(t)} = \sum_{i=1}^{+\infty} g_{yi}\, y_{(t-i)} + \sum_{i=1}^{+\infty} g_{ui}\, u_{(t-i)} + \sum_{i=1}^{+\infty} g_{oi}\, o_{(t-i)} + e_{(t)} \tag{5.14}$$

Where $\sum_{i=1}^{+\infty} g_{yi}$, $\sum_{i=1}^{+\infty} g_{ui}$ and $\sum_{i=1}^{+\infty} g_{oi}$ are the transfer function reflecting the dynamics of the model driven by the input signals. $y_{(t)}$ is the measured CO_2 concentration level, $u_{(t)}$is the actual level of fresh air system which is presented as electricity consumption, kWh, $o_{(t)}$ represent the number of occupants in the room at each time instant, $e_{(t)}$ is the zero-mean Gaussian white noise with unknown variance σ^2.

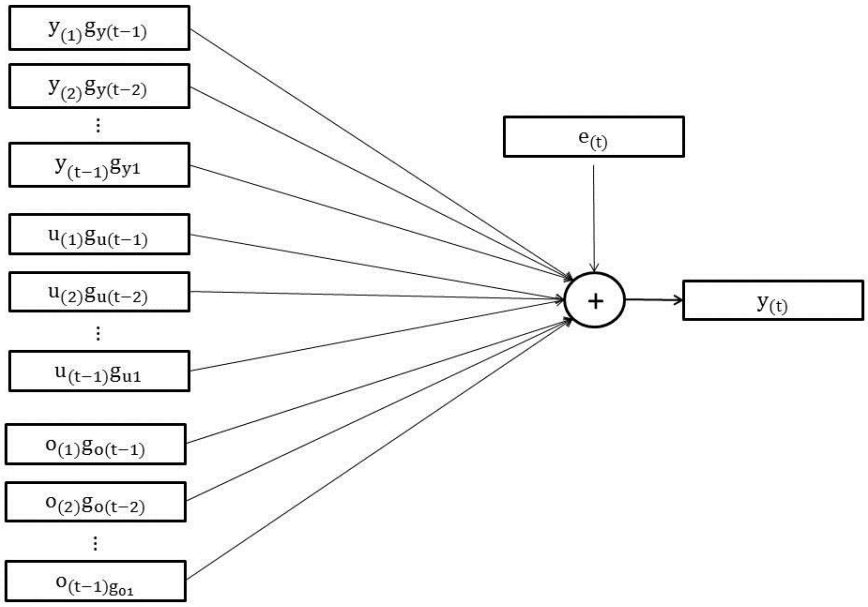


Fig. 5-1: block scheme of the MISO model

Define the following vectors as:

$$\mathbf{y}:=\begin{bmatrix}\mathrm{y}(1)\\ \mathrm{y}(2)\\ \vdots\\ \mathrm{y}(\mathrm{n})\end{bmatrix},\ \mathbf{u}:=\begin{bmatrix}\mathrm{u}(1)\\ \mathrm{u}(2)\\ \vdots\\ \mathrm{u}(\mathrm{n})\end{bmatrix},\ \mathbf{o}:=\begin{bmatrix}\mathrm{o}(1)\\ \mathrm{o}(2)\\ \vdots\\ \mathrm{o}(\mathrm{n})\end{bmatrix},\ \mathbf{e}:=\begin{bmatrix}\mathrm{e}(1)\\ \mathrm{e}(2)\\ \vdots\\ \mathrm{e}(\mathrm{n})\end{bmatrix} \tag{5.15}$$

Then the operator $\mathbf{T}_n(\cdot)$ is used to map above vectors as the $n \times n$ Toeplitz matrix, e.g.,

$$\mathbf{T}_n(\boldsymbol{y})=\begin{bmatrix} y_o & 0 & & 0\\ y_1 & y_o & \cdots & 0\\ \vdots & y_1 & \cdots & 0\\ y_{n-2} & \vdots & \cdots & \vdots\\ y_{n-1} & y_{n-2} & & y_o \end{bmatrix} \in \mathbb{R}^{n\times n}$$

After revise all the input signals **y**, **u** and **o** into the Toeplitz matrix as $\mathbf{T}_\mathrm{n}(\mathbf{y})$, $\mathbf{T}_\mathrm{n}(\mathbf{u})$ and $\mathbf{T}_\mathrm{n}(\mathbf{o})$, we reserve symbol **Y**, **U** and **O** for the matrix $\mathbf{T}_\mathrm{n}(\mathbf{y})$, $\mathbf{T}_\mathrm{n}(\mathbf{u})$ and $\mathbf{T}_\mathrm{n}(\mathbf{o})$ respectively, then equation (5.15) can be rewritten as

$$\mathbf{y} = \mathbf{Y}\mathbf{g}_y + \mathbf{U}\mathbf{g}_u + \mathbf{O}\mathbf{g}_o + \mathbf{e} \tag{5.16}$$

The transfer function $\mathbf{g}_y, \mathbf{g}_u$ and $\mathbf{g}_o$ follow the Gaussian distribution as [5.50]:

$$\mathbf{g}_y \sim \mathcal{N}(0, \lambda_y K_{\beta y})$$

$$\mathbf{g}_u \sim \mathcal{N}(0, \lambda_u K_{\beta u})$$

$$\mathbf{g}_o \sim \mathcal{N}(0, \lambda_o K_{\beta o}) \tag{5.17}$$

Where $K_{\beta y}$, $K_{\beta u}$ and $K_{\beta o}$ are $n \times n$ covariance matrix obeying Weibull distribution, whose structure depends on the shaping parameter βy, βu and βo respectively, which regulate how fast the signal decay in time series. βy, βu and βo are the scalar in the interval [0,1).The scaling factors λ_y, λ_u and λ_o tune the amplitude of the responses from the input signals.

Then, define the unknown parameters as:

$$\theta \coloneqq \left[\mathbf{o}(t) \;\; \beta y \;\; \beta u \;\; \beta o \;\; \lambda_y \;\; \lambda_u \;\; \lambda_o \;\; \sigma^2 \right]^T \tag{5.18}$$

The $\mathbf{g}_y$ follows the posterior distribution of given y and θ as Gaussian distribution, namely

$$P(\mathbf{g}_y | y, \theta) = \mathcal{N}(C_y, P_y) \tag{5.19}$$

Where $P_y = (\frac{\mathbf{Y^T Y}}{\sigma^2} + {K_{\beta y}}^{-1})^{-1}$, $C_y = P_y \frac{\mathbf{Y^T}}{\sigma^2}$. From (5.18) the transfer function $\mathbf{g}_y$ can be estimated as:

$$\hat{\mathbf{g}}_y = E\left(\mathbf{g}_y \middle| y, \theta \right) = C_y y \tag{5.20}$$

Similarly, the other two transfer function $\mathbf{g}_u$ and $\mathbf{g}_o$ can also be derived as (5.20). Clearly, such estimators are carried out based on the function of θ .Thus, the initialization of unknown parameter θ is fundamental in the Bayesian estimation algorithm.

The method to estimate θ is to maximization of the marginal likelihood [5.51] as follows:

$$\hat{\theta} = \arg\max_{\theta} \log p(y|\theta) \tag{5.21}$$

However, solving the nonlinear equation (5.17) in that form is difficult, because **y**, **u**, **o** and $\mathbf{g}_y$, $\mathbf{g}_u$, $\mathbf{g}_o$ are both obeyed Gaussian distribution. For this reason, an iterative solution is adopted by maximizing the complete log-likelihood:

$$L(y, g|\theta) = \log P(y, g|\theta) = \log P(y|g, \theta) + \log P(g|\theta) \tag{5.22}$$

The expansion of (3.28) is composed with eight components. Readers could refer to [5.52] for more detailed calculation.**Table 5.2** shows the steps of Bayesian estimation in MATLAB.Partial code programming could be found in the appendix.

Table 5-2: The calculation steps of Bayesian estimation.

Input:$\{y\}_{t=1}^{n}, \{u\}_{t=1}^{n}$

Output:$\{\hat{o}\}_{t=1}^{n}, \{\hat{g}_y\}_{t=1}^{n}, \{\hat{g}_u\}_{t=1}^{n}, \{\hat{g}_o\}_{t=1}^{n}$

(1) Initialization: set value to $\hat{\theta}^0 := \left[\, o^0(t)\ \beta y^0\ \beta u^0\ \beta o^0\ \lambda_y{}^0\ \lambda_u{}^0\ \lambda_o{}^0\ \sigma^{2,0}\right]^T$

(2) Repeat until convergence:

(a) Calculate $\hat{P}_y{}^k$, $\hat{C}_y{}^k$, $\hat{\mathbf{g}}_y{}^k$, $\hat{P}_u{}^k$, $\hat{C}_u{}^k$, $\hat{\mathbf{g}}_u{}^k$, $\hat{P}_o{}^k$, $\hat{C}_o{}^k$, $\hat{\mathbf{g}}_o{}^k$ from (5.19) and (5.20)

(b) Update the parameter $\hat{\theta}^{k+1}$ from maximizing (5.22)

5.3 Data sets

We tested the proposed occupancy estimators in an office room, located in a commercial building in Beijing. The conditioned floor area of the office room is 152m^2 with design occupants of 36 people, as depicted in **Fig.5-2**. Ventilation of the office room is provided by independent fresh air system combined with variable refrigerant volume (VRV) air-conditioning system. In the office, the fresh air system and VRV system were active during occupied periods. In this work, we did not measure the flow rate of the fresh air, but corresponding value is obtained indirectly from electricity consumption of fresh air system. For example, the fresh air system is set as medium /high/off state, corresponding flow rate of each state is estimated as 400m^3/h, 700m^3/h and 0m^3/h, respectively. The room is cooled by VRV system and setting value of indoor temperature is centrally controlled. What is more, there are four ceiling fans installed in the office space, occupants could use it when necessary. The occupancy schedule in weekday is from 9:00 a.m. to 18:00 p. m., Monday-Friday. The door keeps closed most of time due to the entrance guard system.

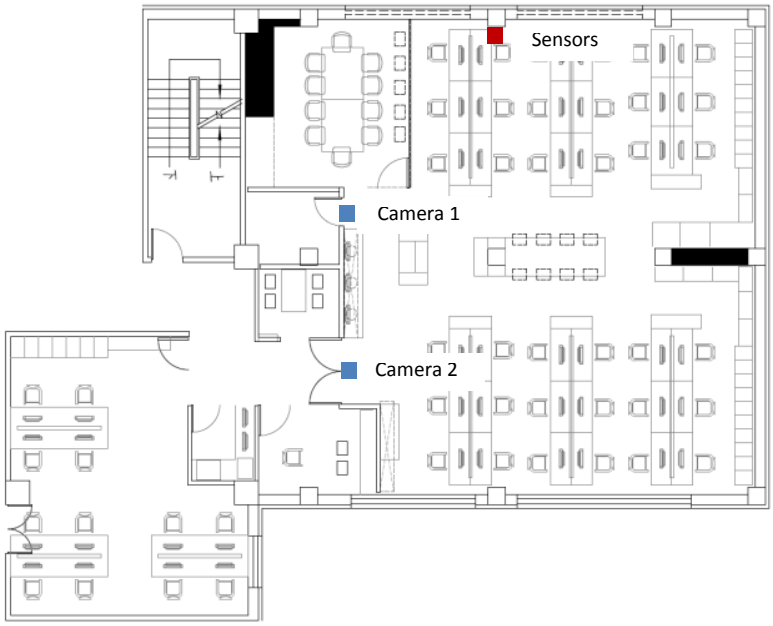


Fig. 5-2: Area of reference office room for model test

As shown in **Fig. 5-2**, the set-up for the sensor network includes:

- Two video cameras are installed facing down on two doors of the office room to record the real occupant number, which is used as the ground truth to assess the proposed models. The entering/leaving action is recorded when the two sides of target area are cut through by the people. The occupancy counts are updated only when the number of the occupants changes hence there is no fixed resolution time. The detection accuracy of the occupancy detection camera is above 90%. However, the performance of video camera was not stable during early installation stage. As discussed before, the accumulative detective error is the cause of unstable performance. Therefore, only five-day occupancy counting data is used in this study.

- An ambiance sensor network is used to measure four types of environmental parameters, including CO_2 concentrations, PM2.5, temperature and relative humidity. We select a location where the sensor is easily fixed and rarely impacted by outdoor disturbance. The height of the sensor is 1.4m just above the nose level of seated occupants. There is no clear answer for the best placement of environmental sensor for occupancy estimation. Hence the location of the sensor is out of the scope of this paper.
- Energy use data, such as the electricity consumption of fresh air system, appliances, lighting, fan and air conditioning system.
- A centralized database with the web application for continuously record data from different sources.

The sampling frequency is 1 sample per 30 minutes and there are 48 time samples in each day. The choice of sampling time is context-dependent according to the granularity required. Environmental variables recorded by the measurement sensors are accumulated over a time interval. Hence, variance feature of environmental variables do not allow arbitrarily small or extreme large duration. The studies in [5.24] and [5.25] are based on interval of 30min, which is adopted in this case.

For continuous five working days (01/15/2018-01/19/2018), 240 samples are available for calculation and validation. Occupancy data are recorded on working days from 8:00 a.m. to 7:00 p.m. **Table 5-3** presents the main features of the monitoring sensors.

Table 5-3: Monitoring equipment

parameters	Range	Accuracy	Brand
Temperature	-40-125℃	±0.5℃	SHT20
Humidity	0-100%RH	±3%RH	SHT20
CO_2	0-5000ppm	±75ppm	Telaire T6703
Occupancy	-	>90%	/

A visual demonstration of the relationship between occupancy and various environmental variables is shown in **Fig. 5-3**. It could be easily seen that CO_2 concentration tracks the true occupancy with a small response delay. The other environemntal variables, such as indoor temperature, relative humidity and PM2.5, do not provide significate insights into the occupancy level. In this model, we estimate the occupancy level using only one environmental vairables (CO_2 concentration). In **Fig. 5-4**, the distribution of measured CO_2 concentration in this case is shown.

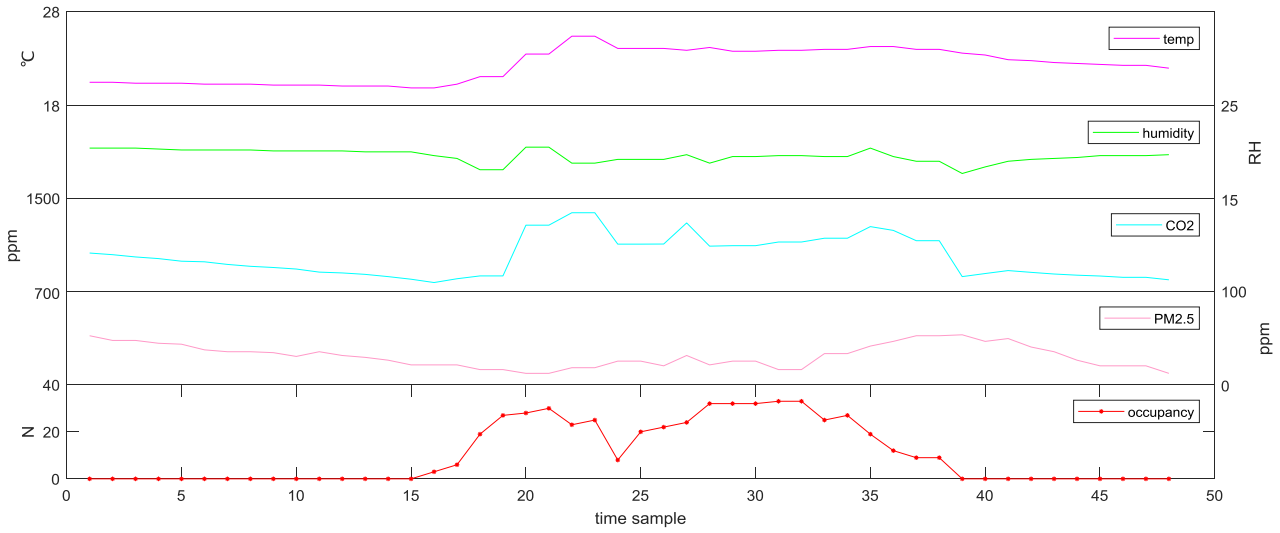


Fig. 5-3: True occupant number with corresponding environmental parameters (January 15, 2018)

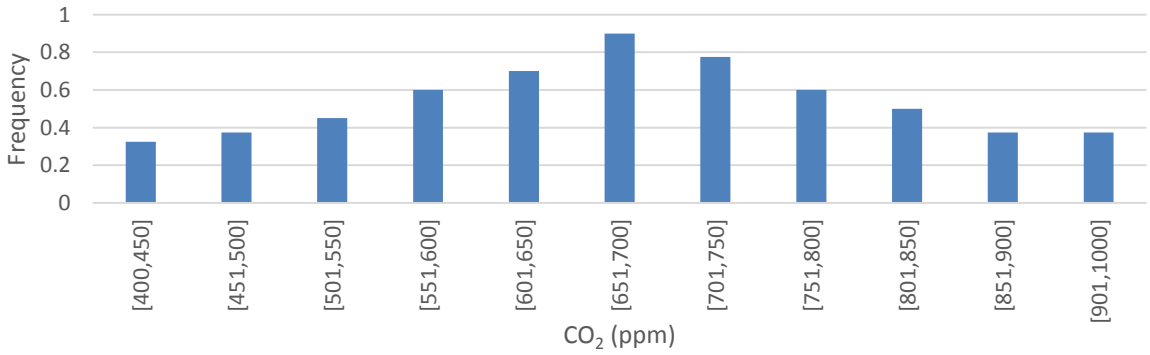


Fig. 5-4: Histograms based on CO_2 concentration considering periods when the room is positively occupied

As mentioned before, the measured CO_2 concentration with spikes cannot correctly reflect the real environmental situation of the whole room. Besides measurement error $e(t)$, data analysis for smoothed CO_2 concentration is conducted as the important data pre-processing before calculating the number of occupants.

Three methods have been adopted to smooth the CO_2 concentration:

- Two-hour moving average
- Two-hour bin
- Globally smooth

Simply saying, the moving average and bin is the unweighted mean and medium value of several continuous data. It yields that the variances with the time-serious data are smoothed and aligned. As for globally smooth method, we define the measured CO_2 data as $\mathbf{y} = [y(1), y(2) \dots y(n)]^T$. The smoothed CO_2 data $\mathbf{y}_s$ can be obtained by searching for the minimum solution of the energy function, shown in following:

$$\mathrm{E}(\mathbf{y}_s) = \|\mathbf{y} - \mathbf{y}_s\|_2^2 + \omega\|\nabla\mathbf{y}_s\|_2^2 \tag{5.23}$$

Where $\nabla\mathbf{y}_s$ is the gradient of $\mathbf{y}_s$, $\nabla = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n-1\times n}$,and ω is the regular factor to balance the weights between the two terms in the equation (5.23), the large value of ω would lead to the smoothing output of $\mathbf{y}_s$.

Set the derivative of $\mathrm{E}(\mathbf{y}_s)$ respect to $\mathbf{y}_s$ as zero, we can obtain the smoothed CO_2 data $\mathbf{y}_s$.

Figure 5-5 shows the raw data and the smoothed CO_2 concentrations filtered by three methods. As for globally smoothed data, the values of start point and end point are basically equal to the raw data. Additionally, the spikes of the raw data are obviously smoothed by two-hour moving average, two-hour bin and globally smooth. What is more importantly, the fluctuation of the raw data is shifted to an early time step by two-hour moving average and two-hour bin.

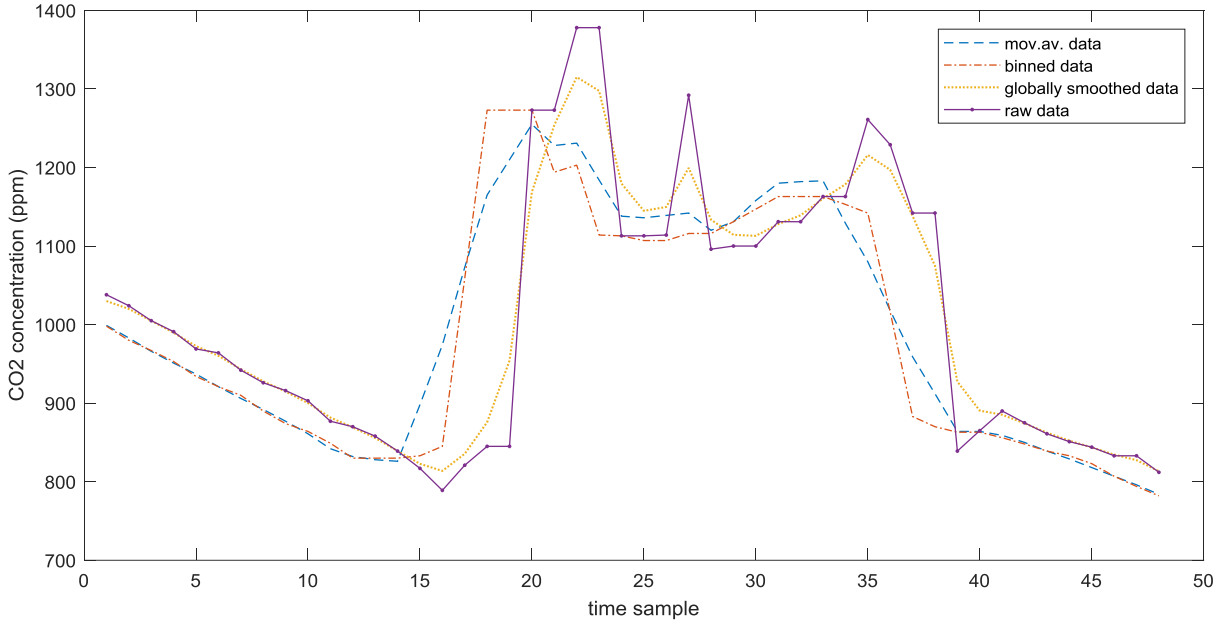


Fig. 5-5: One day CO_2 concentrations of the reference office room

5.4 Evaluate criteria

To assess the performance of the proposed models of the occupancy estimators, we consider the three performance indices:

- *normalized root mean square error* (NRMSE) [5.37], indicating the relative mean magnitude of the calculation error, as defined in (5.24);
- *x-tolerance accuracy*, reporting the percentage that the occupancy estimation model can provide the results whose calculation errors are less than *x*, as defined in (5.25);

- *detection rate* (DR) is the calculation accuracy of two states (unoccupied/occupied), representing the rate of accurate detecting whether the room is occupied or not, as defined in (5.27);

These performance indices are computed as follows:

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{\|\hat{\mathbf{o}}-\mathbf{o}\|_2^2}{\mathrm{T}}}}{\mathbf{o}_{\max}-\mathbf{o}_{\min}} \tag{5.24}$$

Where $\hat{\mathbf{o}} \in \mathbb{R}^n$ is the estimated number of occupants in n sampling time, and $\mathbf{o}$ present the real occupancy number, $\|\cdot\|_2^2$is the l_2-norm, T is the number of the samples in a day.

x-tolerance accuracy is defined as:

$$\tau(\hat{\mathbf{o}}, \mathrm{x}) = \frac{\mathrm{n}-\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathbb{x}(|\hat{\mathrm{o}}_\mathrm{k}-\mathrm{o}_\mathrm{k}|,\mathrm{x})}{\mathrm{n}} \tag{5.25}$$

Where $\mathbb{x}(\cdot)$is the indicator function, $\mathbb{x}(\mathrm{x}) = \begin{cases} 1, & \text{if } |\hat{\mathrm{o}}_\mathrm{k} - \mathrm{o}_\mathrm{k}| \le \mathrm{x} \\ 0, & \text{otherwise} \end{cases}$

It can be seen that there is a special case of x-tolerance accuracy when $\mathrm{x} = 0$, such as $\tau(\hat{\mathbf{o}}, 0) = \mathrm{Acc}$. In most studies, Acc has been employed as an important performance index, which represents the rate of accurate detecting in a day, defined as:

$$\mathrm{Acc} = \frac{\mathrm{n}-\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathbb{I}(\hat{\mathrm{o}}_\mathrm{k}-\mathrm{o}_\mathrm{k})}{\mathrm{n}} \tag{5.26}$$

Where $\mathbb{I}(\cdot)$is the indicator function, $\mathbb{I}(\mathrm{x}) = \begin{cases} 1, & \text{if } \mathrm{x} > 0 \\ 0, & \text{otherwise} \end{cases}$

DR is defined as:

$$DR = \frac{n - \sum_{k=1}^{n} |\mathbb{I}(\hat{o}_k) - \mathbb{I}(o_k)|}{n} \tag{5.27}$$

Where $\mathbb{I}(\cdot)$is the indicator function, $\mathbb{I}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$

These three indices are used to assess the performance of occupancy estimators from different aspects.

5.5 Comparison between ML approach and Bayesian estimation approach

5.5.1 The ground truth of occupancy schedule

The daily schedule of real occupancy on working days is plotted and analyzed. **Fig.5-6** shows the boxplot and average occupancy schedule from 01/15/2018 to 01/19/2018. In the building design phase, the internal load is estimated based on the assumption that the office room is full-occupied. In order to reveal the discrepancy between building design condition and building actual operation, prototype occupancy schedule is presented and compared with the true occupancy. The dot line plot represents the prototype occupancy schedule in the energy simulation software (the room is filled to capacity during working period). It could be seen that, occupancy variations with nearly 25 people are observed from 2:00p.m. to 4:00 p.m., which is much larger than those in the morning. A primary reason could be that the most occupants are architecture engineers, onsite investigation for half-day leaving is more usual in the afternoon than other times. Another observation is that the lunch break pattern is fairly consistent, starting from 11:30 a.m. until 13:00 p.m. The lowest

occupancy rate due to lunch break occurs at 12:00 p.m. with average 7 people inside the office.

As for weekly analysis, the daily occupancy differences between real schedule and prototype schedule are shown in **Table 5-4**. As shown in **Fig.5-7**, there is a clear observation that Monday and Wednesday have higher occupancy rate while Tuesday and Thursday have lower occupancy rate, and Friday has medium occupancy rate. The magnitude of the occupancy rates has large variance for different days within a week, which offers opportunities for the facility management to make better control strategy towards building energy efficiency.

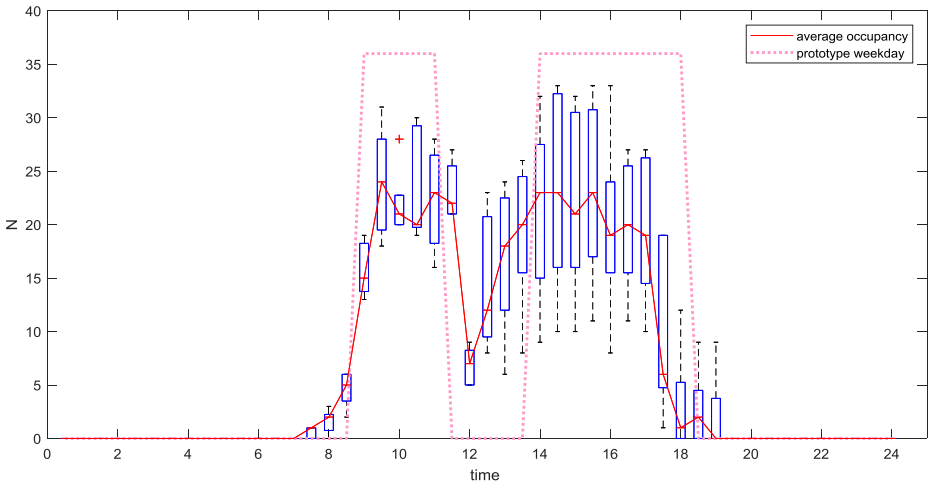


Fig. 5-6: Boxplot of the daily schedule of true occupancy in the reference office

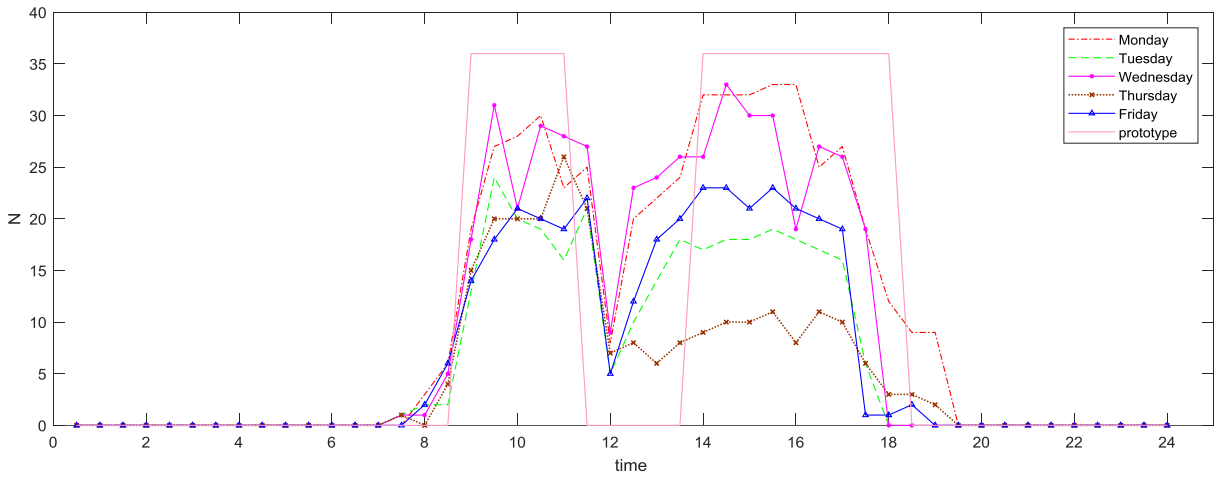


Fig. 5-7: True occupancy schedule in the reference office

Table 5-4: Daily occupancy difference compared to the prototype weekday schedule

Monday	Tuesday	Wednesday	Thursday	Friday
-1.19%	-41.67%	-10.11%	-52.57%	-34.32%

5.5.2 Result from parameter estimation models

The performance indices (NRMSE, *x-tolerance* accurancy, and DR) of the five days occupancy estimation are shown in **Table 5-5**, **Table 5-6** and **Table 5-7**, respectively. **Table 5-5** shows NRMSE value of two parameter estimation models, where the average (over the days) daily results are reported. The Bayesian estimation outperforms frequentist ML in all cases except Thursday, giving a low calculation error ranging from 0.0885 to 0.1391 in NRMSE. While the NRMSE of frequentist ML model ranges from 0.0937 to 0.1614. Smoothing the CO_2 concentration data by

moving average approach and globally smoothing approach can reduce the calculation errors of both parameter estimation models. However, the binned CO_2 data fails to give an improvement in terms of NRMSE with Bayesian estimation.

Table 5-5: NRMSE of frequentist ML and Bayesian estimation

Methods	CO_2 concentration data	NRMSE					
		Mo	Tu	We	Th	Fr	Average
frequentist ML	Raw data	0.1513	0.1105	0.1514	0.1141	0.1184	0.1291
	Moving average data	0.1399	0.0989	0.1456	0.0976	0.1162	0.1196
	Binned data	0.1614	0.1108	0.1483	0.0963	0.1078	0.1249
	Globally smoothed data	0.1562	0.1069	0.1404	0.0937	0.1043	0.1203
	Average	0.1522	0.1067	0.1464	0.1004	0.1116	**0.1235**
Bayesian estimation	Raw data	0.1391	0.0917	0.1023	0.1298	0.0928	0.1114
	Moving average data	0.1162	0.1030	0.1031	0.1312	0.0869	0.1080
	Binned data	0.1409	0.1101	0.1145	0.1252	0.0814	0.1144
	Globally smoothed data	0.1190	0.0975	0.1065	0.1211	0.0932	0.1125
	Average	0.1288	0.1005	0.1129	0.1268	0.0885	**0.1115**

The *x-tolerance accuracy* of the result is shown in **Table 5-6**. For both two parameter estimation models, the average *3-tolerance accuracy* is no more than 70%. Frequentist ML model with globally smoothed CO_2 data gives best accuracy result 80.00% than the other three data pre-processing methods. However, for Bayesian

estimation model, the highest accuracy 84.16% is conducted by using raw CO_2 data. For both parameter estimation models, the binned data fails to give an improvement in terms of *x-tolerance accuracy*. As show in **Fig.5-8**, the Bayesian estimation provides higher accuracy than frequentist ML in most cases; and the accuracy of the two models are almost equal with the increase of *x* value. With respect to Acc (*0-tolerance accuracy*), the Bayesian estimation provides accuracy no lower than 50% except using binned data. It is also found that the Acc of frequentist ML is not lower than 50% except using raw data. Hence, the performances of these two parameter estimation models are similar in terms of Acc.

Some studies demonstrated that Acc is a suitable criterion when the number of indoor occupants is less than 4 [5.28, 5.41 and 5.42]. But if the number of indoor occupants is large, as 36 in this case, Acc is not a suitable performance index. For example, $\hat{o}_k = 15$ while $o_k = 14$ would be classified as false result using Acc. In fact, 14 or 15 occupants have no large differences for optimization the control strategy of appliances [5.39]. While the estimation of the fine exact number of occupancy is excellent, the intended application in energy efficiency and control optimization which does not require exact number of occupants. For office space with more than 4 occupants, several misestimated occupants have insignificant influence on the operation control of the AC system. Additionally, too large tolerance accuracy is also meaningless because it would cause estimation error on building energy consumption. Hence, *4-tolerance accuracy* has been adopted as one critical index in this study, as shown in **Fig.5-9**. The purpose of utilizing 4 is to balance the weights between estimation accurate and control optimization. We can easily find out that the Bayesian estimation

provides consistent and highly-accurate results, while the frequentist ML suffers large variance in *4-tolerance accuracy*. The *4-tolerance accuracy* is relatively low in two parameter estimation models on Monday and Thursday. The miscalculation is probably aroused by the random of windows open or stochastic occupant behaviours. What's more important, the Bayesian estimation gives better accuracy result when using raw data. This is probably due to the preprocessed CO_2 data is too smooth compared to the raw data. We remark that the data pre-smoothing is a tradeoff between real-time measurement and stability of estimation. Namely, to eliminate the spikes of raw CO_2 concentration data, it is effective to increase the number of samples when calculating moving average/binned values, or use larger regular factor during globally smoothed calculation. The preprocessed data can be adopted to generate the occupancy model with high stability. However, some important variances of raw data are probability lost during pre-smoothing process, then the calculation accuracy of the occupancy estimation model is degraded, and vice versa.

As for the DR index shown in **Table 5-7**, both two parameter estimation models can give a good estimation of the room state (e.g. occupied/unoccupied), because the real-time electricity consumption of appliances is adopted to modify the number of occupancy. Specifically, the number of occupancy is set to be zero when the electricity consumption is lower than 0.5KWh.

Table 5-6: x-tolerance accuracy of frequentist ML and Bayesian estimation

Methods	CO_2 concentration data	*x* value	
			Average

		3	4	5	6	7	
frequentist ML	Raw data	64.58	72.91	77.08	81.25	85.41	76.24
	Moving average data	58.33	70.83	77.03	83.33	85.41	78.33
	Binned data	66.67	70.83	81.25	83.33	89.58	74.98
	Globally smoothed data	66.67	77.08	81.25	87.50	87.50	80.00
	Average	64.06	72.91	79.15	83.85	86.97	**77.39**
Bayesian estimation	Raw data	66.67	83.33	87.50	91.67	91.67	84.16
	Moving average data	72.91	79.17	83.33	87.50	87.50	82.08
	Binned data	66.67	79.16	81.25	85.41	87.50	79.99
	Globally smoothed data	70.83	79.16	87.50	89.58	91.67	83.74
	Average	69.27	80.20	84.89	88.54	89.58	**82.49**

Table 5-7: DR of frequentist ML and Bayesian estimation

Methods	DR(%)			
	Raw data	Moving average data	Binned data	Globally smoothed data
frequentist ML	97.08	97.50	97.08	97.08
Bayesian estimation	97.08	97.08	97.08	97.08

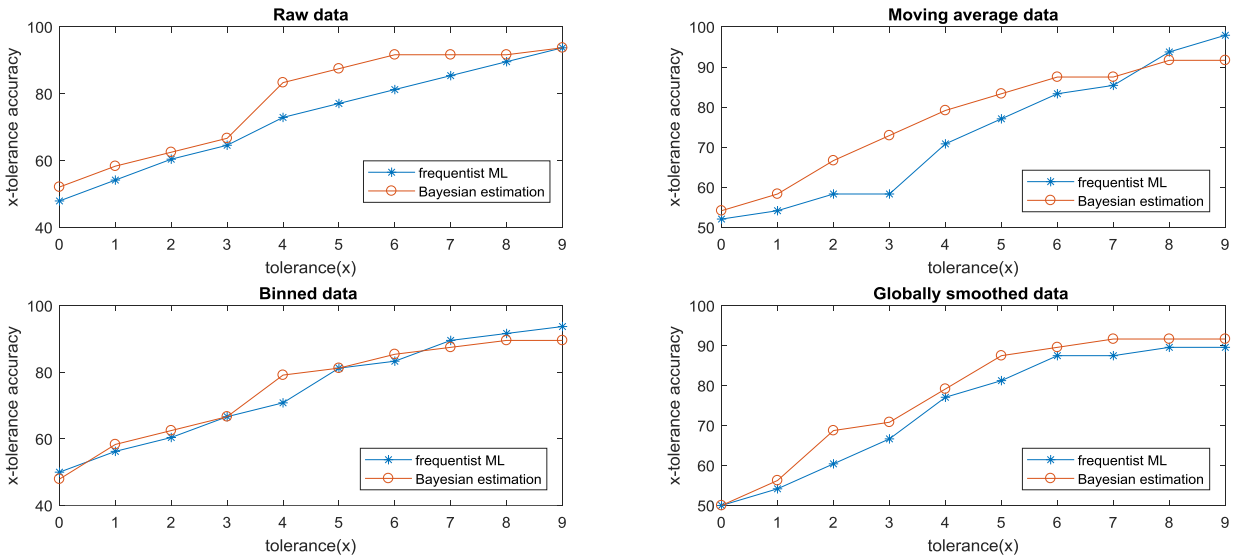


Fig. 5-8: x-tolerance accuracy of the models developed based on different data of CO_2 concentration

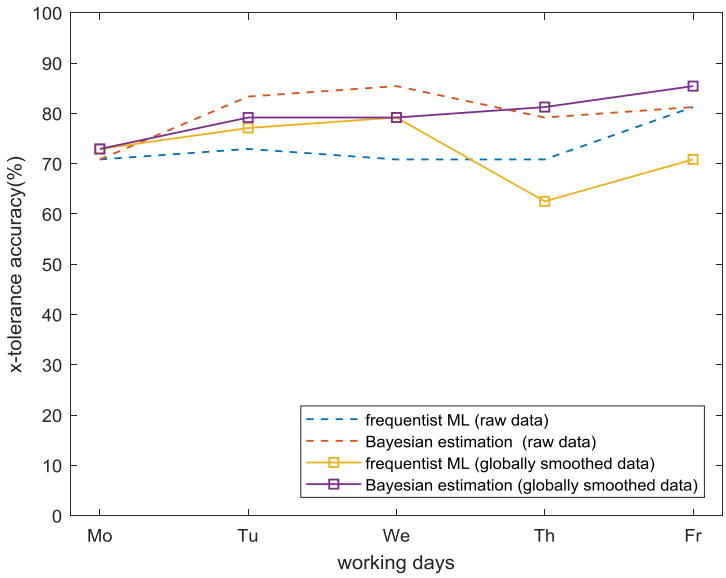


Fig. 5-9: x-tolerance accuracy of the models developed based on raw data of CO_2

concentration and globally smoothed data of CO_2 concentration. Note that x=4 in this case.

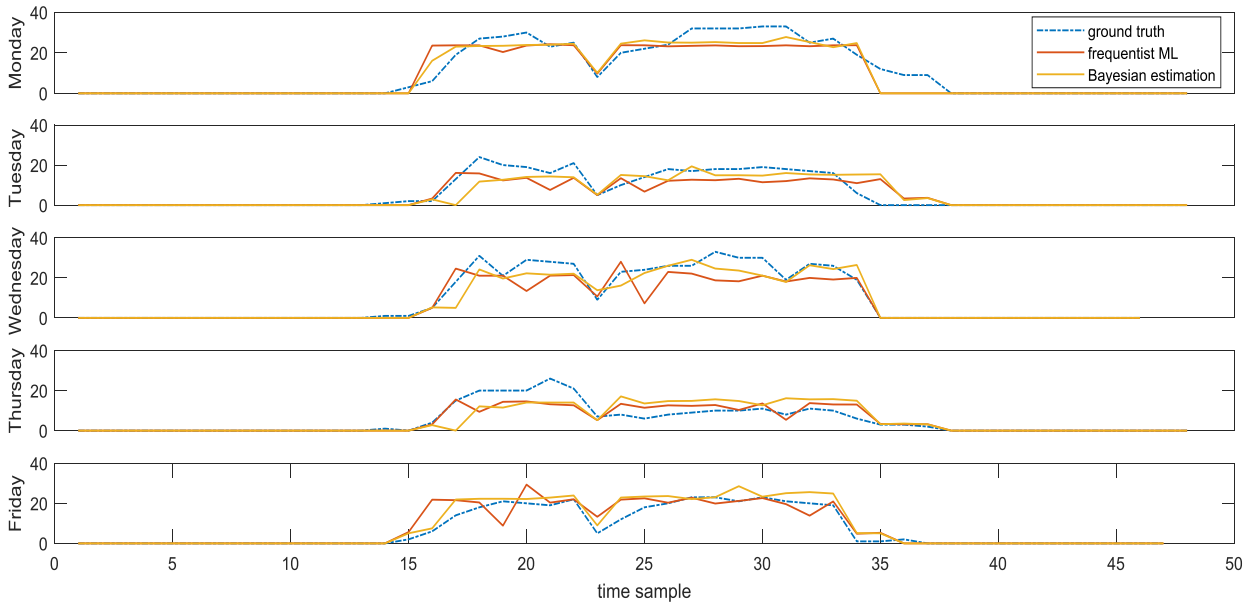


Fig. 5-10: Occupancy estimation results of the five working days by using moving average data of CO_2 concentration.

The visualization comparisons of the occupancy estimated by two parameter estimation models are presented in **Fig.5-10**. There are several spikes in the true occupancy profile which are not identified by both parameter estimation models. These spikes represent a sudden change in the number of occupants in short time, such as stopping by or temporary leavings. In addition, the indoor CO_2 concentration is a cumulative parameter with time lags; it is difficult to track every spike in occupancy profile. We can conclude that the estimated profiles can track the changes in occupancy fairly well on Tuesday, Wednesday as well as Friday. As for Monday

and Thursday, the estimated profiles present the "smoothed" version of true occupancy. What's more, the two parameter estimation models fail to track the time of leaving on Monday evening, because occupants may work overtime occasionally.

5.5.3 Comparison with results reported in the literature

Because the experimental data reported in the literature (section 5.1) is not available, and the metering condition and equipment are not the same, it is not possible to fully compare all the models and their performance indices. Several proposed models, such as HMM, DT, LDA and ELM, report the NRMSE between 0.1912-0.2509 which are higher than those of the frequentist ML model (0.1235) and the Bayesian estimation (0.1115). For DR index, the existed models put the accuracy between 88.54%-93.63%, which are lower than the parameter estimation models. It can be easily seen that the both frequentist ML and Bayesian estimation give more accurate results compared to existed models.

5.6 Chapter summary

As discussed earlier, the development of the most occupancy estimators depends on the people-counting data. The fundamental task of these statistical models is to figure out the probabilistic/numerical relationship between people-counting data and corresponding state variables. Though providing acceptable results with good accuracy, all the previous relative studies did not include deep investigation of physical relationship between occupant number and environmental parameters. Therefore, large amount of measurement data must be obtained at the beginning, leading to tediously long time for data collection. To bridge the gap between

occupancy estimation from environmental data and independent calculation without people-counting data, two novel parameter estimation models are introduced to calculate the unknown variables blindly based on the dynamic of indoor CO_2 concentration. Therefore, the robustness of the proposed parameter estimation models offers a new method for occupancy estimation, which further benefits the prediction of building energy consumption and development of control strategies for building automation.

Indoor occupancy estimation from the non-iterative environmental parameters is a challenging project, especially for the large office space where a few tens occupants inside. Different from most of works which depend on training sets, our parameter estimation models are able to compute the number of occupants blindly.

Specifically, the model starts from identification of indoor CO_2 dynamics, which is derived from mass-conservation law and ventilation level. In order to offset the negative influence of CO_2 concentration with spikes on occupancy estimation, two-hour moving average, two-hour bin and globally smooth have been adopted as data pre-process to smooth the CO_2 data. Then, frequentist ML and Bayesian estimation are applied to estimate the number of occupants. The significant conclusions show that the Bayesian estimation outperforms frequentist ML giving a low calculation error. Compared to other proposed models, our parameter estimation models can give lower calculation error and higher calculation accuracy.

In conclusion, the calculation results show that both frequentist ML and Bayesian estimation can give reliable and accurate estimation in a real-time and non-iterative

way. A nature extension of the current work is to improve the calculation accuracy of the parameter estimation models, as well as the occupancy interconnection of multiple rooms.

CHAPTER 6 ENERGY PREDICTION MODEL

6.1 Overall review of prediction model of electricity consumption in office building

As reviewed in section 2.1, the downside of fundamental ANN and SVM require the long training time for excessively large amounts of data. Another limitation of ANN and SVM is due to the difficulty in the determination of model parameters and kernel function. There is no uniform standard to select suitable model parameters and kernel functions. Researchers must make decisions based on the characteristics of the data and/or their experience. The ELM, as an emergent technology that overcomes the challenges faced by fundamental ANN and SVM, has been proposed recently. Unlike the fundamental ANN, the hidden layer of ELM does not need to be tuned. This approach is capable of solving the problem without a back-propagation training process. Naji et al. [6.1] and Li et al. [6.2] have developed a building energy estimating model based on ELM, the results revealed the robustness of the ELM method.

Many researchers have analyzed the influence of input parameters on prediction accuracy and observed that the indoor occupancy plays a critical role in energy prediction model. A fixed occupancy schedule or simple day type (weekdays/weekends) was usually adopted as an input parameter by some models [6.3, 6.4]. These results indicated that the fixed profile of the occupancy was able to increase the prediction accuracy compared to that of the cases without consideration of the occupancy schedule. Other approaches to simulate dynamic internal load

variance caused by occupants were based on the time factor [6.5, 6.6] and historical load of electricity consumption [6.7, 6.8]. For example, a multi-layered feedforward network with Bayesian regularisation was adopted by Y.T. Chae et al. [6.8] to predict the sub-hourly energy consumption. In their work, both the day type and time of day were selected as input elements. As for the large-scale campus building, the fit of a load-forecasting model was dramatically improved when time of day and day of week were introduced as inputs [6.9]. In another study [6.10, 6.11], three kinds of factors were introduced to represent the internal load features: the operating schedules of air units, hour of day, and occupancy space power demand. Although use of the occupancy space power demand would enhance the model accuracy, the prediction was not sufficiently accurate for winter. Concerning implementation of historical load as an influential factor of internal load, An et al. [6.12] used previous electricity consumption as the input of a multi-output prediction model to forecast the energy demand for the next half hour. In Ref. [6.13], transitional characteristics and power-level characteristics of the heating system were adopted as the input variables of the ANN for load prediction. These variables were validated using a precise description of transitional delay for fluctuation of occupancy in buildings. In recent research [6.14], the authors concluded that the ANN was able to be used for forecasting the next day's energy use based on the five previous days' data with acceptable accuracy. In all the above studies, the fixed occupancy schedule, time factor, and historical internal load were considered as input parameters of the energy-prediction models.

However, variance of the internal load aroused by randomness due to occupancy variation cannot be represented comprehensively and accurately with a fixed

occupancy schedule, time factor, or historical load. Most previous works show that there is a considerable performance gap between the predicted and measured energy use. One cause of poor performance is the interaction between occupancy and building systems. By introducing occupant information as a component of building performance, energy consumption can be very different according to the various occupant actions [6.15]. Therefore, introducing a real occupant schedule into the building energy-prediction model has been brought to the forefront. Some energy-prediction models have been proposed based on the estimated occupancy profile [6.16]. However, in reality, few energy-prediction models use true occupant presence or interactions (adjusting thermostats, lighting control, etc.) as the inputs to the model design because the key parameters regarding occupancy is difficult to collect. Some basic models were proposed based on a few test buildings, such as an airport terminal [6.17] or single office room [6.18].

After reviewing peer research, a few key points are identified:

- Despite the extensive application of ANNs in prediction, ANN training is time-consuming and greatly affected by model parameters. It is necessary to improve the generality and prediction performance of the fundamental ANN models.
- As one of the leading influential factors for energy consumption in buildings, the occupancy factor is usually introduced as a fixed occupancy schedule, time factor, or historical load as an input parameter of the energy-prediction model. Few studies have focused on using a dynamic occupancy profile to predict the

building load with ANN. The neglect of occupancy presence and behaviour is one cause of the performance gap between the predicted and measured energy use.

Fundamental ANN models, ELM models, and ensemble models are constructed as the prediction performance of building energy consumption. Finally, we compare the performance of these three models with the different supplementary inputs, i.e., true occupancy, estimated occupancy from BSI models, and without inputs of occupancy. Overall, this paper focuses on bridging the gap between energy-prediction models and the dynamic occupancy profile estimated from indoor CO_2 concentration.

6.2 Methodologies for energy prediction

6.2.1 Architecture of FFNN model

In this study, we use an FFNN with one hidden layers of neurons and a single linear output to predict the electricity consumption of an AC system. The number of hidden neurons is determined when the accuracy is satisfied as changing the structure of the FFNN. The structure of FFNN (BP network) is shown in section 2.1.1. The establishment of the energy prediction models was based on the MATLAB program.

6.2.2 Architecture of ELM model

As mentioned before, it is known that fundamental ANN faces some challenging issues, such as slow learning speed and necessity of human intervention [6.19]. Compared with those traditional computational intelligence techniques, ELM

provides better generalisation performance at a much faster learning speed and with less intervention. Unlike fundamental ANN, ELM not only tends to reach the smallest training error, but also the smallest norm of the output weights. According to the neural network theory [6.20], the generalisation performance of the model would be improved when the smaller training error and smaller norm of weights are both considered.

ELM is a tool of learning algorithm for the single-layer FFNN architecture [6.21]. The essence of ELM is that the hidden layer need not be tuned. Therefore, this algorithm requires less calculation time compared to the traditional FFNN because it determines the network weights and minimises the sum of the error simultaneously without iterative training. In the ELM, the least-squares method is adopted to optimise the output weighting matrix β.

A set of inputs x_i $(i = 1, 2, \dots\dots, m)$ corresponds to a known representation $\sigma_k(x_i)$. The desired output is represented as z_q $(q = 1, 2, \dots\dots, n)$. The estimated output $\hat{z}_q$ is expected to be calculated to minimise the estimation error to zero. The process can be expressed as

$$\sum_{q=1}^{n} \left\| \hat{z}_q - z_q \right\| = 0. \tag{6.1}$$

The problem can be described as

$$\sigma_k(x_i)\, \beta = z_q, \tag{6.2}$$

where $\sigma_k(x_i)$ is the input matrix of the ELM, which is represented as

$$\sigma_k(\mathrm{x_i}) = [\sigma_k(\mathrm{x_1}), \sigma_k(\mathrm{x_2}) \dots \dots \sigma_k(\mathrm{x_m})] \tag{6.3}$$

$$= \begin{pmatrix} f(b_1^k + w_1^k f(\dots w_1^2 f(b_1^1 + w_1^1 \mathrm{x_1}))) \\ f(b_1^k + w_1^k f(\dots w_1^2 f(b_1^1 + w_1^1 \mathrm{x_2}))) \\ \vdots \\ f(b_1^k + w_1^k f(\dots w_1^2 f(b_1^1 + w_1^1 \mathrm{x_m}))) \end{pmatrix}^T_{m \times n},$$

$$\beta = [\beta_1, \beta_2 \dots \dots \beta_m]^T, \tag{6.4}$$

$$\mathrm{z_q} = [\mathrm{z_1}, \mathrm{z_2} \dots \dots \mathrm{z}_n]^T. \tag{6.5}$$

According to matrix theory [6.22], the optimal matrix β in Eq. (6.2) is derived as

$$\beta = \sigma_k(\mathrm{x_i})^{+}\mathrm{z_q}, \tag{6.6}$$

where $\sigma_k(\mathrm{x_i})^{+}$is the Moore–Penrose generalised inverse of $\sigma_k(\mathrm{x_i})$.

Hence, the extreme learning of the weighting matrix takes place of the iterative training process of traditional back-propagation neural networks, such like FFNN. ELM tends to obtain the least training error at once. Ease of use, compatibility with various types of activation functions, fast calculating speed, and superior performance are the advantages of this algorithm.

6.2.3 Architecture of ensemble model

Engineering problems such as energy-use prediction is perhaps difficult for a single neural network. An ensemble of results from two neural networks is conducted for the possible improvement of prediction accuracy. In this research, the neural-network ensemble is developed based on the average value of model outputs from FFNN and ELM. In **Fig. 6.1**, the neural-network ensemble in this study is shown.

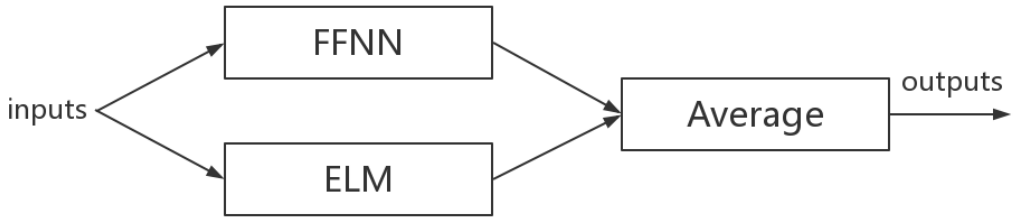


Fig. 6-1: Neural-network ensemble structure

6.3 Data sets

Reference room and dataset used in energy prediction model are same with the occupancy estimation model described in chapter 5. Readers could refer to section 5.1 for more detailed information.

6.4 Evaluation criteria

To evaluate the obtained results from prediction model and compare the performance of different FFNN models, the coefficient of determination (R^2), root-mean-square error (RMSE), and mean absolute percentage error (MAPE) are usually adopted as evaluating criteria. These coefficients are defined as

$$R^2 = \frac{\sum_{q=i}^{n}(\hat{y}_q - y_q)^2}{\sum_{q=i}^{n}(y_q)^2}, \tag{6.7}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{q=i}^{n}(\hat{y}_q - y_q)^2}{n}} \times 100, \tag{6.8}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{q=1}^{n}\left|\frac{\hat{y}_q - y_q}{y_q}\right| \times 100\%. \tag{6.9}$$

In addition to MAPE, defined by Eq. (6.9), two other indices are adopted for evaluate the model performance: $\mathrm{MAPE}_{\mathrm{peak}}$ and $\mathrm{MAPE}_{\mathrm{simple-peak}}$.

$$\mathrm{MAPE}_{\mathrm{peak}} = \left|\frac{L_{mpl}-\hat{L}_{mpl}}{L_{mpl}}\right| \times 100, \tag{6.10}$$

where$\mathrm{L}_{\mathrm{mpl}}$ and $\hat{\mathrm{L}}_{\mathrm{mpl}}$ are the actual and predicted electricity consumption of the AC system. $\mathrm{MAPE}_{\mathrm{peak}}$ is introduced to locate the actual daily consumption peak and its occurrence time, and to compare the magnitude to the predicted value at the same time. The relative error is calculated as the performance index.

Then, the $\mathrm{MAPE}_{\mathrm{simple-peak}}$ is adopted to compare the daily peak value of the actual and predicted electricity consumption, without considering the occurrence time, defined as

$$\mathrm{MAPE}_{\mathrm{peak}} = \left|\frac{L_{opl}-\hat{L}_{ppl}}{L_{opl}}\right| \times 100, \tag{6.11}$$

where $\mathrm{L}_{\mathrm{opl}}$ and $\hat{\mathrm{L}}_{\mathrm{ppl}}$ are the peak values of the actual and predicted electricity consumption of the AC system.

6.5 Parameter selection analysis

In this section, we discuss the effect of different parameter choices on the prediction accuracy of our model, including the number of input parameters and structure parameters of the FFNN model.

6.5.1 Principal component analysis

In this study, a full-scale site measurement on the environment parameters, energy-consumption data, and estimated number of occupants, including 12 variables, for the reference office room is conducted. In reality, it is not always possible to obtain all the variables that are collected in this study. Hence, the importance is that prediction of the energy consumption with acceptable accuracy is still achievable even if using only a few of the most influencing parameters.

In order to identify the most important factors which have large influences on the power consumption of AC system, principal component analysis (PCA) has been used to assess these factors. PCA is a popular multivariate statistical analysis [6.23] method and has been successfully adopted in various applications [6.24, 6.25]. The analysis result is shown in **Table 6-1**. The contribution rate and cumulative contribution rate of PCA are displayed in **Fig. 6-2**. It is indicated that the four principal components explain 81.479% of the total variance. Hence, these four components are used to represent the raw variables.

Table 6-1: Principal component analysis results

Component	Eigenvalues	Contribution rate (%)	Cumulative contribution rate (%)
1	5.474	45.615	45.615
2	2.369	19.742	65.357
3	1.059	8.822	74.179
4	0.876	7.300	81.479
5	0.656	5.470	87.060

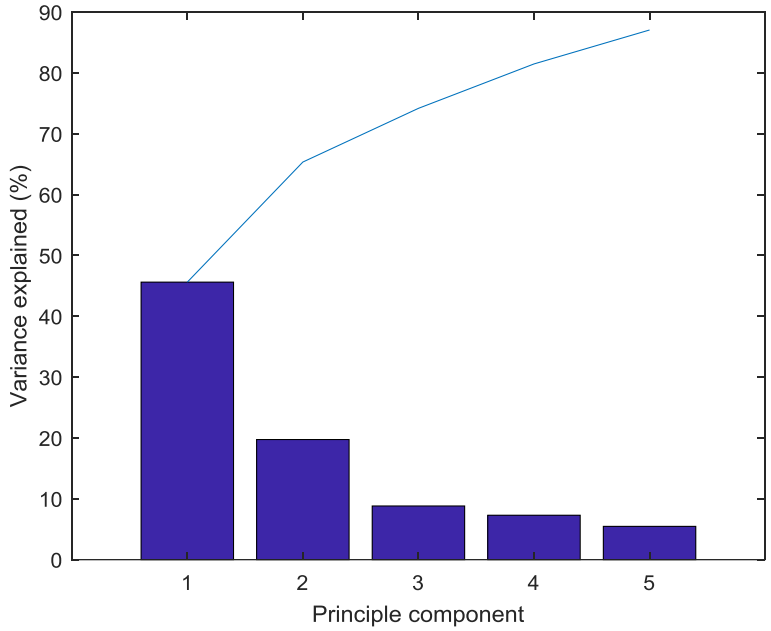


Fig. 6-2: Contribution rate and cumulative contribution rate of PCA

Based on the PCA result, these 12 independent variables are listed with the most significant at the top in **Table 6-2**. The values of the component matrix are shown in **Table 6-3**. Variables having coefficients with absolute values greater than 0.7 from these top 5 components were considered statistically significant. It should be mentioned that the importance ranking conducted by PCA is merely statistical relationships within the dimensionless datasets, rather than the physical influence of theses parameters on energy use. In this study, the former six independent variables, including electricity consumption of appliances, number of occupants, electricity consumption of lighting, solar radiation, electricity consumption of the fresh-air system, and outdoor temperature are selected as the inputs to the predication model. The reason for omitting the other six variables is that certain variables have less effect

on building energy consumption and too many predictors that are highly correlated may cause a decrease in the accuracy.

Table 6-2: Statistical importance ranked by PCA

variables	Unit/index
Electricity consumption of appliances (P)	kWh
Number of occupants (O)	$0-36$
Electricity consumption of lighting (L)	kWh
Solar radiation (S)	W/m^2
Electricity consumption of fresh air system (F)	W/m^2
Outdoor dry-bulb temp. (T)	℃
Indoor relative humidity (r)	%
Indoor dry-bulb temp. (b)	℃
Outdoor relative humidity (R)	%
Indoor carbon dioxide (C)	ppm
Electricity consumption of ceiling fan (A)	kWh
Wind speed (W)	m/s

Table 6-3: Component matrix

	Components				
	1	2	3	4	5
Electricity consumption of appliances (P)	0.928	0.170	-0.031	0.002	0.095
Number of occupants (O)	0.925	0.212	0.032	-0.043	0.105
Electricity consumption of lighting (L)	0.859	-0.142	0.144	0.052	-0.164

Solar radiation (S)	0.822	-0.041	0.248	0.019	0.179
Electricity consumption of fresh air system (F)	0.817	-0.416	-0.086	0.039	0.090
Outdoor dry-bulb temp. (T)	0.754	0.382	-0.221	-0.277	0.109
Indoor relative humidity (r)	-0.156	0.643	0.179	0.342	0.145
Indoor dry-bulb temp. (b)	-0.556	0.552	0.147	0.219	0.150
Outdoor relative humidity (R)	-0.138	-0.543	-0.032	0.305	0.549
Indoor carbon dioxide (C)	0.458	0.626	-0.460	-0.078	0.159
Electricity consumption of ceiling fan (A)	0.653	0.082	0.669	0.061	-0.089
Wind speed (W)	0.583	-0.069	-0.363	0.632	-0.310

Furthermore, the electricity consumption of the AC system with three time-step delays are also parameterised as inputs to predict the energy consumption at the current time. Additionally, the hour of the day is coded by a sine value, as below, and fed into the energy-prediction model.

$$\mathrm{sh} = \sin(\frac{\pi t}{48}) \tag{6.12}$$

6.5.2 Effect of structure parameters of FFNN model

To save computational time, an FFNN model with fewer neurons in the hidden layer is more favourable if it can meet the requirement of accuracy. The values of the structure parameters are identified by changing the neurons in the hidden layer.

In order to find the effect of the FFNN structure on its accuracy in prediction, we investigate the performance of different neuron sizes. We start from three-layer

FFNNs, for which the number of neurons in the hidden layer increases from 10 to 160. Each network configuration has been tested by repeating 10 times to evaluate the robustness of the performance, which should eliminate the effect of the randomness of the initial setting of weights and bias on the prediction accuracy. The performance of these FFNNs is identified by using the value of MAPE after 1500 iterations. **Figure 6-3** shows the value of MAPE in training (80% datasets) and validating (20% datasets) for all these FFNNs. It is shown that MAPE is the lowest (6.43% for training, 7.29% for validation) when the hidden-neuron size is 120, which is regarded as the best structure of the model.

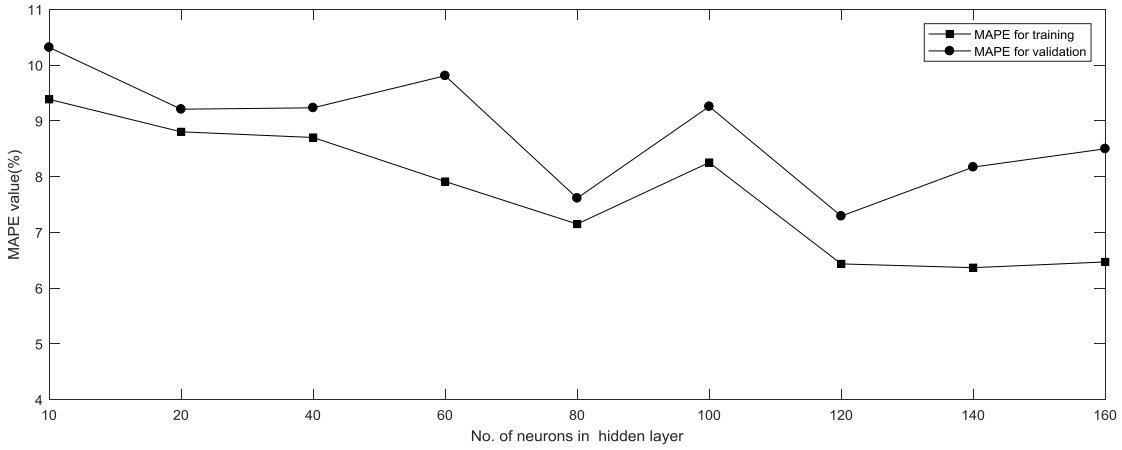


Fig. 6-3: MAPE performance with different number of neurons in the hidden layer

6.6 Prediction results

In this section, the aforementioned FFNN and ELM are applied to predict the electricity consumption of the AC system. To highlight some aspects of the

benchmark test and introduce occupancy as the input of the energy-prediction model, we conduct following three studies:

1) Occupancy-estimation results.

2) Energy-prediction result with the input of true occupant counts.

3) Energy-prediction result with the input of estimated occupant counts.

6.6.1 Energy-prediction result with true occupant counts

In order to investigate the prediction performance of the models with the input of true occupant counts, other models without the input of occupant number are also constructed for comparison. Additionally, the effects of PCA pre-treatment on prediction accuracy are analyzed, and a total of twelve models are proposed to predict the energy consumption of the AC system in the next time-step $z(k+1)$ shown in **Table 6-4**. As shown in **Table 6-4**, the model named FFNN10 refers to the energy-prediction model with 10 inputs that adopts FFNN as the neural-network structure.

Table 6-4: Input parameters of prediction models (with true occupant counts)

Model	Input parameters	Output parameter
FFNN10, ELM10, Ensemble10	$P(k), O(k), L(k), S(k), F(k), T(k), z(k), z(k-1), z(k-2), sh$	$z(k+1)$
FFNN9, ELM9, Ensemble9	$P(k), L(k), S(k), F(k), T(k), z(k), z(k-1), z(k-2), sh)$	

FFNN15, ELM15, Ensemble15	$P(k), L(k), S(k), F(k), T(k), W(k), r(k), b(k), R(k), C(k), A(k), z(k), z(k-1), z(k-2), sh$
FFNN16, ELM16, Ensemble16	$P(k), O(k), L(k), S(k), F(k), T(k), W(k), r(k), b(k), R(k), C(k), A(k), z(k), z(k-1), z(k-2), sh$

In **Table 6-4**, $z(k), z(k-1)$, and $z(k-2)$ are the electricity consumption of the AC system at times k, $k-1$, and k-2. $O(k)$ is the true occupant count and sh is the sine value representing the hour of the day.

Electricity consumption of the AC system and true occupancy profile during the workday is depicted in **Fig. 6-4**. From **Fig. 6-4**, the occupancy profile provides major information about electricity-consumption variances. However, the electricity consumption is not in accordance with the occupancy profile at 18:00. Thus, it further shows that occupancy cannot be regarded as the only indicator to predict energy demand; other parameters such as time indicator and electricity consumption of the appliance might be considered as indicators as well.

Table 6-6 and **Table 6-7** show the comparison of the performance of the proposed models, including FFNN, ELM, and the ensemble models in training datasets and validation datasets, respectively. Average results from cross validation are reported to avoid the biased condition. It is clear that the ensemble models generate the best prediction results and validation results with 10 inputs. The reason may be that the data of true occupant counts is necessarily close to the electricity consumption of AC system and hence improves the prediction performance. It could be seen from the

result that models developed using the PCA selection slightly outperform the models using all the inputs. In addition, it is validated that the ensemble model in general can improve the prediction accuracy more than the individual neural-network model.

If we look more deeply into the performance indices of prediction models, the boxplots of **Table 6-5** and **Table 6-6** are plotted in **Fig. 6-5** and **Fig. 6-6** according to the different numbers of input parameters and prediction models, respectively. The predictions are evaluated based on the minimum, maximum, median, 25^{th}-percentile, and 75^{th}-percentile values of the performance indices. By considering the variance of the R^2, MAPE, and RMSE, the prediction models with 10 inputs yields the least variance. Furthermore, it is obvious from **Fig. 6-5** that better prediction is obtained by the prediction models with 16 inputs rather than those with 9 inputs. This means that models with 10 inputs > models with 16 inputs > models with 9 inputs > models with 15 inputs, where > indicates 'performs better than'. The results illustrate that the model performance is more sensitive to the number of occupants rather than the PCA selection. As for the performance of different prediction models shown in **Fig. 6-6**, it can be easily seen that FFNN models provide the lowest R^2 value and highest MAPE and RMSE. However, it is difficult to draw the conclusion for the best-performing method between the ELM model and ensemble model from **Fig. 6-6**, as ELM models provide more stable performance with smaller variances of R^2 and MAPE.

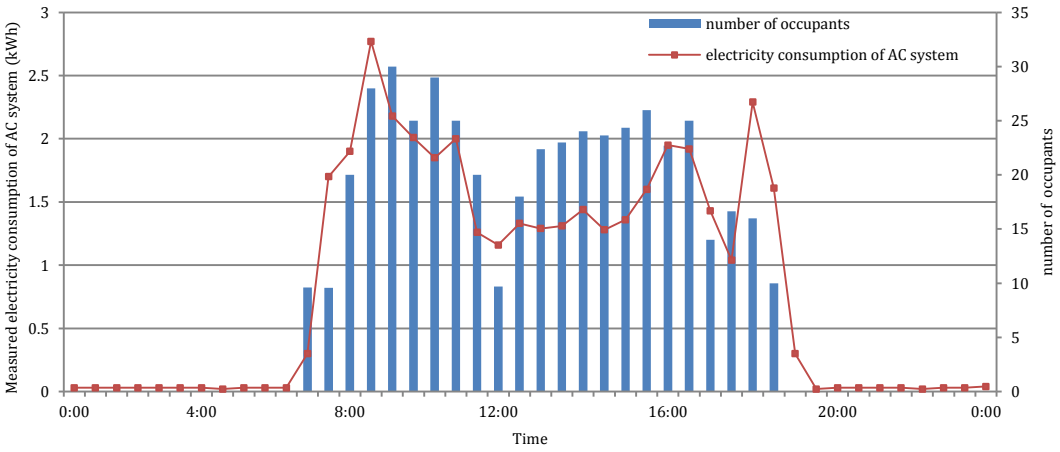


Fig. 6-4: Measured electricity consumption of AC system and true occupancy profile during workday

Table 6-5: R^2, MAPE, and RMSE of the prediction models in the training datasets (with true occupant counts)

	R^2				MAPE (%)				RMSE			
Number of inputs	9	10	16	15	9	10	16	15	9	10	16	15
FFNN	0.9035	0.9463	0.9329	0.9268	8.9297	5.3017	6.4489	7.4219	4.2430	2.3001	3.3053	3.3711
ELM	0.9222	0.9773	0.9542	0.9466	7.8592	4.1950	4.4273	7.7359	3.7112	1.5749	2.6393	4.0630
ensemble	0.9363	0.9831	0.9624	0.9048	5.7525	3.0384	3.8203	7.1314	3.2596	1.5230	2.3353	3.3028

Table 6-6: R^2, MAPE, and RMSE of the prediction models in the validation datasets (with true occupant counts)

	R^2	MAPE (%)	RMSE

Number of inputs	9	10	16	15	9	10	16	15	9	10	16	15
FFNN	0.9174	0.9398	0.9072	0.8758	7.7964	6.9675	8.6625	10.3566	3.9548	2.1590	2.7754	4.9012
ELM	0.9170	0.9501	0.9355	0.9330	7.2806	5.8652	7.7034	8.1813	3.8555	2.2792	2.9170	3.4351
ensemble	0.9315	0.9639	0.9384	0.9231	6.6259	4.7923	5.6668	8.4668	3.4996	1.9179	2.6227	3.7502

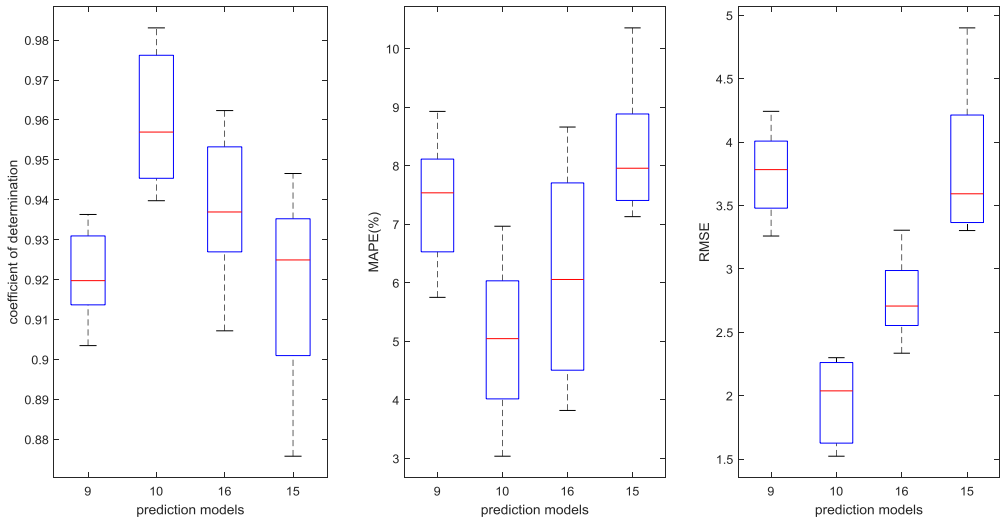


Fig. 6-5: Boxplot R^2 (left), MAPE (middle), and RMSE (right) according to different number of input parameters

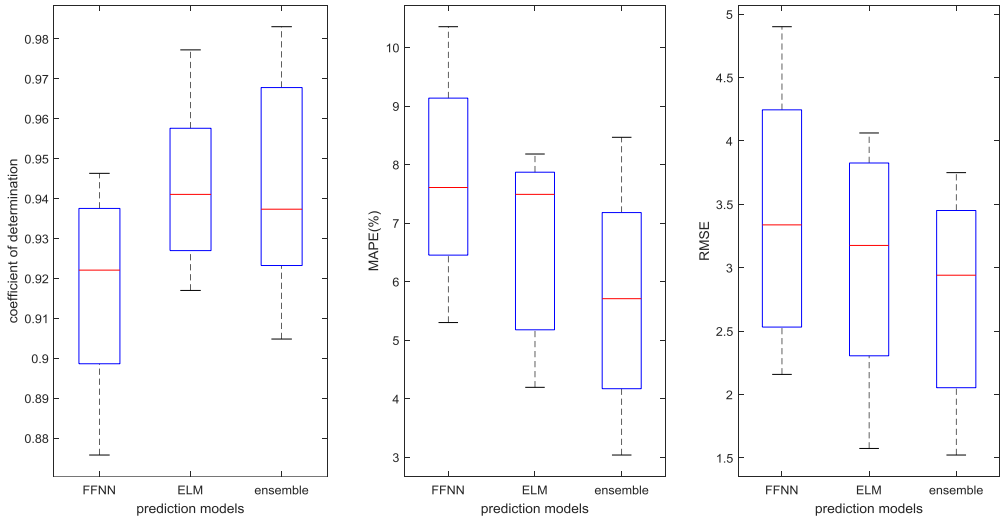


Fig. 6-6: Boxplot R^2 (left), MAPE (middle), and RMSE (right) according to different prediction models

6.6.2 Energy prediction result with estimated occupant counts

In Section 6.6.1, the ensemble model with the input of true occupant counts has been validated with high accuracy compared to other models. However, the real-time data of occupant counts is not always available for an office building. By using the BSI model, the number of occupants is blindly estimated through a non-intrusive method. In this section, the prediction performance of the prediction models with the input of the estimated number of occupants is investigated. Twelve models with different input parameters are proposed to predict the energy consumption of the AC system in the next time-step, as shown in **Table 6-7**.

Table 6-7: Input parameters of prediction models (with estimated number of occupants)

Model	Input parameters	Output parameter
FFNN10, ELM10, Ensemble10	$P(k), \hat{O}(k), L(k), S(k), F(k), T(k), z(k), z(k-1), z(k-2), sh$	$z(k+1)$
FFNN9, ELM9, Ensemble9	$P(k), L(k), S(k), F(k), T(k), z(k), z(k-1), z(k-2), sh)$	
FFNN15, ELM15, Ensemble15	$P(k), L(k), S(k), F(k), T(k), W(k), r(k), b(k), R(k), C(k), A(k), z(k), z(k-1), z(k-2), sh$	
FFNN16, ELM16, Ensemble16	$P(k), \hat{O}(k), L(k), S(k), F(k), T(k), W(k), r(k), b(k), R(k), C(k), A(k), z(k), z(k-1), z(k-2), sh$	

*$\hat{O}(k)$ is the estimated number of occupants.

Table 6-8 and **Table 6-9** show the R^2, MAPE, and RMSE of the proposed models in the training and validation datasets. The best training performance with the highest R^2 (0.9594) and lowest MAPE (4.7972%) and RMSE (2.3977) is obtained with the ensemble model, which outperforms FFNN and ELM in all configurations. As illustrated in **Table 6-8**, the overall training performance of ELM is superior to all FFNNs with different numbers of input variables. Furthermore, when using 10 variables, including the estimated number of occupants, as inputs, the FFNN, ELM, and ensemble models yield better training qualities than when using 9, 15, or 16 inputs. It is suggested in this research that using too many predictors that are highly correlated is likely to cause a decrease in the accuracy.

In the validation dataset shown in **Table 6-9**, adding the estimated number of occupants as an input variable leads to higher prediction accuracy as well. For example, R^2, MAPE, and RMSE corresponding to the ensemble model with 10 input

variables are 0.9435, 6.1852%, and 3.1280, respectively, whereas using only 9 input variables gives performance indices as 0.9315, 6.6259%, and 3.4996. Additionally, the prediction accuracy decreases with increasing numbers of input variables. It is shown that the predictor selection conducted by PCA plays a crucial role in the energy-prediction model. Therefore, using fewer input variables has several advantages, such as high accuracy and simple measurement. In addition, it is validated that the ensemble model can compensate for the predicting error of the single-FFNN or ELM model.

Table 6-5 and **Table 6-9** show the comparison of the performance of prediction models with the input of true occupant counts and the prediction models with the input of estimated number of occupants. Although the neural-network models using the estimated number of occupants as input have improved the prediction accuracy to some extent, the R^2 of the ensemble model with true occupant counts is higher (0.9639), which is strong evidence supporting that the indoor occupancy is an important factor for the electricity-consumption prediction of the AC system.

Figure 6-7 and **Fig. 6-8** show an overview of R^2, MAPE, and RMSE of the prediction models in the training and testing datasets. It can be seen that the variance of the RMSE value is the largest among these three indices. When the number of inputs is relatively large, such as 15 or 16, the RMSE value of FFNN models is extremely large in testing datasets. The results show that FFNN models probably suffer an overfitting problem when the number of inputs is large.

Table 6-8:R^2, MAPE, and RMSE of the prediction models in the training datasets (with estimated occupant counts)

	R^2				MAPE (%)				RMSE			
Number of inputs	9	10	16	15	9	10	16	15	9	10	16	15
FFNN	0.9035	0.9350	0.9326	0.9273	8.9297	6.4329	8.0790	7.4219	4.2430	3.3526	3.2626	3.3711
ELM	0.9222	0.9682	0.9400	0.9466	7.8592	5.2422	6.3590	7.7359	3.8555	3.1633	3.2824	3.4351
ensemble	0.9363	0.9594	0.9607	0.9048	5.7525	4.7972	6.0657	7.1314	3.2596	2.3977	2.8677	3.3028

Table 6-9: R^2, MAPE, and RMSE of the prediction models in the validating datasets (with estimated occupant counts)

	R^2				MAPE (%)				RMSE			
Number of inputs	9	10	16	15	9	10	16	15	9	10	16	15
FFNN	0.9174	0.9224	0.8884	0.8758	7.7964	7.2981	8.7161	10.3566	3.9548	3.7841	4.5786	4.9012
ELM	0.9170	0.9412	0.9412	0.9330	7.2806	7.0690	7.8973	8.1813	3.7112	2.4239	4.4015	4.0630
ensemble	0.9315	0.9435	0.9272	0.9231	6.6259	6.1852	7.6433	8.4668	3.4996	3.1280	3.6373	3.7502

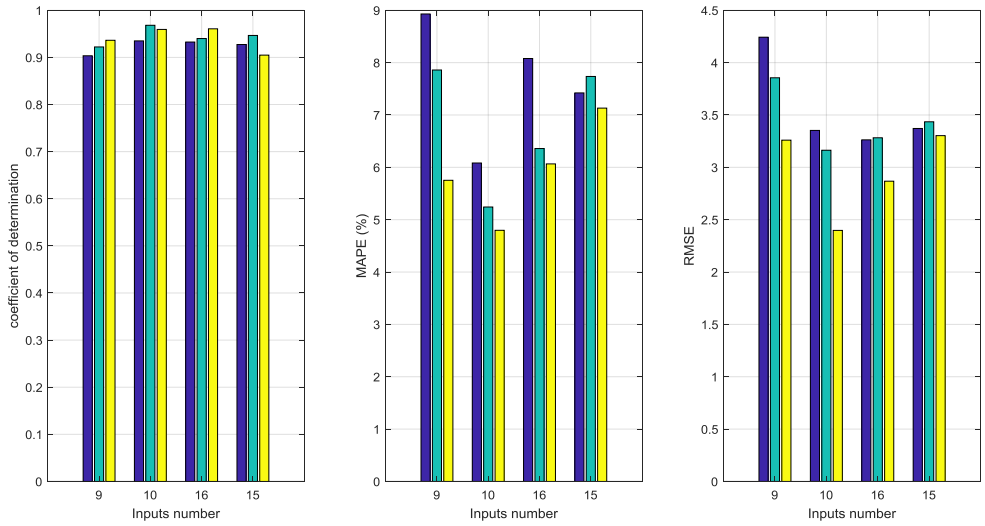


Fig. 6-7: Graphic representation of R^2 (left), MAPE (middle), and RMSE (right) of the prediction models in the training datasets

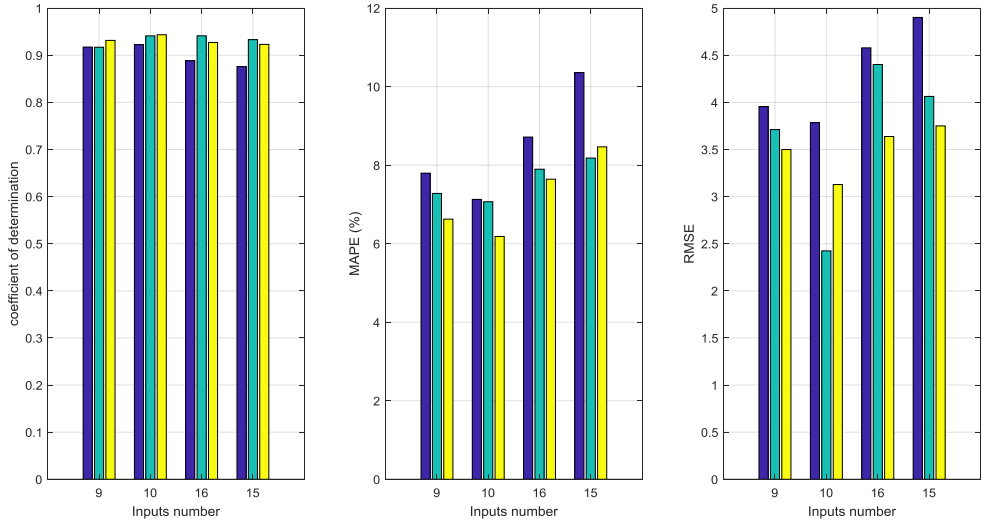


Fig. 6-8: Graphic representation of R^2 (left), MAPE (middle), and RMSE (right) of the prediction models in the testing datasets

Figure 6-9 shows the comparison results of the measured and predicted electricity consumption of AC system for training process (27 Aug. 2017 to 31 Aug. 2017). The measurement and prediction are in good agreement for the three comparative methods during the training process. However, the measured data fluctuates uncertainly during the day, especially on the mornings of the second, third, and fourth days. Hence, the absolute error between the measured and predicted values sometimes becomes large. This is because some unexpected sudden changes in input variables deteriorate the prediction performance of the models, especially when the electricity consumption becomes zero or suffers unpredictable spikes. For example, the electricity consumption of the VRV system became zero at 18:20 on the second training day, when the indoor lighting was not switched off until 18:50. Therefore, the prediction value of the electricity consumption of the VRV system fluctuated slightly and was non-zero during this half hour, because the electricity consumption of lighting was adopted as one input parameter.

Figure 6-10 shows the comparison of the measured and predicted electricity consumption of the AC system for the validation process (1 Sep. 2017). It is shown that the predicted result of the ensemble model is in the best accordance with the actual target, even though there are some fluctuations (e.g., from 11:00 to 14:00). **Table 6-10** lists the actual electricity consumption, predicted electricity consumption, and predicted error in detail. The improvement is expected because the ensemble model combines outputs from different networks, which may be offset; hence, the average error, which is the error for ensemble network, is the lowest, as shown in **Fig. 6-11**. Hence, a better result with higher accuracy is provided by the ensemble model.

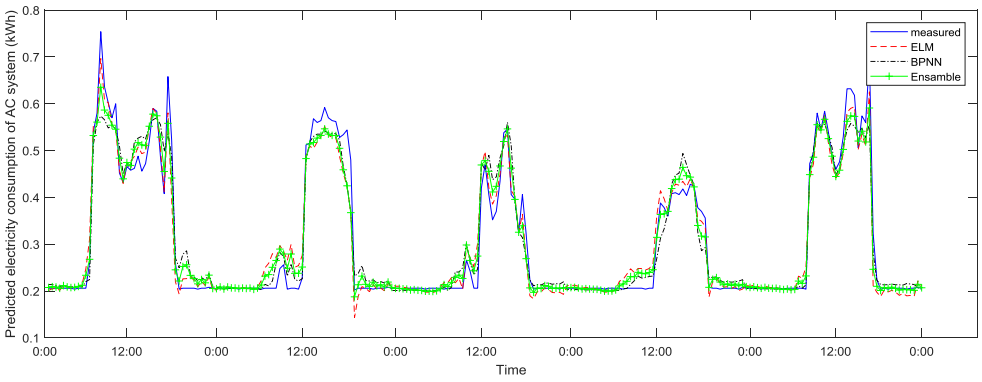


Fig. 6-9: Comparison of measured and predicted electricity consumption for training process (10 input variables)

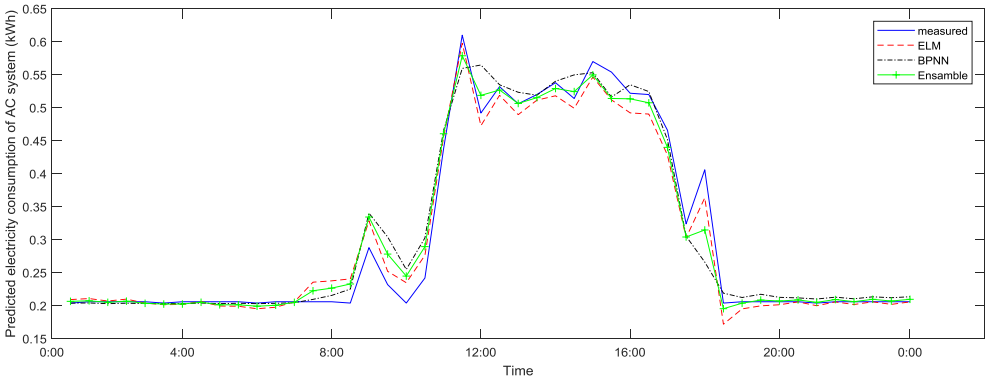


Fig. 6-10: Comparison of measured and predicted electricity consumption for validation process (10 input variables)

Table 6-10: The actual value, predicted value, and predictor error on 1 Sep. 2017 (10 input variables)

Time	Actual value (kWh)	Predicted value (kWh)			Predicted error (kWh)		
		ELM	FFNN	ensemble	ELM	FFNN	ensemble
00:30	0.206	0.210	0.204	0.207	-0.004	0.002	-0.001

01:00	0.204	0.209	0.204	0.207	-0.005	0.000	-0.003
01:30	0.206	0.211	0.204	0.207	-0.005	0.002	-0.001
02:00	0.206	0.207	0.203	0.205	-0.001	0.003	0.001
02:30	0.206	0.210	0.204	0.207	-0.004	0.002	-0.001
03:00	0.206	0.204	0.203	0.204	0.002	0.003	0.002
03:30	0.204	0.201	0.204	0.203	0.003	0.000	0.001
04:00	0.206	0.202	0.203	0.203	0.004	0.003	0.003
04:30	0.206	0.206	0.204	0.205	0.000	0.002	0.001
05:00	0.206	0.200	0.203	0.201	0.006	0.003	0.005
05:30	0.206	0.199	0.204	0.201	0.007	0.002	0.005
06:00	0.204	0.195	0.203	0.199	0.009	0.001	0.005
06:30	0.206	0.198	0.204	0.201	0.008	0.002	0.005
07:00	0.206	0.206	0.205	0.205	0.000	0.001	0.001
07:30	0.206	0.236	0.210	0.223	-0.030	-0.004	-0.017
08:00	0.206	0.238	0.216	0.227	-0.032	-0.010	-0.021
08:30	0.204	0.241	0.225	0.233	-0.037	-0.021	-0.029
09:00	0.288	0.329	0.340	0.334	-0.041	-0.052	-0.046
09:30	0.232	0.252	0.304	0.278	-0.02	-0.072	-0.046
10:00	0.204	0.235	0.255	0.245	-0.031	-0.051	-0.041
11:30	0.242	0.276	0.303	0.290	-0.034	-0.061	-0.048
11:00	0.438	0.455	0.466	0.461	-0.017	-0.028	-0.023
11:30	0.610	0.598	0.560	0.579	0.012	0.050	0.031
12:00	0.492	0.473	0.565	0.519	0.019	-0.073	-0.027
12:30	0.532	0.519	0.535	0.527	0.013	-0.003	0.005
13:00	0.506	0.490	0.523	0.506	0.016	-0.017	0.000
13:30	0.520	0.511	0.519	0.515	0.009	0.001	0.005
14:00	0.538	0.518	0.540	0.529	0.020	-0.002	0.009

14:30	0.514	0.499	0.550	0.525	0.015	-0.036	-0.011
15:00	0.570	0.547	0.553	0.550	0.023	0.017	0.020
15:30	0.554	0.512	0.516	0.514	0.042	0.038	0.040
16:00	0.522	0.492	0.535	0.514	0.030	-0.013	0.008
16:30	0.520	0.490	0.525	0.507	0.030	-0.005	0.013
17:00	0.466	0.428	0.452	0.440	0.038	0.014	0.026
17:30	0.324	0.303	0.305	0.304	0.021	0.019	0.02
18:00	0.406	0.363	0.266	0.315	0.043	0.140	0.091
18:30	0.204	0.172	0.219	0.195	0.032	-0.015	0.009
19:00	0.206	0.195	0.212	0.204	0.011	-0.006	0.002
19:30	0.206	0.200	0.217	0.209	0.006	-0.011	-0.003
20:00	0.206	0.201	0.213	0.207	0.005	-0.007	-0.001
20:30	0.206	0.206	0.212	0.209	0.000	-0.006	-0.003
21:00	0.204	0.200	0.210	0.205	0.004	-0.006	-0.001
21:30	0.206	0.206	0.213	0.209	0.000	-0.007	-0.003
22:00	0.206	0.202	0.210	0.206	0.004	-0.004	0.000
22:30	0.206	0.206	0.213	0.210	0.000	-0.007	-0.004
23:00	0.206	0.202	0.212	0.207	0.004	-0.006	-0.001
23:30	0.206	0.206	0.214	0.210	0.000	-0.008	-0.004

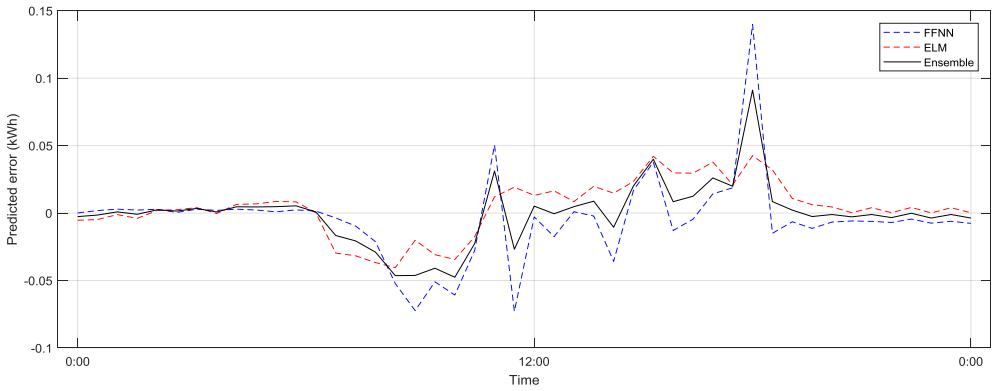


Fig. 6-11: Predicted errors of different models for validation process (10 input variables)

In order to demonstrate the merits of the proposed ensemble model on a more definite basis, **Fig. 6-12** depicts the overall validation performance of ensemble models with various numbers of input variables. Compared to ensemble models with 9 and 10 inputs, the ensemble models with 15 and 16 inputs fail to offer highly accurate prediction of electricity consumption. Specifically, the estimated energy uses of these two models are lower than the actual measurement during the day. Although high correlation of the measured and predicted results is verified by using an ensemble model with 10 inputs, the peaks observed at samples 17, 21, and 35 are not captured effectively by this model. This suggests that unusual data spikes in the testing day that were not observed during the training days are more difficult for prediction models to manage.

As for the peak forecasting shown in **Table 6-11** and **Fig. 6-10**, it is evident that the peak value (0.61 kWh) of measured electricity consumption is captured effectively by

the ELM model. Additionally, the improvements are more dramatic with ELM models, such that the $\text{MAPE}_{peak}/\text{MAPE}_{simple-peak}$ is the lowest, i.e., 1.9770%.

As in the case of overall performance, the ELM models are more able to learn peak behaviour than the FFNN and the ensemble models, whereas ensemble models provide best prediction performance in evaluating criteria, such as R^2, MAPE, and RMSE. In summary, the advantages of the various network models are different; the model selection should be determined case by case. For example, the ELM models could be used to predict the peak electricity demand. Based on the load forecasting of regional buildings, the energy market operator could determine the dispatch strategy of electricity to match the maximum supply capacity of generators and regional peak demand. In addition, the estimation of electricity usage at off-peak hours could be conducted based on ensemble models. This information is also important to guide consumers to develop their own energy-saving plans and improve the reliability of the power network at the same time.

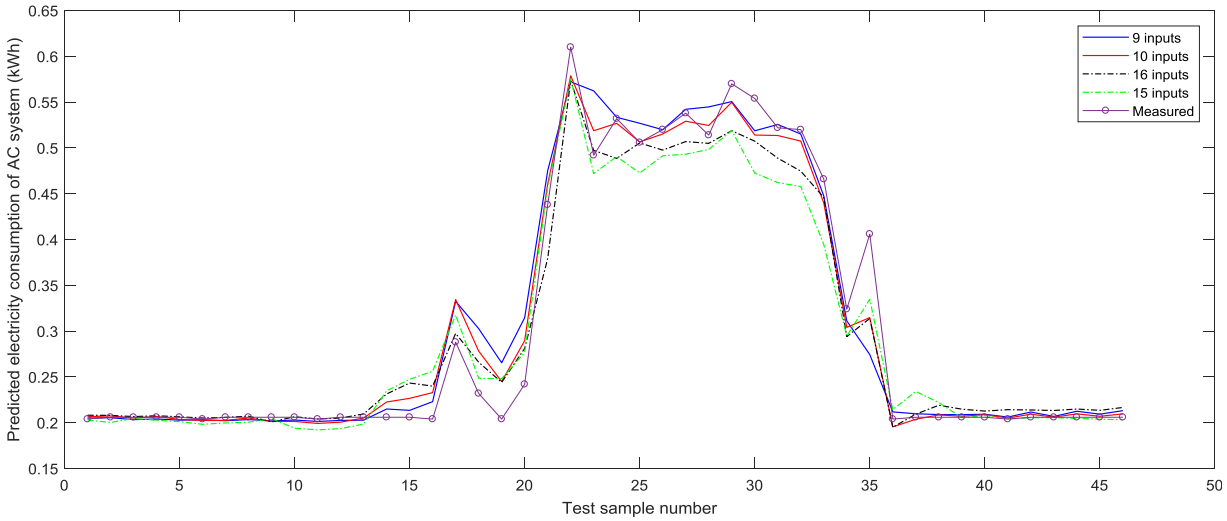


Fig. 6-12: Comparison of measured and predicted electricity consumption for ensemble models

Table 6-11: $\mathbf{MAPE_{peak}}$ and $\mathbf{MAPE_{simple-peak}}$v of the prediction models in the validating datasets

	$\mathrm{MAPE}_{peak}/\mathrm{MAPE}_{simple-peak}$			
Number of inputs	9	10	16	15
FFNN	14.7253/10.0687	8.2606/7.4109	15.8267/15.8267	13.6686/12.2969
ELM	2.0219/2.0219	1.9770/1.9770	3.8770/3.8770	2.3503/2.3503
Ensemble	6.2041/6.2041	2.6054/2.6054	5.9748/5.9748	5.6591/5.6591

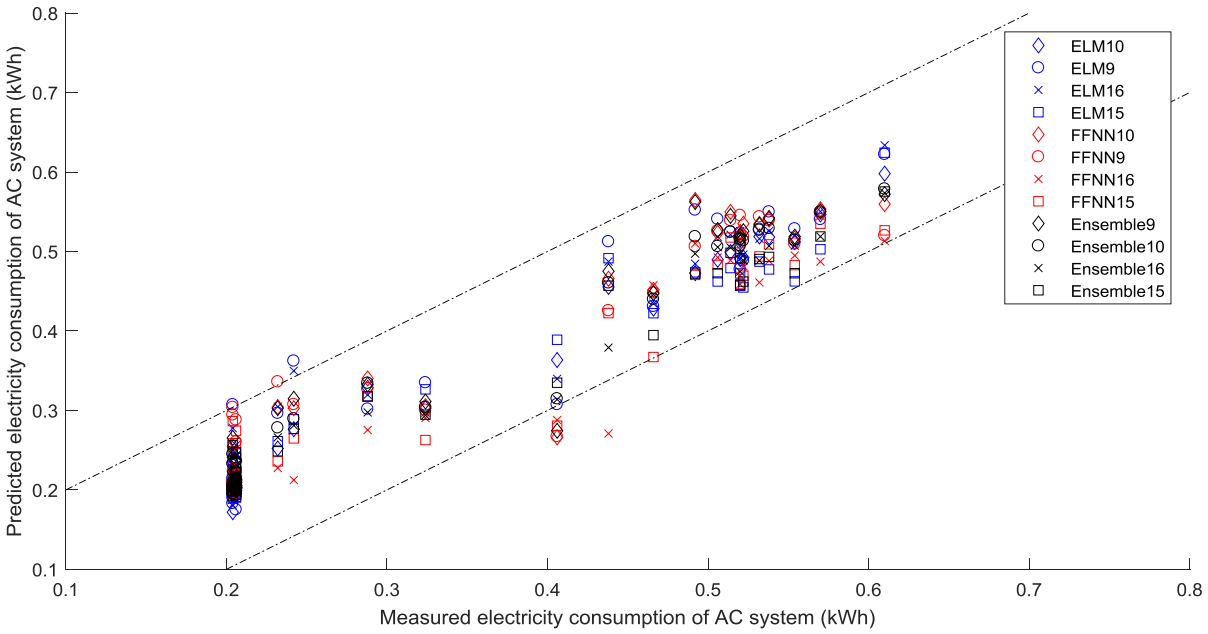


Fig. 6-13: Comparison between the measured and predicted electricity consumption of twelve models

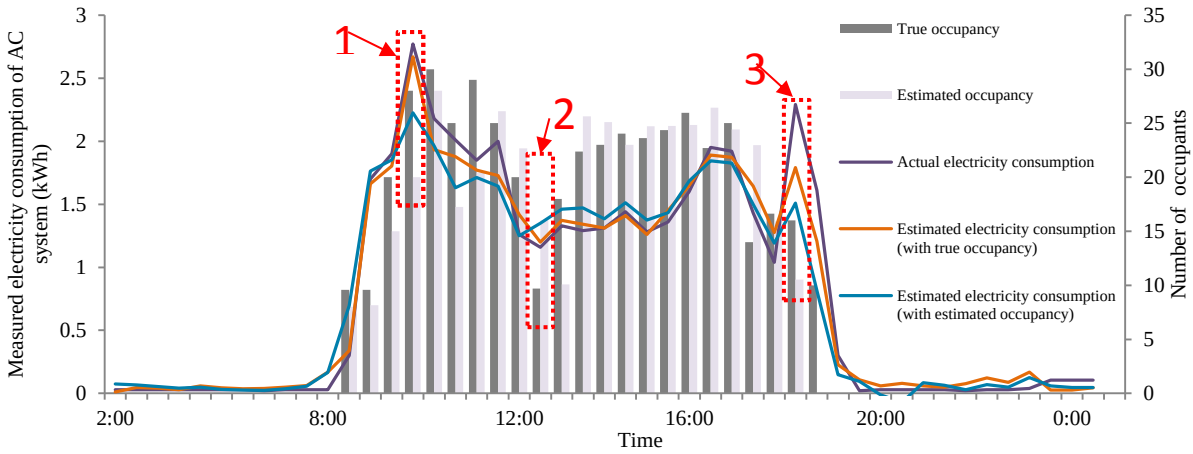


Fig. 6-14: Comparison of ensemble models (10 inputs) with true/estimated occupant counts as input parameter

Figure 6-13 shows the comparison of the measured and predicted electricity consumption of the twelve proposed models. It can be noted that, for ensemble models, 97.8% of the datasets have been included in the error range of ±10%, whereas 93.5% and 86.9% of the datasets are within the range for ELM models and FFNN models, respectively. Specifically, the validation datasets of the ensemble model with 10 inputs all fall within the error range. However, the ensemble models with 9, 15, and 16 inputs are not good compared to that with 10 inputs. The results show that the electricity-consumption prediction model requires the predictor selection conducted by PCA and the input of occupant numbers so as to capture the characteristic of the indoor load variance, which has significant influence on the energy consumption of the building.

In order to demonstrate that the effects of estimation error of occupant counts on the accuracy of energy-prediction models in a more tangible way, the predicting

performance of ensemble model with estimated occupancy as input parameter was compared to the predicting performance of ensemble model with true occupancy, which was used as benchmark model. **Figure 6-14** graphically shows the comparison results for the testing day. Differences can be noticed between the ensemble models and the benchmark model at morning peak, noon break and afternoon peak (point 1, 2 and 3). Concretely, the estimated occupant number is less than true value at 9:30 (point 1) and 18:30 (point 3), corresponding energy consumption predicted by ensemble model is slightly smaller than the benchmark model and the true value. Moreover, the time lag of occupancy estimation at noon break is obvious, hence the energy-prediction result of ensemble model is not satisfying as benchmark model during lunch time.

6.7 Chapter summary

In this research, we develop a prediction model of electricity consumption of an AC system based on feed-forward neural network, extreme learning machine, and ensemble models, with an input of the occupancy determined by BSI estimation.

To analyse some aspects of the benchmark test for the effect of structural parameters and input-selection alternatives, three studies are conducted: 1) the effect of the predictor selection conducted by principal component analysis, 2) the effect of the estimated occupancy as the supplementary input, and 3) the effect of the neural-network ensemble.

It is shown that consideration of too many supplemental input variables is likely to deteriorate the accuracy of prediction. Predictor selection conducted by principal

component analysis plays crucial role in the energy-prediction model. In addition, the consideration of extra inputs of occupancy can improve the accuracy of training and validation for the proposed models. Hence, the electricity-consumption prediction model requires the occupant number to be input so as to capture the characteristic of the indoor load variance, which has a significant influence on the energy consumption of the building. The best performance with the highest R^2 and lowest MAPE and RMSE can be obtained with the ensemble model, which is better than the feed-forward neural network and extreme learning machine in all neural-network structures. As for the peak forecasting, the extreme learning machine is better able to learn peak behaviour than the feed-forward neural network and ensemble models.

One limitation of this study is that the application of energy prediction model combining dynamic occupant profile is only validated during summer period. The model performance should be deeply investigated for one-year period since the occupancy interactions with building system and energy profiles of sub-components would be varied according to different seasons. Comprehensive study and more elaborate approaches should be carried out to investigate the model's prediction performance.

CHAPTER 7 CONCLUSION AND FUTURE WORK

7.1 Conclusion

The research has presented an in-depth investigation into the application of data-driven models in building energy analysis and occupant behaivor. The research was undertaken through a critical literature review, establishment of models concerning window behaviour/occupancy estimation/energy prediction/load profiling, and validation of these models.

The overall achievements include (1) residential behaviour model extracted based on electricity load patterns; (2) window behaivor models developed based on various algrothms, which were validated with satisfying accuracy; (3) an occupancy estimation model that could be used to compute the occupancy level blindly and (4) an energy prediction model with the input of estimated occupant counts.

7.2 Future work

A number of future works following the completion of the project reported in this thesis are proposed:

(1) Possibilities for improvements of residents' occupant behavior include: the isolation of occupant behaviors from other influencing factors—in this regard, further information needs to be collected about the physical characteristics of buildings; and increasing the size of the dataset to enhance the accuracy and reliability of model predictions. On the other hand, future research should be focused on the quantification of each end-user's contribution and on the

identification of other parameters that significantly influence occupant behavior (e.g. thermal comfort, social and economic factors, etc.). Thus, characterizing the clustering analysis-derived occupant behavior patterns of various energy sources (such as electricity, hot water, space heating etc.) and integrating them into building simulation software should be a future goal. In this regard, different levels of simplifications should be tested to identify the appropriate level of smart metering required to balance accuracy and efficiency.

(2) We highlight that the case study in this project focusing on occupant window behaviour during transition season. A next step of this research is to use an extensive database for whole-year to develop a generalizable model. In the further work, it is necessary to further explore the applicability of machine learning based model for predicting occupant window behaviour. More importantly, contextual information including occupant types (e.g. age and gender), social factors (energy-related habit and knowledge) and psychological factors should be observed and analysed.

(3) The estimation result of occupancy in this study is acceptable because the test-bed office is a closed and regular space which is easily measured by a single sensor. More sensors should be installed generally at the center and corners in office room to supply sufficient information about the real-time variance of indoor environment. A similar study conducted with a comprehensive sensor network may produce better results. Because the placement of sensors is vital, we shall investigate the optimal placement of sensors in our future works. In

addition, the variance of CO_2 emission rates of occupants should be considered in the future work, as difference can be caused by a number of factors, such as age, weights, gender and activity level.

(4) Further optimisation of building system operation should be guided by occupant interaction. The present work is the first step in addressing a challenging problem in short-term prediction models for electricity consumption of HVAC system with occupant profile. In future work, the results obtained in this study can be beneficial toward developing a predictive controller of HVAC systems for energy conservation and thermal comfort. Compared with the traditional feedback controller, the predictive controller is designed to maintain the indoor temperature with lower thermal violations and minimizing the energy consumption. Although the time step of current prediction model is 10 min, the selection of control period for improving the operating performance of HVAC system would be investigated in the future work. Hence, dedicated energy-prediction models with consideration of occupancy provide an opportunity to couple the electric grid and the building's control actions, and to be utilised by buildings and utility companies to simultaneously optimise their performance.

APPENDIX

Partial code of occupancy estimation model:

```
        %optimaze sigma
        fun = @(sigma)(0.5*log((sigma^2))+1/(2*(sigma^2)).*y'*y);
        options = optimset('PlotFcns',@optimplotfval,'MaxIter',10);
        [sigma,fval] = fminbnd(fun,0,1,options);
        save 'Untitled2', 'sigma';

        %optimaze beta_y
        options = optimset('MaxIter',1);
        beta_y = fminbnd(@f1,0,0.7,options)

        %optimaze beta_u
        options = optimset('MaxIter',1);
        beta_u = fminbnd(@f2,0,0.8,options)

        for iter = 1:1
        %optimaze beta_o
        options = optimset('MaxIter',1);
        beta_o = fminbnd(@f3,0,1,options)

        load('f1.mat')
        load('f2.mat')
        load('f3.mat')

        fun = @(o)(gu'*U*(toeplitz(I0*o,zeros(1,sets)))*go  +
        gy'*Y*(toeplitz(I0*o,zeros(1,sets)))*go +
        go'*(toeplitz(I0*o,zeros(1,sets)))'*(toeplitz(I0*o,zeros(1,sets)))*go -
        y'*(toeplitz(I0*o,zeros(1,sets)))*go);
        o0 = o;
        options = optimset('PlotFcns',@optimplotfval,'TolFun',5*10^22,'MaxIter',500);
        o = fminsearch(fun,o0,options);
        save 'Untitled2', 'o';
        end

%identitity matrix for time-shift operation
    I = eye(sets);
    A(1,(1:sets-1))=1;
    I0 =  diag(A,-1);

 %define the y and cov_y
 cov_y = (((I - (1-bu).*I0)^-1)*((I - (1-bu).*I0)^-1)');
 y = ((I - (1-bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I -
(1-bu).*I0)^-1)*e;


fun = @(bu)(log(det(sigma.*((I - (1-bu).*I0)^-1)*((I - (1-bu).*I0)^-1)'))+(((I - (1-
bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I - (1-bu).*I0)^-
1)*e)'*(((I - (1-bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I -
(1-bu).*I0)^-1)*e));
options = optimset('PlotFcns',@optimplotfval,'MaxIter', 10);
bu = fminbnd(fun,0,1,options);

fun = @(e)(log(det(sigma.*((I - (1-bu).*I0)^-1)*((I - (1-bu).*I0)^-1)'))+(((I - (1-
bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I - (1-bu).*I0)^-
1)*e)'*(((I - (1-bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I -
(1-bu).*I0)^-1)*e));
e0 = e;
options = optimset('PlotFcns',@optimplotfval,'MaxIter', 10);
e = fminsearch(fun,e0,options);

fun = @(bo)(log(det(sigma.*((I - (1-bu).*I0)^-1)*((I - (1-bu).*I0)^-1)'))+(((I - (1-
bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I - (1-bu).*I0)^-
1)*e)'*(((I - (1-bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I -
(1-bu).*I0)^-1)*e));
```

```
options = optimset('PlotFcns',@optimplotfval,'MaxIter', 10);
[bo0,fval] = fminbnd(fun,0,1,options);

fun = @(x)(log(det(sigma.*((I - (1-bu).*I0)^-1)*((I - (1-bu).*I0)^-1)'))+(((I - (1-
bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I - (1-bu).*I0)^-
1)*e)'*(((I - (1-bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I -
(1-bu).*I0)^-1)*e));
x0 = x;
options = optimset('PlotFcns',@optimplotfval,'MaxIter', 500);
x = fminsearch(fun,x0,options);

for iter = 1:1
fun = @(x)(log(det(sigma.*((I - (1-bu).*I0)^-1)*((I - (1-bu).*I0)^-1)'))+(((I - (1-
bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I - (1-bu).*I0)^-
1)*e)'*(((I - (1-bu).*I0)^-1).*bu*I0*coutdoor + ((I - (1-bu).*I0)^-1).*bo*I0*x + ((I -
(1-bu).*I0)^-1)*e));
x0 = x;
options = optimset('PlotFcns',@optimplotfval,'MaxIter', 50);
x = fminsearch(fun,x0,options);
end
```

REFERENCES

Reference for chapter 1

[1.1] Pérez-Lombard, Luis, José Ortiz, and Christine Pout. "A Review on Buildings Energy Consumption Information." Energy and buildings 40, no. 3 (2008): 394-98.

[1.2] Mathew, Paul A, Laurel N Dunn, Michael D Sohn, Andrea Mercado, Claudine Custudio, and Travis Walter. "Big-Data for Building Energy Performance: Lessons from Assembling a Very Large National Database of Building Energy Use." Applied Energy 140 (2015): 85-93.

[1.3] Nikolaidis, Yiannis, Petros A Pilavachi, and Alexandros Chletsis. "Economic Evaluation of Energy Saving Measures in a Common Type of Greek Building." Applied Energy 86, no. 12 (2009): 2550-59.

[1.4] Zhao, Hai-xiang, and Frédéric Magoulès. "A Review on the Prediction of Building Energy Consumption." Renewable and Sustainable Energy Reviews 16, no. 6 (2012): 3586-92.

[1.5] Kang, Zhaoyi, Ming Jin, and Costas J Spanos. "Modeling of End-Use Energy Profile: An Appliance-Data-Driven Stochastic Approach." Paper presented at the IECON 2014-40th Annual Conference of the IEEE Industrial Electronics Society, 2014.

[1.6] Bojić, M, and N Lukić. "Numerical Evaluation of Solar-Energy Use through Passive Heating of Weekend Houses in Yugoslavia." Renewable energy 20, no. 2 (2000): 207-22.

[1.7] Farahbakhsh, Hassan, VI Ugursal, and AS Fung. "A Residential End-Use Energy Consumption Model for Canada." International Journal of Energy Research 22, no. 13 (1998): 1133-43.

[1.8] Huang, Yu Joe, and Jim Brodrick. "A Bottom-up Engineering Estimate of the Aggregate Heating and Cooling Loads of the Entire Us Building Stock." (2000).

[1.9] Shimoda, Yoshiyuki, Takuro Fujii, Takao Morikawa, and Minoru Mizuno. "Residential End-Use Energy Simulation at City Scale." Building and environment 39, no. 8 (2004): 959-67.

[1.10] Tardioli, Giovanni, Ruth Kerrigan, Mike Oates, O'Donnell James, and Donal Finn. "Data Driven Approaches for Prediction of Building Energy Consumption at Urban Level." Energy Procedia 78 (2015): 3378-83.

[1.11] ISO. ISO Standard 12655: Energy performance of buildings – presentation of real energy use of buildings; 2013

[1.12] Hong, Tianzhen, Le Yang, David Hill, and Wei Feng. "Data and Analytics to Inform Energy Retrofit of High Performance Buildings." Applied Energy 126 (2014): 90-106.

[1.13] Nikolaou, Triantafyllia G, Dionysia S Kolokotsa, George S Stavrakakis, and Ioannis D Skias. "On the Application of Clustering Techniques for Office Buildings' Energy and Thermal Comfort Classification." IEEE Transactions on Smart Grid 3, no. 4 (2012): 2196-210.

[1.14] Mathew, Paul A, Laurel N Dunn, Michael D Sohn, Andrea Mercado, Claudine Custudio, and Travis Walter. "Big-Data for Building Energy Performance: Lessons from Assembling a Very Large National Database of Building Energy Use." Applied Energy 140 (2015): 85-93.

[1.15] Cavalheiro, José, and Paulo Carreira. "A Multidimensional Data Model Design for Building Energy Management." Advanced Engineering Informatics 30, no. 4 (2016): 619-32.

[1.16] Al-Homoud, Mohammad Saad. "Computer-Aided Building Energy Analysis Techniques." Building and environment 36, no. 4 (2001): 421-33.

[1.17] Barnaby, Charles S, and Jeffrey D Spitler. "Development of the Residential Load Factor Method for Heating and Cooling Load Calculations." ASHRAE Transactions 111, no. 1 (2005).

[1.18] Paudel, Subodh, Phuong H Nguyen, Wil L Kling, Mohamed Elmitri, Bruno Lacarrière, and Olivier Le Corre. "Support Vector Machine in Prediction of Building Energy Demand Using Pseudo Dynamic Approach." arXiv preprint arXiv:1507.05019 (2015).

[1.19] Li, Zhengwei, Yanmin Han, and Peng Xu. "Methods for Benchmarking Building Energy Consumption against Its Past or Intended Performance: An Overview." Applied Energy 124 (2014): 325-34.

[1.20] International Energy Agency, Total energy use in buildings: analysis and evaluation methods, 2013. http://www.iea-ebc.org/fileadmin/user upload/images/Pictures/EBC Annex 53 Main Report.pdf.

[1.21] D. Yan, T. Hong, IEA EBC Annex 66: Definition and Simulation of Occupant Behavior in Buildings, 2014. http://www.annex66.org/.

[1.22] Belafi, Zsofia Deme, Federica Naspi, Marco Arnesano, Andras Reith, and Gian Marco Revel. "Investigation on window opening and closing behavior in schools through measurements and surveys: A case study in Budapest." Building and Environment 143 (2018): 523-531.

[1.23] Emery, A. F., and C. J. Kippenhan. "A long term study of residential home heating consumption and the effect of occupant behavior on homes in the Pacific Northwest constructed according to improved thermal standards." Energy 31, no. 5 (2006): 677-693.

[1.24] Rijal, Hom B., Paul Tuohy, Michael A. Humphreys, J. Fergus Nicol, Aizaz Samuel, and Joseph Clarke. "Using results from field surveys to predict the effect of open windows on thermal comfort and energy use in buildings." Energy and buildings 39, no. 7 (2007): 823-836.

[1.25] Andersen, Rune Vinther, Jørn Toftum, Klaus Kaae Andersen, and Bjarne W. Olesen. "Survey of occupant behaviour and control of indoor environment in Danish dwellings." Energy and Buildings 41, no. 1 (2009): 11-16.

[1.26] Hu, Shan, Da Yan, Siyue Guo, Ying Cui, and Bing Dong. "A survey on energy consumption and energy usage behavior of households and residential building in urban China." Energy and Buildings 148 (2017): 366-378.

[1.27] Takasu, Marina, Ryozo Ooka, Hom B. Rijal, Madhavi Indraganti, and Manoj Kumar Singh. "Study on adaptive thermal comfort in Japanese offices under various operation modes." Building and Environment 118 (2017): 273-288.

[1.28] Wei, Yixuan, Xingxing Zhang, Yong Shi, Liang Xia, Song Pan, Jinshun Wu, Mengjie Han, and Xiaoyun Zhao. "A review of data-driven approaches for prediction and classification of building energy consumption." Renewable and Sustainable Energy Reviews 82 (2018): 1027-1047.

[1.29] Zhou, Xin, Da Yan, Tianzhen Hong, and Xiaoxin Ren. "Data analysis and stochastic modeling of lighting energy use in large office buildings in China." Energy and Buildings 86 (2015): 275-287.

[1.30] Todd, Annika, Elizabeth Stuart, Steven R. Schiller, and Charles A. Goldman. "Evaluation, measurement, and verification (EM&V) of residential behavior-based energy efficiency programs: Issues and recommendations." State and Local Energy Efficiency Action Network (2012).

[1.31] Sun, Kaiyu, and Tianzhen Hong. "A simulation approach to estimate energy savings potential of occupant behavior measures." Energy and Buildings 136 (2017): 43-62.

[1.32] Diao, Longquan, Yongjun Sun, Zejun Chen, and Jiayu Chen. "Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation." Energy and Buildings 147 (2017): 47-66.

[1.33] D'Oca, Simona, Valentina Fabi, Stefano P. Corgnati, and Rune Korsholm Andersen. "Effect of thermostat and window opening occupant behavior models on energy use in homes." In Building Simulation, vol. 7, no. 6, pp. 683-694. Tsinghua University Press, 2014.

[1.34] D'Oca, Simona. "Influence of occupants' behaviour on heating energy consumption and thermal comfort in residential buildings." PhD diss., Politecnico di Torino, 2012.

[1.35] Zhou, Xin, Da Yan, Xiaohang Feng, Guangwei Deng, Yiwen Jian, and Yi Jiang. "Influence of household air-conditioning use modes on the energy performance of residential district cooling systems." In Building Simulation, vol. 9, no. 4, pp. 429-441. Tsinghua University Press, 2016.

[1.36] Hong, Tianzhen, Sarah C. Taylor-Lange, Simona D'Oca, Da Yan, and Stefano P. Corgnati. "Advances in research and applications of energy-related occupant behavior in buildings." Energy and buildings 116 (2016): 694-702.

[1.37] Schakib-Ekbatan, Karin, Fatma Zehra Cakıcı, Marcel Schweiker, and Andreas Wagner. "Does the occupant behavior match the energy concept of the building?–Analysis of a German naturally ventilated office building." Building and Environment 84 (2015): 142-150.

[1.38] Daniel, Lyrian, Veronica Soebarto, and Terence Williamson. "House energy rating schemes and low energy dwellings: The impact of occupant behaviours in Australia." Energy and Buildings 88 (2015): 34-44.

[1.39] Wang, Liping, and Steve Greenberg. "Window operation and impacts on building energy consumption." Energy and Buildings 92 (2015): 313-321.

[1.40] Santin, Olivia Guerra, Laure Itard, and Henk Visscher. "The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock." Energy and buildings 41, no. 11 (2009): 1223-1232.

[1.41] Annex 66 Final Report: Definition and Simulation of Occupant Behavior in Buildings. https://annex66.org/?q=Publication/2018FinalReport/, 2018 (accessed May 2018)

[1.42] Fabi, Valentina, Rune Vinther Andersen, Stefano Paolo Corgnati, Bjarne W. Olesen, and Marco Filippi. "Description of occupant behaviour in building energy simulation: state-of-art and concepts for improvements." In 12th Conference of International Building Performance Simulation Association, vol. 14, pp. 2882-2889. 2011.

[1.43] Yan, Da, William O'Brien, Tianzhen Hong, Xiaohang Feng, H. Burak Gunay, Farhang Tahmasebi, and Ardeshir Mahdavi. "Occupant behavior modeling for building performance simulation: Current state and future challenges." Energy and Buildings 107 (2015): 264-278.

[1.44] Schweiker, Marcel. "Understanding occupants' behaviour for energy efficiency in buildings." Current Sustainable/Renewable Energy Reports 4, no. 1 (2017): 8-14.

[1.45] Hong, Tianzhen, Sarah C. Taylor-Lange, Simona D'Oca, Da Yan, and Stefano P. Corgnati. "Advances in research and applications of energy-related occupant behavior in buildings." Energy and buildings 116 (2016): 694-702.

[1.46] Nord, Natasa, Tymofii Tereshchenko, Live Holmedal Qvistgaard, and Ivar S. Tryggestad. "Influence of occupant behavior and operation on performance of a residential Zero Emission Building in Norway." Energy and Buildings 159 (2018): 75-88.

[1.47] Vasilyev, A., and I. Yarmoshenko. "Effect of energy-efficient measures in building construction on indoor radon in Russia." Radiation protection dosimetry 174, no. 3 (2017): 419-422.

[1.48] Shi G Z. Energy-Efficient Measures in Architecture and its Application[J]. Journal of Anyang Institute of Technology, 2006.

[1.49] Yong-Zheng Y I. Energy-efficiency design and measures of architecture[J]. Shanxi Architecture, 2007.

[1.50] Saari, Arto, Targo Kalamees, Juha Jokisalo, Rasmus Michelsson, Kari Alanne, and Jarek Kurnitski. "Financial viability of energy-efficiency measures in a new detached house design in Finland." Applied energy 92 (2012): 76-83.

[1.51] Heravi, Gholamreza, and Mahsa Qaemi. "Energy performance of buildings: The evaluation of design and construction measures concerning building energy efficiency in Iran." Energy and Buildings 75 (2014): 456-464.

[1.52] Huang, Yu, Jian-lei Niu, and Tse-ming Chung. "Study on performance of energy-efficient retrofitting measures on commercial building external walls in cooling-dominant cities." Applied energy 103 (2013): 97-108.

[1.53] Vine, Edward L., Paul P. Craig, James C. Cramer, Thomas M. Dietz, Bruce M. Hackett, Dan J. Kowalczyk, and Mark D. Levine. "The applicability of energy models to occupied houses: Summer electric use in Davis." Energy 7, no. 11 (1982): 909-925.

[1.54] Yi Jiang, Qingpeng Wei, Xiu Yang. Results based on data-scientific development of building energy conservation. Construction technology, 2009, 7:20-24.

[1.55] Yi Jiang, Xiu Yang. The state of building energy consumption and problems in building energy conservation in China. China construction, 2006, 2:12-18.

[1.56] Yu Liu. Classification and system development of green building tools [J]. Journal of architecture, 2006, 07: 36-40.

[1.57] Crawley, Drury B., Jon W. Hand, Michaël Kummert, and Brent T. Griffith. "Contrasting the capabilities of building energy performance simulation programs." Building and environment 43, no. 4 (2008): 661-673.

[1.58] De Wilde, Pieter. "The gap between predicted and measured energy performance of buildings: A framework for investigation." Automation in Construction 41 (2014): 40-49.

[1.59] Menezes, Anna Carolina, Andrew Cripps, Dino Bouchlaghem, and Richard Buswell. "Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap." Applied energy 97 (2012): 355-364.

[1.60] Andrews, Clinton J., Daniel Yi, Uta Krogmann, Jennifer A. Senick, and Richard E. Wener. "Designing buildings for real occupants: An agent-based approach." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 41, no. 6 (2011): 1077-1091.

[1.61] Meyers, Robert J., Eric D. Williams, and H. Scott Matthews. "Scoping the potential of monitoring and control technologies to reduce energy use in homes." Energy and Buildings 42, no. 5 (2010): 563-569.

Reference for chapter 2

[2.1] Kuo W J, Chang R F, Chen D R, et al. Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. Breast Cancer Research and Treatment, 2001, 66(1):51.

[2.2] Sides J. The Victory Lab: The Secret Science of Winning Campaigns. Public Opinion Quarterly, 2014, 78(S1):363-364.

[2.3] Alhamazani K, Ranjan R, Mitra K, et al. An overview of the commercial cloud monitoring tools: research dimensions, design issues, and state-of-the-art. Computing, 2015, 97(4):357-377.

[2.4] Magoules F, Zhao H X. Data Mining and Machine Learning in Building Energy Analysis. John Wiley & Sons, Inc. 2016.

[2.5] Kalogirou S A. Artificial neural networks in renewable energy systems applications: a review. Renewable & Sustainable Energy Reviews, 2001, 5(4):373-401.

[2.6] Kalogirou S A, Bojic M. Artificial neural networks for the prediction of the energy consumption of a passive solar building. Energy, 2000, 25(5):479–491.

[2.7] Vapnik V, Golowich S E, Smola A. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing[C]. Advances in Neural Information Processing Systems 9. 1996:281--287.

[2.8] Zhao H X, Magoulès F. Parallel Support Vector Machines Applied to the Prediction of Multiple Buildings Energy Consumption. Journal of Algorithms & Computational Technology, 2010, 4(2):231-250.

[2.9] Dong B, Cao C, Lee S E. Applying support vector machines to predict building energy consumption in tropical region. Energy & Buildings, 2005, 37(5):545-553.

[2.10] Li Q, Meng Q, Cai J, et al. Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. Energy Conversion & Management, 2009, 50(1):90-96.

[2.11] Swan L G, Ugursal V I. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. Renewable & Sustainable Energy Reviews, 2009, 13(8):1819-1835.

[2.12] Quinlan, J. R. Induction of decision trees Machine Learning. in Data: Goals and General Description of the IN L.EN System." in (1986):257--264.

[2.13] Goldberg D E. The Genetic Algorithm Approach: Why, How, and What Next? [M]. Adaptive and Learning Systems. Springer US, 1986:247-253.

[2.14] Beyer H G. Evolutionary algorithms in noisy environments: theoretical issues and guidelines for practice. Computer Methods in Applied Mechanics & Engineering, 2000, 186(2–4):239-267.

[2.15] Panapakidis I P, Papadopoulos T A, Christoforidis G C, et al. Pattern recognition algorithms for electricity load curve analysis of buildings. Energy & Buildings, 2014, 73(2):137-145.

[2.16] J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 1973, 3(3):32-57.

[2.17] Vesanto, J., and E. Alhoniemi. Clustering of the self-organizing map. IEEE Transactions on Neural Networks 11.3(2000):586.

[2.18] Setiawan A, Koprinska I, Agelidis V G. Very short-term electricity load demand forecasting using support vector regression[C]// International Joint Conference on Neural Networks, IJCNN 2009, Atlanta, Georgia, Usa, 14-19 June. DBLP, 2009:2888-2894.

[2.19] J.L. Mathieu, P.N. Price, S. Kiliccote, M.A. Piette, Quantifying changes in building electricity use, with application to demand response, IEEE Transactions onSmart Grid 2 (2011) 507–518.

[2.20] Neto A H, Fiorelli F A S. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. Energy & Buildings, 2008, 40(12):2169-2176.

[2.21] Sözen A, Arcaklioglu E. Prediction of net energy consumption based on economic indicators (GNP and GDP) in Turkey. Energy Policy, 2007, 35(10):4981-4992.

[2.22] Yang J, Rivard H, Zmeureanu R. On-line building energy prediction using adaptive artificial neural networks. Energy & Buildings, 2005, 37(12):1250-1259.

[2.23] Canyurt O E, Ozturk H K, Hepbasli A, et al. Estimating the Turkish residential–commercial energy output based on genetic algorithm (GA) approaches. Energy Policy, 2005, 33(8):1011-1019.

[2.24] An N, Zhao W, Wang J, et al. Using multi-output feedforward neural network with empirical mode decomposition based signal filtering for electricity demand forecasting. Energy, 2013, 49(1):279-288.

[2.25] Yezioro A, Dong B, Leite F. An applied artificial intelligence approach towards assessing building performance simulation tools. Energy & Buildings, 2008, 40(4):612-620.

[2.26] Yan C W, Yao J. Application of ANN for the prediction of building energy consumption at different climate zones with HDD and CDD[C]// International Conference on Future Computer and Communication. IEEE, 2010:V3-286-V3-289.

[2.27] Yokoyama R, Wakui T, Satake R. Prediction of energy demands using neural network with model identification by global optimization. Energy Conversion & Management, 2009, 50(2):319-327.

[2.28] Olofsson T, Andersson S. Overall heat loss coefficient and domestic energy gain factor for single-family buildings. Building & Environment, 2002, 37(11):1019-1026.

[2.29] Li X, Deng Y, Ding L, et al. Building cooling load forecasting using fuzzy support vector machine and fuzzy C-mean clustering[C]// International Conference on Computer and Communication Technologies in Agriculture Engineering(cctae 2010)(volume. 2010:438-441.

[2.30] Li Q, Meng Q, Cai J, et al. Applying support vector machine to predict hourly cooling load in the building. Applied Energy, 2009, 86(10):2249-2256.

[2.31] Li, Zhengwei, and G. Huang. Re-evaluation of building cooling load prediction models for use in humid subtropical area. Energy & Buildings 62.3(2013):442–449..

[2.32] Amjady N. Short-term hourly load forecasting using time-series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(3):498-505.

[2.33] Mejri O, Barrio E P D, Ghrab-Morcos N. Energy performance assessment of occupied buildings using model identification techniques. Energy & Buildings, 2011, 43(2):285-299.

[2.34] Wauman B, Breesch H, Saelens D. Evaluation of the accuracy of the implementation of dynamic effects in the quasi steady-state calculation method for school buildings. Energy & Buildings, 2013, 65(10):173-184.

[2.35] Tso G K F, Yau K K W. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 2007, 32(9):1761-1768.

[2.36] Yu Z, Haghighat F, Fung B C M, et al. A decision tree method for building energy demand modeling. Energy & Buildings, 2010, 42(10):1637-1646.

[2.37] Sadeghi H, Zolfaghari M, Heydarizade M. Estimation of Electricity Demand in Residential sector using Genetic Algorithm Approach. 2011.

[2.38] Azadeh A, Ghaderi S F, Tarverdian S, et al. Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption.. Applied Mathematics & Computation, 2007, 186(2):1731–1741.

[2.39] Li K, Su H. Forecasting building energy consumption with hybrid genetic algorithm-hierarchical adaptive network-based fuzzy inference system. Energy & Buildings, 2010, 42(11):2070-2076.

[2.40] Tsekouras G J, Hatziargyriou N D, Dialynas E N. Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers. IEEE Transactions on Power Systems, 2007, 22(3):1120-1128.

[2.41] Xiao H, Wei Q, Jiang Y. The reality and statistical distribution of energy consumption in office buildings in China. Energy & Buildings, 2012, 50(50):259–265.

[2.42] Heidarinejad M, Dahlhausen M, Mcmahon S, et al. Cluster analysis of simulated energy use for LEED certified U.S. office buildings. Energy & Buildings, 2014, 85:86-97.

[2.43] Arambula Lara R, Cappelletti F, Romagnoni P, et al. Selection of Representative Buildings through Preliminary Cluster Analysis. 2014.

[2.44] Tiedemann K H. Using conditional demand analysis to estimate residential energy use and energy savings. Proceedings of the Cdeee, 2007.

[2.45] Aydinalp-Koksal M, Ugursal V I. Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector. Applied Energy, 2008, 85(4):271-296.

[2.46] Aydinalp M, Ugursal V I, Fung A S. Modeling of the appliance, lighting, and space-cooling energy consumptions in the residential sector using neural networks. Applied Energy, 2002, 71(2):87-110.

[2.47] Caputo P, Costa G, Ferrari S. A supporting method for defining energy strategies in the building sector at urban scale. Energy Policy, 2013, 55(249):261-270.

[2.48] Larivière, Isabelle, and G. Lafrance. Modelling the electricity consumption of cities: effect of urban density. Energy Economics21.1(1999):53-66.

[2.49] Mastrucci A, Baume O, Stazi F, et al. Estimating energy savings for the residential building stock of an entire city: A GIS-based statistical downscaling approach applied to Rotterdam. Energy & Buildings, 2014, 75(2):358–367.

[2.50] Howard B, Parshall L, Thompson J, et al. Spatial distribution of urban building energy consumption by end use. Energy & Buildings, 2011, 45:141-151.

[2.51] Jones P, Patterson J, Lannon S. Modelling the built environment at an urban scale—Energy and health impacts in relation to housing. Landscape & Urban Planning, 2007, 83(1):39-49.

[2.52] Yamaguchi, Y., Y. Shimoda, and M. Mizuno. Proposal of a modeling approach considering urban form for evaluation of city level energy management. Energy & Buildings 39.5(2007):580-592.

[2.53] Fonseca J A, Schlueter A. Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. Applied Energy, 2015, 142(15 March 2015):247-265.

[2.54] T. Nikolaou, D. Kolokotsa, G. Stavrakakis. Review on methodologies for energy benchmarking, rating and classification of buildings. Advances in Building Energy Research, 2011, 5(1):53-70.

[2.55] Chung W, Hui Y V, Lam Y M. Benchmarking the energy efficiency of commercial buildings. Applied Energy, 2005, 83(1):1-14.

[2.56] Efron B, Tibshirani R. An introduction to bootstrap. New York: Chapman & Hall; 1993.

[2.57] Yalcintas M. An energy benchmarking model based on artificial neural network method with a case example for tropical climates. International Journal of Energy Research, 2006, 30(14):1158-1174.

[2.58] Yalcintas M, Ozturk U A. An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database[M]// International Journal of Energy Research. 2006:412–421.

[2.59] Santamouris M, Mihalakakou G, Patargias P, et al. Using intelligent clustering techniques to classify the energy performance of school buildings. Energy & Buildings, 2007, 39(1):45-51.

[2.60] Park H S, Lee M, Kang H, et al. Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques. Applied Energy, 2016, 173:225-237.

[2.61] Wang E. Benchmarking whole-building energy performance with multi-criteria technique for order preference by similarity to ideal solution using a selective objective-weighting approach. Applied Energy, 2015, 146:92-103.

[2.62] Yalcintas M. Energy-savings predictions for building-equipment retrofits. Energy & Buildings, 2008, 40(12):2111-2120.

[2.63] Juan, Yi Kai, et al. GA-based decision support system for housing condition assessment and refurbishment strategies. Automation in Construction 18.4(2009):394-401.

[2.64] Asadi E, Silva M G D, Antunes C H, et al. Multi-objective optimization for building retrofit: A model using genetic algorithm and artificial neural network and an application. Energy & Buildings, 2014, 81(na):na.

[2.65] Lannon S, Georgakaki A, Macdonald S. Modelling urban scale retrofit, pathways to 2050 low carbon residential building stock. Ibpsa, 2013.

[2.66] Wang, Chuang, et al. A generalized probabilistic formula relating occupant behavior to environmental conditions. Building and Environment. 95 (2016): 53-62.

[2.67] Feng, Xiaohang, et al. A preliminary research on the derivation of typical occupant behavior based on large-scale questionnaire surveys. Energy and Buildings. 117 (2016): 332-340

[2.68] Rory V. Jones , Alba Fuertes, Elisa Gregori, Alberto Giretti, Stochastic behavioural models of occupants' main bedroom window operation for UK residential buildings. Building and Environment. 118 (2017) 144-158

[2.69] Bongchan Jeong, Jae-Weon Jeong, J.S. Park, Occupant behavior regarding the manual control of windows in residential buildings. Energy and Buildings. 127 (2016) 206–216

[2.70] Rune Andersen, Valentina Fabi, Jorn Toftum , Stefano P. Corgnati , Bjarne W. Olesen, Window opening behaviour modelled from measurements in Danish dwellings. Building and Environment. 69 (2013) 101-113

[2.71] Davide Calì, Rune Korsholm Andersen, Dirk Müller, Bjarne W. Olesen, Analysis of occupants' behavior related to the use of windows in German households. Building and Environment. 103 (2016) 54-69

[2.72] Federica Naspi , Marco Arnesano , Lorenzo Zampetti , Francesca Stazi , Gian Marco Revel , Marco D'Orazi, Experimental study on occupants' interaction with windows and lights in Mediterranean offices during the non-heating season. Building and Environment. 127 (2018) 221–238

[2.73] Francesca Stazi, Federica Naspi, Marco D'Orazio, Modelling window status in school classrooms. Results from a case study in Italy. Building and Environment. 111 (2017) 24-32

[2.74] Yufan Zhang, Peter Barrett, Factors influencing the occupants' window opening behaviour in a naturally ventilated office building. Building and Environment. 50 (2012) 125-134

[2.75] Valentina Fabi, Rune Korsholm Andersen, Stefano Corgnati, Verification of stochastic behavioural models of occupants' interactions with windows in residential buildings. Building and Environment. 94 (2015) 371-383

[2.76] Shen Wei, Richard Buswell, Dennis Loveday, Factors affecting 'end-of-day' window position in a non-air-conditioned office building. Energy and Buildings. 62 (2013) 87–96

[2.77] Geun Young Yun, Paul Tuohy, Koen Steemers, User behaviour of window-control in offices during summer and winter. Energy and Buildings. 41(5). pp. 489-499

[2.78] R. Fritsch et al. A Stochastic Model of User Behaviour Regarding Ventilation. Building and Environment. 25 (1990) 173-181

[2.79] Geun Young Yun, Paul Tuohy, Koen Steemers, Thermal performance of a naturally ventilated building using a combined algorithm of probabilistic occupant behaviour and deterministic heat and mass balance models. Energy and Buildings. 41 (2009) 489–499

[2.80] Frédéric Haldi, Darren Robinson, Interactions with window openings by office occupants. Building and Environment. 44 (2009) 2378–2395

[2.81] Davide Calì, Mark Thomas Wesseling, Dirk Müller, WinProGen: A Markov-Chain-based stochastic window status profile generator for the simulation of realistic energy performance in buildings. Building and Environment. 136 (2018) 240–258

[2.82] Song Pana, Yiye Han, Shen Wei, Yixuan Wei,∗ , Liang Xia, Lang Xie, Xiangrui Kong, Wei Yu, A model based on Gauss Distribution for predicting window behavior in building. Building and Environment 149 (2019) 210–219

[2.83] Romana Markovic, Eva Grintal, Daniel Wölki, Jérôme Frisch, Christoph van Treeck, Window Opening Model using Deep Learning Methods. Building and Environment. 145 (2018) 319–329

[2.84] Verena M. Barthelmes, Yeonsook Heo, Valentina Fabi, Stefano P. Corgnati, Exploration of the Bayesian Network framework for modelling window control behavior. Building and Environment 126 (2017) 318–330

[2.85] Simona D'Oca, Tianzhen Hong, A data-mining approach to discover patterns of window opening and closing behaviour in offices. Building and Environment 82 (2014) 726-739

[2.86] Farhang Tahmasebi, Ardeshir Mahdavi, An inquiry into the reliability of window operation models in building performance simulation. Building and Environment. 105 (2016) 343-357

[2.87] SongPan,Yingzi Xiong, Yiye Han, Xingxing Zhang, Liang Xia, Shen Wei, Jinshun Wu, Mengjie Han, A study on influential factors of occupant window-opening behavior in an office building in China. Building and Environment. 133 (2018) 41–50

[2.88] Mingyao Yao, Bin Zhao, Factors affecting occupants' interactions with windows in residential buildings in Beijing, China. Procedia Engineering. 205 (2017) 3428–3434

[2.89] Mingyao Yao, Bin Zhao, Window opening behavior of occupants in residential buildings in Beijing. Building and Environment 124 (2017) 441-449

[2.90] Zhenni Shi, Hua Qian, Xiaohong Zheng, Zhengfei Lv, Yuguo Li, Li Liu, Peter V. Nielsen, Seasonal variation of window opening behaviors in two naturally ventilated hospital wards. Building and Environment. 130 (2018) 85–93

[2.91] D. Liu, X. Guan, Y. Du, Q. Zhao, Measuring indoor occupancy in intelligent buildings using the fusion of vision sensors, Meas. Sci. Technol. 24 (7) (2013)074023.

[2.92] Y. Benezeth, H. Laurent, B. Emile, C. Rosenberger, Towards a sensor for detecting human presence and characterizing activity, Energy Build. 43 (2) (2011) 305–314.

[2.93] Wang, Fulin, et al. Predictive Control of Indoor Environment Using Occupant Number Detected by Video Data and CO 2, Concentration. Energy & Buildings 145(2017).

[2.94] P. Liu, S.K. Nguang, A. Partridge, Occupancy inference using pyroelectric infrared sensors through hidden markov models, IEEE Sens. J. 16 (4) (2016)1062–1068.

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

[2.96] Gunay, H Burak, et al. DETECTING OCCUPANTS' PRESENCE IN OFFICE SPACES: A CASE STUDY. Esim 2016.

[2.97] Jin, Ming, R. Jia, and C. Spanos. Virtual Occupancy Sensing: Using Smart Meters to Indicate Your Presence. IEEE Transactions on Mobile Computing PP.99(2017):1-1.

[2.98] S. Depatla, A. Muralidharan, Y. Mostofi, Occupancy estimation using only WiFi power measurements, IEEE J. Sel. Areas Commun. 33 (7) (2015) 1381–1393.

[2.99] N. Li, G. Calis, B. Becerik-Gerber, Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations, Autom. Constr. 24(2012) 89–99.

[2.100] Zhao, Yang, et al. Virtual occupancy sensors for real-time occupancy information in buildings. Building & Environment 93.2(2015):9-20.

[2.101] K. Padmanabh, V.A. Malikarjuna, S. Sen, S.P. Katru, A. Kumar, S.K. Vuppala, S.Paul, et al., iSense: a wireless sensor network based conference room management system, in: Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-efficiency in Buildings, ACM, 2009, pp.37–42.

[2.102] Shih, Oliver, and A. Rowe. Occupancy estimation using ultrasonic chirps. ACM/IEEE Sixth International Conference on Cyber-Physical Systems ACM, 2015:149-158.

[2.103] Amayri, Manar, et al. Estimating occupancy in heterogeneous sensor environment. Energy & Buildings 129(2016):46-58.

[2.104] Mckenna, Eoghan, M. Krawczynski, and M. Thomson. Four-state domestic building occupancy model for energy demand simulations. Energy & Buildings 96.8(2015):30-39

[2.105] Richardson, Ian, M. Thomson, and D. Infield. A high-resolution domestic building occupancy model for energy demand simulations. Energy & Buildings 40.8(2008):1560-1566.

[2.106] Chen, Zhenghua, J. Xu, and Y. C. Soh. Modelling regular occupancy in commercial buildings using stochastic models. Energy & Buildings 103(2015):216-223

[2.107] Andersen, Philip Delff, et al. Dynamic modelling of presence of occupants using inhomogeneous Markov chains. Energy & Buildings 69.69(2014):213-223.

[2.108] Erickson, Varick L., M. Á. Carreira-Perpiñán, and A. E. Cerpa. OBSERVE: Occupancy-based system for efficient reduction of HVAC energy. International Conference on Information Processing in Sensor Networks IEEE, 2011:258-269.

[2.109] Dong, Bing, and K. P. Lam. A real-time model predictive control for building heating and cooling systems based on the occupancy behaviour pattern detection and local weather forecasting. Building Simulation 7.1(2014):89-106.

[2.110] Mahdavi, Ardeshir, and F. Tahmasebi. Predicting people's presence in buildings: An empirically based model performance analysis. Energy & Buildings 86(2015):349-355.

[2.111] Chen, Zhenghua, J. Xu, and Y. C. Soh. Modelling regular occupancy in commercial buildings using stochastic models. Energy & Buildings 103(2015):216-223.

[2.112] Shi, Jie, N. Yu, and W. Yao. Energy Efficient Building HVAC Control Algorithm with Real-time Occupancy Prediction. Energy Procedia 111(2017):267-276.

[2.113] Dedesko, Sandra, et al. Methods to assess human occupancy and occupant activity in hospital patient rooms. Building & Environment90.3(2015):136-145.

[2.114] Ansanay-Alex, Guillaume. Estimating Occupancy Using Indoor Carbon Dioxide Concentrations Only in an Office Building: a Method and Qualitative Assessment. Rehva World Congress energy Efficient, Smart and Healthy Buildings: Clima 2013.

[2.115] Ito, S, and H. Nishi. Estimation of the number of people under controlled ventilation using a CO 2, concentration sensor. IECON 2012 -, Conference on IEEE Industrial Electronics Society IEEE, 2012:4834-4839.

[2.116] Jin, Ming, et al. Sensing by Proxy: Occupancy Detection Based on Indoor CO2 Concentration. The, International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies 2015.

[2.117] Weekly, Kevin, et al. Modelling and Estimation of the Humans' Effect on the CO2 Dynamics Inside a Conference Room. IEEE Transactions on Control Systems Technology 23.5(2015):1770-1781.

[2.118] Dong, Bing, et al. An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network. Energy & Buildings 42.7(2010):1038-1046.

[2.119] Ekwevugbe, Tobore, et al. Real-time building occupancy sensing using neural-network based sensor network. IEEE International Conference on Digital Ecosystems and Technologies IEEE, 2013:114-119.

[2.120] Chen, Zhenghua, et al. Environmental Sensors-Based Occupancy Estimation in Buildings via IHMM-MLR. IEEE Transactions on Industrial Informatics 13.5(2017):2184-2193.

[2.121] badat, A., et al. Multi-room occupancy estimation through adaptive gray-box models. IEEE Conference on Decision and Control IEEE, 2015:3705-3711.

[2.122] Ebadat, Afrooz, et al. Estimation of building occupancy levels through environmental signals deconvolution. ACM Workshop on Embedded Systems for Energy-Efficient Buildings ACM, 2013:1-8.

[2.123] Szczurek, Andrzej , M. Maciejewska , and T. Pietrucha . Occupancy determination based on time series of CO 2, concentration, temperature and relative humidity. Energy & Buildings 147(2017):142-154.

[2.124] Ang, Irvan Bastian Arief, F. D. Salim, and M. Hamilton. Human occupancy recognition with multivariate ambient sensors. IEEE International Conference on Pervasive Computing and Communication Workshops IEEE, 2016:1-6.

[2.125] E. Hailemariam, R. Goldstein, R. Attar, A. Khan, Real-time occupancy detection using decision trees with multiple sensor types, in: Proceedings of the 2011Symposium on Simulation for Architecture and Urban Design, Society for Computer Simulation International, San Diego, CA, 2011, pp. 141–148.

[2.126] Candanedo, Luis M., and V. Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and CO 2, measurements using statistical learning models. Energy & Buildings 112(2016):28-39.

[2.127] Yang, Zheng, et al. A systematic approach to occupancy modelling in ambient sensor-rich buildings. Simulation 90.8(2014):960-977.

[2.128] Ariefang, Irvan B., F. D. Salim, and M. Hamilton. CD-HOC: Indoor Human Occupancy Counting using Carbon Dioxide Sensor Data. (2017).

[2.129] Jiang, Chaoyang, et al. Indoor occupancy estimation from carbon dioxide concentration. Energy & Buildings 131(2016):132-141.

[2.130] Zhu, Qingchang, et al. Occupancy estimation with environmental sensing via non-iterative LRF feature learning in time and frequency domains. Energy & Buildings 141(2017):125-133.

[2.131] Risuleo, R. S., Molinari, M., Bottegal, G., Hjalmarsson, H., & Johansson, K. H. (2015). A benchmark for data-based office modelling: challenges related to CO2 dynamics. IFAC-PapersOnLine, 48(28), 1256-1261.

[2.132] Zikos, Stylianos, et al. Conditional Random Fields - based approach for real-time building occupancy estimation with multi-sensory networks."Automation in Construction 68(2016):128-145.

[2.133] Naji, S., Keivani, A., Shamshirband, S., Alengaram, U.J., Jumaat, M.Z.,

Mansor, Z. and Lee, M., 2016. Estimating building energy consumption using extreme learning machine method. Energy, 97, pp.506-516.

[2.134] Li, C., Ding, Z., Zhao, D., Yi, J. and Zhang, G., 2017. Building energy consumption prediction: An extreme deep learning approach. Energies, 10(10), p.1525.

[2.135] Cui, C., Wu, T., Hu, M., Weir, J.D. and Li, X., 2016. Short-term building energy model recommendation system: a meta-learning approach. Applied Energy, 172, pp.251-263.

[2.136] Wong, S. L., K. K. W. Wan, and T. N. T. Lam. Artificial neural networks for energy analysis of office buildings with daylighting. Applied Energy 87.2(2010):551-557.

[2.137] Zuo, W., Huang, S., & Sohn, M. D. (2016). A BAYESIAN NETWORK MODEL FOR PREDICTING THE COOLING LOAD OF EDUCATIONAL FACILITIES. IBPSA-USA Journal, 6(1).

[2.138] Osman, Z. H, M. L. Awad, and T. K. Mahmoud. Neural network based approach for short-term load forecasting. Power Systems Conference and Exposition, 2009. PSCE '09. IEEE/PES IEEE, 2009:1-8.

[2.139] Mena, R., Rodríguez, F., Castilla, M. and Arahal, M.R., 2014. A prediction model based on neural networks for the energy consumption of a bioclimatic building. Energy and Buildings, 82, pp.142-155.

[2.140] Chae, Y.T., Horesh, R., Hwang, Y. and Lee, Y.M., 2016. Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. Energy and Buildings, 111, pp.184-194.

[2.141] Powell, K.M., Sriprasad, A., Cole, W.J. and Edgar, T.F., 2014. Heating, cooling, and electrical load forecasting for a large-scale district energy system. Energy, 74, pp.877-885.

[2.142] Leung, M.C., Norman, C.F., Lai, L.L. and Chow, T.T., 2012. The use of occupancy space electrical power demand in building cooling load prediction. Energy and Buildings, 55, pp.151-163.

[2.143] Kwok, Simon S. K., R. K. K. Yuen, and E. W. M. Lee. An intelligent approach to assessing the effect of building occupancy on building cooling

load prediction. Building & Environment 46.8(2011):1681-1690.

[2.144] An, N., Zhao, W., Wang, J., Shang, D. and Zhao, E., 2013. Using multi-output feedforward neural network with empirical mode decomposition based signal filtering for electricity demand forecasting. Energy, 49, pp.279-288.

[2.145] Paudel, S., Elmtiri, M., Kling, W.L., Le Corre, O. and Lacarrière, B., 2014. Pseudo dynamic transitional modeling of building heating energy demand using artificial neural network. Energy and Buildings, 70, pp.81-93.

[2.146] Deb, C., Eang, L.S., Yang, J. and Santamouris, M., 2016. Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks. Energy and Buildings, 121, pp.284-297.

[2.147] Feng, Xiaohang, Da Yan, Chuang Wang, and Hongsan Sun. A preliminary research on the derivation of typical occupant behavior based on large-scale questionnaire surveys. Energy and Buildings 117 (2016): 332-340.

[2.148] Virote, J. and Neves-Silva, R., 2012. Stochastic models for building energy prediction based on occupant behavior assessment. Energy and Buildings, 53, pp.183-193.

[2.149] Huang, H., Chen, L. and Hu, E., 2015. A neural network-based multi-zone modelling approach for predictive control system design in commercial buildings. Energy and buildings, 97, pp.86-97.

[2.150] Gruber, M., Trüschel, A. and Dalenbäck, J.O., 2014. Model-based controllers for indoor climate control in office buildings–complexity and performance evaluation. Energy and Buildings, 68, pp.213-222.

Reference for chapter 3

[3.1] Panapakidis, Ioannis P., Theofilos A. Papadopoulos, Georgios C. Christoforidis, and Grigoris K. Papagiannis. "Pattern recognition algorithms for electricity load curve analysis of buildings." *Energy and Buildings* 73 (2014): 137-145.

[3.2] Tsekouras, George J., Nikos D. Hatziargyriou, and Evangelos N. Dialynas. "Two-stage pattern recognition of load curves for classification of electricity

customers." *IEEE Transactions on Power Systems* 22, no. 3 (2007): 1120-1128.

[3.3] Chicco, Gianfranco, Roberto Napoli, and Federico Piglione. "Comparisons among clustering techniques for electricity customer classification." *IEEE Transactions on Power Systems* 21, no. 2 (2006): 933-940.

[3.4] López, José J., José A. Aguado, F. Martín, F. Munoz, A. Rodríguez, and José E. Ruiz. "Hopfield–K-Means clustering algorithm: A proposal for the segmentation of electricity customers." *Electric Power Systems Research* 81, no. 2 (2011): 716-724.

[3.5] Mutanen, Antti, Maija Ruska, Sami Repo, and Pertti Jarventausta. "Customer classification and load profiling method for distribution systems." *IEEE Transactions on Power Delivery* 26, no. 3 (2011): 1755-1763.

[3.6] Chicco, Gianfranco, and Irinel-Sorin Ilie. "Support vector clustering of electrical load pattern data." *IEEE Transactions on Power Systems* 24, no. 3 (2009): 1619-1628.

[3.7] Peng, Chen, Da Yan, Ruhong Wu, Chuang Wang, Xin Zhou, and Yi Jiang. "Quantitative description and simulation of human behavior in residential buildings." In *Building simulation*, vol. 5, no. 2, pp. 85-94. Tsinghua Press, 2012.

[3.8] Hartkopf, Volker H., Vivian E. Loftness, and Peter AD Mill. "The concept of total building performance and building diagnostics." In *Building performance: Function, preservation, and rehabilitation*. ASTM International, 1986.

[3.9] Jian, Yi-wen, Qing-rui Li, Zhen Bai, and Xiang-dong Kong. "Study on influences of usage behavior of residential air handling unit on energy consumption in summer." *Building Science* 27, no. 12 (2011): 16-20.

Reference for chapter 4

[4.1] Rory V. Jones , Alba Fuertes, Elisa Gregori, Alberto Giretti, Stochastic behavioural models of occupants' main bedroom window operation for UK residential buildings. Building and Environment. 118 (2017) 144-158

[4.2] Bongchan Jeong, Jae-Weon Jeong, J.S. Park, Occupant behavior regarding the manual control of windows in residential buildings. Energy and Buildings. 127 (2016) 206–216

[4.3] Rune Andersen, Valentina Fabi, Jorn Toftum , Stefano P. Corgnati , Bjarne W. Olesen, Window opening behaviour modelled from measurements in Danish dwellings. Building and Environment. 69 (2013) 101-113

[4.4] Davide Calì, Rune Korsholm Andersen, Dirk Müller, Bjarne W. Olesen, Analysis of occupants' behavior related to the use of windows in German households. Building and Environment. 103 (2016) 54-69

[4.5] Federica Naspi , Marco Arnesano , Lorenzo Zampetti , Francesca Stazi , Gian Marco Revel , Marco D'Orazi, Experimental study on occupants' interaction with windows and lights in Mediterranean offices during the non-heating season. Building and Environment. 127 (2018) 221–238

[4.6] Francesca Stazi, Federica Naspi, Marco D'Orazio, Modelling window status in school classrooms. Results from a case study in Italy. Building and Environment. 111 (2017) 24-32

[4.7] Yufan Zhang, Peter Barrett, Factors influencing the occupants' window opening behaviour in a naturally ventilated office building. Building and Environment. 50 (2012) 125-134

[4.8] Dutton, Spencer, and Li Shao. "Window opening behaviour in a naturally ventilated school." Proceedings of SimBuild 4, no. 1 (2010): 260-268.

[4.9] H.B. Rijal, P. Tuohy, M.A. Humphreys, J.F. Nicol , A. Samuel, J. Clarke, Using results from field surveys to predict the effect of open windows on thermal comfort and energy use in buildings, Energy and Buildings. 39 (2007) 823–836

[4.10] Herkel, Sebastian, Ulla Knapp, and Jens Pfafferott. Towards a model of user behaviour regarding the manual control of windows in office buildings. Building and environment. 43, no. 4 (2008): 588-600

[4.11] Valentina Fabi, Rune Korsholm Andersen, Stefano Corgnati, Verification of stochastic behavioural models of occupants' interactions with

windows in residential buildings. Building and Environment. 94 (2015) 371-383

[4.12] Shen Wei, Richard Buswell, Dennis Loveday, Factors affecting 'end-of-day' window position in a non-air-conditioned office building. Energy and Buildings. 62 (2013) 87–96

[4.13] Kim, Hakpyeong, Taehoon Hong, and Jimin 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

[4.14] Zhang, Yufan, and Peter Barrett. Factors influencing occupants' blind-control behaviour in a naturally ventilated office building. Building and Environment. 54 (2012): 137-147

[4.15] Haldi, Frédéric, and Darren Robinson. Adaptive actions on shading devices in response to local visual stimuli. Journal of Building Performance Simulation. 3, no. 2 (2010): 135-153

[4.16] Zhao, Jie, Bertrand Lasternas, Khee Poh Lam, Ray Yun, and Vivian Loftness. Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining. Energy and Buildings. 82 (2014): 341-355

[4.17] Geun Young Yun, Paul Tuohy, Koen Steemers, User behaviour of window-control in offices during summer and winter. Energy and Buildings. 41(5). pp. 489-499

[4.18] R. FRITSCH, A. KOHLER, M. NYGÅRD-FERGUSON, J.-L. SCARTEZZINI. A Stochastic Model of User Behaviour Regarding Ventilation. Building and Environment. 25 (1990) 173-181

[4.19] Geun Young Yun, Paul Tuohy, Koen Steemers, Thermal performance of a naturally ventilated building using a combined algorithm of probabilistic occupant behaviour and deterministic heat and mass balance models. Energy and Buildings. 41 (2009) 489–499

[4.20] Frédéric Haldi, Darren Robinson, Interactions with window openings by office occupants. Building and Environment. 44 (2009) 2378–2395

[4.21] Haldi, Frédéric, Davide Calì, Rune Korsholm Andersen, Mark Wesseling, and Dirk Müller. "Modelling diversity in building occupant behaviour: a novel statistical approach." Journal of Building Performance Simulation 10, no. 5-6 (2017): 527-544

[4.22] Davide Calì, Mark Thomas Wesseling, Dirk Müller, WinProGen: A Markov-Chain-based stochastic window status profile generator for the simulation of realistic energy performance in buildings. Building and Environment. 136 (2018) 240–258

[4.23] Yun, Geun Young, Paul Tuohy, and Koen Steemers. Thermal performance of a naturally ventilated building using a combined algorithm of probabilistic occupant behaviour and deterministic heat and mass balance models. Energy and buildings. 41, no. 5 (2009): 489-499

[4.24] Li, Nan, Juncheng Li, Ruijuan Fan, and Hongyuan Jia. Probability of occupant operation of windows during transition seasons in office buildings. Renewable Energy. 73 (2015): 84-91

[4.25] Richardson, Ian, Murray Thomson, and David Infield. A high-resolution domestic building occupancy model for energy demand simulations. Energy and buildings. 40, no. 8 (2008): 1560-1566.

[4.26] Chen, Zhenghua, Jinming Xu, and Yeng Chai Soh. Modeling regular occupancy in commercial buildings using stochastic models. Energy and Buildings. 103 (2015): 216-223.

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

[4.28] Widén, Joakim, Annica M. Nilsson, and Ewa Wäckelgård. A combined Markov-chain and bottom-up approach to modelling of domestic lighting demand. Energy and Buildings. 41, no. 10 (2009): 1001-1012.

[4.29] Wilke, Urs, Frédéric Haldi, Jean-Louis Scartezzini, and Darren Robinson. A bottom-up stochastic model to predict building occupants' time-dependent activities. Building and Environment. 60 (2013): 254-264.

[4.30] Günay, M. Erdem. Forecasting annual gross electricity demand by artificial neural networks using predicted values of socio-economic indicators and climatic conditions: Case of Turkey. Energy Policy. 90(2016):92-101.

[4.31] An, Ning, Weigang Zhao, Jianzhou Wang, Duo Shang, and Erdong Zhao. Using multi-output feedforward neural network with empirical mode decomposition based signal filtering for electricity demand forecasting. Energy. 49 (2013): 279-288.

[4.32] Du, Zhimin, Bo Fan, Xinqiao Jin, and Jinlei Chi. Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. Building and Environment. 73 (2014): 1-11.

[4.33] Li, Shun, and Jin Wen. A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform. Energy and Buildings. 68 (2014): 63-71.

[4.34] Zhao, Jie, Ray Yun, Bertrand Lasternas, Haopeng Wang, Khee Poh Lam, Azizan Aziz, and Vivian Loftness. Occupant behavior and schedule prediction based on office appliance energy consumption data mining. In CISBAT 2013 Conference-Clean Technology for Smart Cities and Buildings, pp. 549-554. 2013

[4.35] Song Pan, Yiye Han, Shen Wei, Yixuan Wei , Liang Xia, Lang Xie, Xiangrui Kong, Wei Yu, A model based on Gauss Distribution for predicting window behavior in building. Building and Environment 149 (2019) 210–219

[4.36] Romana Markovic, Eva Grintal, Daniel Wölki, Jérôme Frisch, Christoph van Treeck, Window Opening Model using Deep Learning Methods. Building and Environment. 145 (2018) 319–329

[4.37] Verena M. Barthelmes, Yeonsook Heo, Valentina Fabi, Stefano P. Corgnati, Exploration of the Bayesian Network framework for modelling window control behavior. Building and Environment 126 (2017) 318–330

[4.38] Simona D'Oca, Tianzhen Hong, A data-mining approach to discover patterns of window opening and closing behaviour in offices. Building and Environment 82 (2014) 726-739

[4.39] D'Oca, Simona, and Tianzhen Hong. Occupancy schedules learning process through a data mining framework. Energy and Buildings. 88 (2015): 395-408.

[4.40] Farhang Tahmasebi, Ardeshir Mahdavi, An inquiry into the reliability of window operation models in building performance simulation. Building and Environment. 105 (2016) 343-357

[4.41] Dayi Lai, Yue Qi, Junjie Liu, Xilei Dai, Lei Zhao, Shen Wei, Ventilation behavior in residential buildings with mechanical ventilation systems across different climate zones in China, Building and Environment. 143 (2018) 679–690

[4.42] SongPan,Yingzi Xiong, Yiye Han, Xingxing Zhang, Liang Xia, Shen Wei, Jinshun Wu, Mengjie Han, A study on influential factors of occupant window-opening behavior in an office building in China. Building and Environment. 133 (2018) 41–50

[4.43] Mingyao Yao, Bin Zhao, Factors affecting occupants' interactions with windows in residential buildings in Beijing, China. Procedia Engineering. 205 (2017) 3428–3434

[4.44] Mingyao Yao, Bin Zhao, Window opening behavior of occupants in residential buildings in Beijing. Building and Environment. 124 (2017) 441-449

[4.45] Zhenni Shi, Hua Qian, Xiaohong Zheng, Zhengfei Lv, Yuguo Li, Li Liu, Peter V. Nielsen, Seasonal variation of window opening behaviors in two naturally ventilated hospital wards. Building and Environment. 130 (2018) 85–93

[4.46] Scott Menard, Applied logistic regression analysis. Second edition. Shanghai, Truth & Wisdom Press, 2012

[4.47] Bo Zhang, Hao Shang, Applied Stochastic Processes. Second edition. Beijing, China Renmin University Press, 2009

Reference for chapter 5

[5.1] D. Liu, X. Guan, Y. Du, Q. Zhao, Measuring indoor occupancy in intelligent buildings using the fusion of vision sensors, Meas. Sci. Technol. 24 (7) (2013)074023.

[5.2] Y. Benezeth, H. Laurent, B. Emile, C. Rosenberger, Towards a sensor for detecting human presence and characterizing activity, Energy Build. 43 (2) (2011) 305–314.

[5.3] Wang, Fulin, et al. Predictive Control of Indoor Environment Using Occupant Number Detected by Video Data and CO 2, Concentration. Energy & Buildings 145(2017).

[5.4] P. Liu, S.K. Nguang, A. Partridge, Occupancy inference using pyroelectric infrared sensors through hidden markov models, IEEE Sens. J. 16 (4) (2016)1062–1068.

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

[5.6] Gunay, H Burak, et al. DETECTING OCCUPANTS' PRESENCE IN OFFICE SPACES: A CASE STUDY. Esim 2016.

[5.7] Jin, Ming, R. Jia, and C. Spanos. Virtual Occupancy Sensing: Using Smart Meters to Indicate Your Presence. IEEE Transactions on Mobile Computing PP.99(2017):1-1.

[5.8] S. Depatla, A. Muralidharan, Y. Mostofi, Occupancy estimation using only WiFi power measurements, IEEE J. Sel. Areas Commun. 33 (7) (2015) 1381–1393.

[5.9] N. Li, G. Calis, B. Becerik-Gerber, Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations, Autom. Constr. 24(2012) 89–99.

[5.10] Zhao, Yang, et al. Virtual occupancy sensors for real-time occupancy information in buildings. Building & Environment 93.2(2015):9-20.

[5.11] K. Padmanabh, V.A. Malikarjuna, S. Sen, S.P. Katru, A. Kumar, S.K. Vuppala, S.Paul, et al., iSense: a wireless sensor network based conference

room management system, in: Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-efficiency in Buildings, ACM, 2009, pp.37–42.

[5.12] Shih, Oliver, and A. Rowe. Occupancy estimation using ultrasonic chirps. ACM/IEEE Sixth International Conference on Cyber-Physical Systems ACM, 2015:149-158.

[5.13] Amayri, Manar, et al. Estimating occupancy in heterogeneous sensor environment. Energy & Buildings 129(2016):46-58.

[5.14] Mckenna, Eoghan, M. Krawczynski, and M. Thomson. Four-state domestic building occupancy model for energy demand simulations. Energy & Buildings 96.8(2015):30-39

[5.15] Richardson, Ian, M. Thomson, and D. Infield. A high-resolution domestic building occupancy model for energy demand simulations. Energy & Buildings 40.8(2008):1560-1566.

[5.16] Chen, Zhenghua, J. Xu, and Y. C. Soh. Modelling regular occupancy in commercial buildings using stochastic models. Energy & Buildings 103(2015):216-223

[5.17] Andersen, Philip Delff, et al. Dynamic modelling of presence of occupants using inhomogeneous Markov chains. Energy & Buildings 69.69(2014):213-223.

[5.18] Erickson, Varick L., M. Á. Carreira-Perpiñán, and A. E. Cerpa. OBSERVE: Occupancy-based system for efficient reduction of HVAC energy. International Conference on Information Processing in Sensor Networks IEEE, 2011:258-269.

[5.19] Dong, Bing, and K. P. Lam. A real-time model predictive control for building heating and cooling systems based on the occupancy behaviour pattern detection and local weather forecasting. Building Simulation 7.1(2014):89-106.

[5.20] Mahdavi, Ardeshir, and F. Tahmasebi. Predicting people's presence in buildings: An empirically based model performance analysis. Energy & Buildings 86(2015):349-355.

[5.21] Chen, Zhenghua, J. Xu, and Y. C. Soh. Modelling regular occupancy in commercial buildings using stochastic models. Energy & Buildings 103(2015):216-223.

[5.22] Shi, Jie, N. Yu, and W. Yao. Energy Efficient Building HVAC Control Algorithm with Real-time Occupancy Prediction. Energy Procedia 111(2017):267-276.

[5.23] Dedesko, Sandra, et al. Methods to assess human occupancy and occupant activity in hospital patient rooms. Building & Environment90.3(2015):136-145.

[5.24] Ansanay-Alex, Guillaume. Estimating Occupancy Using Indoor Carbon Dioxide Concentrations Only in an Office Building: a Method and Qualitative Assessment. Rehva World Congress energy Efficient, Smart and Healthy Buildings: Clima 2013.

[5.25] Ito, S, and H. Nishi. Estimation of the number of people under controlled ventilation using a CO 2, concentration sensor. IECON 2012 -, Conference on IEEE Industrial Electronics Society IEEE, 2012:4834-4839.

[5.26] Jin, Ming, et al. Sensing by Proxy: Occupancy Detection Based on Indoor CO2 Concentration. The, International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies 2015.

[5.27] Weekly, Kevin, et al. Modelling and Estimation of the Humans' Effect on the CO2 Dynamics Inside a Conference Room. IEEE Transactions on Control Systems Technology 23.5(2015):1770-1781.

[5.28] Dong, Bing, et al. An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network. Energy & Buildings 42.7(2010):1038-1046.

[5.29] Ekwevugbe, Tobore, et al. Real-time building occupancy sensing using neural-network based sensor network. IEEE International Conference on Digital Ecosystems and Technologies IEEE, 2013:114-119.

[5.30] Chen, Zhenghua, et al. Environmental Sensors-Based Occupancy Estimation in Buildings via IHMM-MLR. IEEE Transactions on Industrial Informatics 13.5(2017):2184-2193.

[5.31] badat, A., et al. Multi-room occupancy estimation through adaptive gray-box models. IEEE Conference on Decision and Control IEEE, 2015:3705-3711.

[5.32] Ebadat, Afrooz, et al. Estimation of building occupancy levels through environmental signals deconvolution. ACM Workshop on Embedded Systems for Energy-Efficient Buildings ACM, 2013:1-8.

[5.33] Szczurek, Andrzej , M. Maciejewska , and T. Pietrucha . Occupancy determination based on time series of CO 2, concentration, temperature and relative humidity. Energy & Buildings 147(2017):142-154.

[5.34] Ang, Irvan Bastian Arief, F. D. Salim, and M. Hamilton. Human occupancy recognition with multivariate ambient sensors. IEEE International Conference on Pervasive Computing and Communication Workshops IEEE, 2016:1-6.

[5.35] E. Hailemariam, R. Goldstein, R. Attar, A. Khan, Real-time occupancy detection using decision trees with multiple sensor types, in: Proceedings of the 2011Symposium on Simulation for Architecture and Urban Design, Society for Computer Simulation International, San Diego, CA, 2011, pp. 141–148.

[5.36] Candanedo, Luis M., and V. Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and CO 2, measurements using statistical learning models. Energy & Buildings 112(2016):28-39.

[5.37] Yang, Zheng, et al. A systematic approach to occupancy modelling in ambient sensor-rich buildings. Simulation 90.8(2014):960-977.

[5.38] Ariefang, Irvan B., F. D. Salim, and M. Hamilton. CD-HOC: Indoor Human Occupancy Counting using Carbon Dioxide Sensor Data. (2017).

[5.39] Jiang, Chaoyang, et al. Indoor occupancy estimation from carbon dioxide concentration. Energy & Buildings 131(2016):132-141.

[5.40] Zhu, Qingchang, et al. Occupancy estimation with environmental sensing via non-iterative LRF feature learning in time and frequency domains. Energy & Buildings 141(2017):125-133.

[5.41] C.S. Cho, S. Lee, Effective five directional partial derivatives-based images moothing and a parallel structure design, IEEE Trans. Image Process. 25 (4)(2016) 1617–1625.

[5.42] Abed-Meraim, K, Wanzhi Qiu, and Hua, Y. Blind system identification. Proceedings of the IEEE 85.8(1997):1310-1322

[5.43] Cullen, Alison C., H. Christopher Frey, and Christopher H. Frey. Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs. Springer Science & Business Media, 1999.

[5.44] Box, George EP, and George C. Tiao. Bayesian inference in statistical analysis. Vol. 40. John Wiley & Sons, 2011

[5.45] Aglan, Heshmat A. Predictive model for CO2 generation and decay in building envelopes. Journal of Applied Physics 93.2(2003):1287-1290.

[5.46] Wang Changke, Wang Yuesi, Liu Guangren, Characteristics of atmospheric CO2 variations and some affecting factors in urban area of Beijing. Environmental Science 24.4(2003):13-20

[5.47] Liu, Xiao Man, X. L. Cheng, and H. U. Fei. Gradient characteristics of CO_2 concentration and flux in Beijing urban area part I: Concentration and virtual temperature. Chinese Journal of Geophysics 58.5(2015):1502-1512.

[5.48] Dougan, D. S. CO2-Based Demand Control Ventilation: Do Risks Outweigh Potential Rewards? Ashrae Journal 46.10(2004):47-55.

[5.49] A., S. H. R. A. E. ASHRAE STANDARD Ventilation for Acceptable Indoor Air Quality. (2007).

[5.50] Association, The Mathematical. The Mathematical Gazette. Mathematical Gazette 1.6(1906):112.

[5.51] Rasmussen, Carl Edward. Gaussian Processes in Machine Learning. Machine Learning Summer School Conference 2004:63-71.

[5.52] Pillonetto, G., and A. Chiuso. Tuning complexity in kernel-based linear system identification: The robustness of the marginal likelihood estimator. Control

Reference for chapter 6

[6.1] Naji, S., Keivani, A., Shamshirband, S., Alengaram, U.J., Jumaat, M.Z., Mansor, Z. and Lee, M., 2016. Estimating building energy consumption using extreme learning machine method. Energy, 97, pp.506-516.

[6.2] Li, C., Ding, Z., Zhao, D., Yi, J. and Zhang, G., 2017. Building energy consumption prediction: An extreme deep learning approach. Energies, 10(10), p.1525.

[6.3] Cui, C., Wu, T., Hu, M., Weir, J.D. and Li, X., 2016. Short-term building energy model recommendation system: a meta-learning approach. Applied Energy, 172, pp.251-263.

[6.4] Wong, S. L., K. K. W. Wan, and T. N. T. Lam. Artificial neural networks for energy analysis of office buildings with daylighting. Applied Energy 87.2(2010):551-557.

[6.5] Zuo, W., Huang, S., & Sohn, M. D. (2016). A BAYESIAN NETWORK MODEL FOR PREDICTING THE COOLING LOAD OF EDUCATIONAL FACILITIES. IBPSA-USA Journal, 6(1).

[6.6] Osman, Z. H, M. L. Awad, and T. K. Mahmoud. Neural network based approach for short-term load forecasting. Power Systems Conference and Exposition, 2009. PSCE '09. IEEE/PES IEEE, 2009:1-8.

[6.7] Mena, R., Rodríguez, F., Castilla, M. and Arahal, M.R., 2014. A prediction model based on neural networks for the energy consumption of a bioclimatic building. Energy and Buildings, 82, pp.142-155.

[6.8] Chae, Y.T., Horesh, R., Hwang, Y. and Lee, Y.M., 2016. Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. Energy and Buildings, 111, pp.184-194.

[6.9] Powell, K.M., Sriprasad, A., Cole, W.J. and Edgar, T.F., 2014. Heating, cooling, and electrical load forecasting for a large-scale district energy system. Energy, 74, pp.877-885.

[6.10] Leung, M.C., Norman, C.F., Lai, L.L. and Chow, T.T., 2012. The use of occupancy space electrical power demand in building cooling load

prediction. Energy and Buildings, 55, pp.151-163.

[6.11] Kwok, Simon S. K., R. K. K. Yuen, and E. W. M. Lee. An intelligent approach to assessing the effect of building occupancy on building cooling load prediction. Building & Environment 46.8(2011):1681-1690.

[6.12] An, N., Zhao, W., Wang, J., Shang, D. and Zhao, E., 2013. Using multi-output feedforward neural network with empirical mode decomposition based signal filtering for electricity demand forecasting. Energy, 49, pp.279-288.

[6.13] Paudel, S., Elmtiri, M., Kling, W.L., Le Corre, O. and Lacarrière, B., 2014. Pseudo dynamic transitional modeling of building heating energy demand using artificial neural network. Energy and Buildings, 70, pp.81-93.

[6.14] Deb, C., Eang, L.S., Yang, J. and Santamouris, M., 2016. Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks. Energy and Buildings, 121, pp.284-297.

[6.15] Feng, Xiaohang, Da Yan, Chuang Wang, and Hongsan Sun. A preliminary research on the derivation of typical occupant behavior based on large-scale questionnaire surveys. Energy and Buildings 117 (2016): 332-340.

[6.16] Virote, J. and Neves-Silva, R., 2012. Stochastic models for building energy prediction based on occupant behavior assessment. Energy and Buildings, 53, pp.183-193.

[6.17] Huang, H., Chen, L. and Hu, E., 2015. A neural network-based multi-zone modelling approach for predictive control system design in commercial buildings. Energy and buildings, 97, pp.86-97.

[6.18] Gruber, M., Trüschel, A. and Dalenbäck, J.O., 2014. Model-based controllers for indoor climate control in office buildings–complexity and performance evaluation. Energy and Buildings, 68, pp.213-222.

[6.19] Huang, Guang-Bin, Dian Hui Wang, and Yuan Lan. "Extreme learning machines: a survey." International journal of machine learning and cybernetics 2.2 (2011): 107-122.

[6.20] Bartlett, P.L., 1997. For valid generalization the size of the weights is more important than the size of the network. In Advances in neural information processing systems (pp. 134-140).

[6.21] Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. Neuro computing 2006;70:489e501.

[6.22] Li, M.B., Huang, G.B., Saratchandran, P. and Sundararajan, N., 2005. Fully complex extreme learning machine. Neurocomputing, 68, pp.306-314.

[6.23] Jackson, Edward J. A User's Guide To Principal Components. Journal of the Operational Research Society 43.6(2005):641-641.

[6.24] Qin, Jianying, and S. Wang. A fault detection and diagnosis strategy of VAV air-conditioning systems for improved energy and control performances. Energy & Buildings 37.10(2005):1035-1048.

[6.25] Platon, Radu, V. R. Dehkordi, and J. Martel. Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis. Energy & Buildings92.1(2015):10-18.