Sustainable Cities and Society 61 (2020) 102247

Contents lists available at ScienceDirect

Sustainable Cities and Society

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

A novel reinforcement learning method for improving occupant comfort via window opening and closing

Mengjie Han[a], Ross May[a], Xingxing Zhang[a,*], Xinru Wang[a], Song Pan[b], Da Yan[c], Yuan Jin[c]

[a] School of Technology and Business Studies, Dalarna University, Falun 79188, Sweden
[b] Beijing Key Laboratory of Green Built Environment and Energy Efficient Technology, Beijing University of Technology, Beijing 100124, China
[c] Building Energy Conservation Research Center, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Keywords:
Markov decision processes
Reinforcement learning
Window control
Indoor comfort
Occupant

ABSTRACT

An occupant's window opening and closing behaviour can significantly influence the level of comfort in the indoor environment. Such behaviour is, however, complex to predict and control conventionally. This paper, therefore, proposes a novel reinforcement learning (RL) method for the advanced control of window opening and closing. The RL control aims at optimising the time point for window opening/closing through observing and learning from the environment. The theory of model-free RL control is developed with the objective of improving occupant comfort, which is applied to historical field measurement data taken from an office building in Beijing. Preliminary testing of RL control is conducted by evaluating the control method's actions. The results show that the RL control strategy improves thermal and indoor air quality by more than 90% when compared with the actual historically observed occupant data. This methodology establishes a prototype for optimally controlling window opening and closing behaviour. It can be further extended by including more environmental parameters and more objectives such as energy consumption. The model-free characteristic of RL avoids the disadvantage of implementing inaccurate or complex models for the environment, thereby enabling a great potential in the application of intelligent control for buildings.

1. Introduction

Indoor comfort, for example, thermal comfort and air quality, have become major concerns for building designers and operators (Roulet et al., 2006). The maintenance of these factors is important for improving the level of an occupant's comfort, health, morale, working efficiency, and productivity (Shaikh et al., 2013; Singh, 1996). A survey shows that more than 80% of occupants are satisfied with their thermal comfort in only 11% of the buildings. For indoor air quality (IAQ), only 26% of the buildings have 80% or more satisfied occupants (Huizenga et al., 2006). In Denmark, 54% of a group of surveyed inhabitants claimed that they have at least one problem related to indoor comfort. A majority of those respondents did not try to search for information on how to solve the problem (Frontczak et al., 2012). Improving the comfort level of occupants is therefore urgent for a sustainable society, and the realisation of it seems to be a joint task between the occupants, the building designers, and the building management system (BMS).

Thermal comfort is used to manifest the thermal state of a human within a given environment (Enescu, 2017). According to ASHRAE-55, the ambient parameter, temperature, is considered as the most relevant one for thermal comfort (*ASHRAE Standard 55*, 2017). For IAQ, air quality index (AQI) - which measures the levels of pollutants in the air – is often used. Kyrkilis et al. (2007) reported a combination of CO, SO_2, NO_2, O_3, and PM10 as the five components of AQI. Cheng et al. (2007) included particulate matter with less than 2.5 μm diameter (PM2.5) since it can trigger cardiovascular disease-related mortality and non-fatal events. Control strategies for maintaining thermal comfort and IAQ at a desired level have been mostly implemented on heating, ventilation, and air conditioning (HVAC) systems since these have a direct influence on both the indoor environment and energy consumption. In a building with natural ventilation, however, indoor comfort depends largely on the control of window opening and closing. Compared to HVAC systems, the control of windows changes the indoor environment through naturally exchanging the air with the outdoor environment and therefore does not demand additional energy. Nevertheless, arbitrary and customary window control by an occupant does not guarantee the improvement of the indoor environment. For example, keeping an open window when the outdoor air quality becomes poor may increase the discomfort level. The occupant can easily fail to sense this slow deterioration of their surroundings. Thus,

* Corresponding author.
E-mail address: xza@du.se (X. Zhang).

https://doi.org/10.1016/j.scs.2020.102247
Received 17 March 2019; Received in revised form 30 April 2020; Accepted 1 May 2020
Available online 25 May 2020

intelligent automation for window control has substantial potential to increase the level of comfort of an occupant.

Machine learning, a subfield of artificial intelligence (AI), has been used in buildings research for many years, and has demonstrated its potential to enhance building performance (Hong et al., 2020). Indeed, a number of previous studies have applied logistic regression as a prediction method for the control of window opening and closing behaviour. In recent times, a prevalent machine learning technique known as model-free reinforcement learning (RL) has made breakthroughs in intelligent controls and decision making (Mnih et al., 2013, 2015; Silver et al., 2016, 2017). An RL agent learns how to optimally act given the environment it interacts with. An early work in this area for buildings is Mozer's Neural Network House (Mozer, 1998). In this groundbreaking piece of work, a residential building's environmental parameters and observations of the occupants' actions are used by an RL agent for optimally controlling the building system. Other applications of RL to building energy systems have followed since this work, for example, in systems such as HVAC (Chen et al., 2019; Fazenda et al., 2014), lighting (Park et al., 2019), heat pump (Nagy et al., 2018; Ruelens et al., 2015), water heaters (Ruelens et al., 2014), and battery and photovoltaic systems (Shi et al., 2017). There are many more such examples of RL being applied in the building literature, however, as the scope of this paper specifically concerns the application of RL used for window control, our focus therefore lies in considering current RL approaches to this building system.

Although there are existing studies using RL in window operating, for example, in the control of HVAC systems, particularly for ventilation purposes (Chen et al., 2018; Dalamagkidis et al., 2007), as well as in a holistic setting among the four subsystems, HVAC, lighting, blind, and window systems (Ding et al., 2019), there exists no research regarding the application of RL in window opening and closing from the aspect of occupant behaviour. This paper, therefore, aims to fill this research gap.

In this study, we train two RL agents to learn when to open or close a window in an office building in Beijing, so as to maximise the comfort level of the indoor environment as measured by a combination of thermal comfort and IAQ, where, respectively, ambient temperature and AQI have been used as proxies. A recurrent neural network (RNN) is used to predict the indoor temperature as a result of an action taken on the window system. This enables the comparison of the agent's window opening/closing behaviour, with that of the actual observed historical occupant window/opening behaviour, under the same environmental conditions as experienced by the occupant at that time. As shown in Fig. 1, the contributions of this paper are as follows. We propose a model-free reinforcement learning method for controlling windows in office buildings. We optimise the opening and closing of a window system with regard to maximising a combination of thermal comfort and air quality – using air temperature and AQI, respectively, as proxies - where a data-driven approach is used for simulating the environmental changes. A comparison is made between the window opening/closing policies of the RL agents and the occupant under identical conditions using a dataset containing the occupant's window behaviour and environmental measurements at that time. A theoretical basis is established for the future live deployment of the developed control method in the physical office, and thus for further incorporation of occupant feedback into its control logic.

The rest of the paper is organised as follows. Section 2 examines drivers for window opening and closing and respective control methods used for controlling the indoor environment via window systems. Section 3 then briefly introduces the RL method and algorithms used in this paper. In Section 4, we summarise the data and implementation details, and in Section 5 the results are discussed. Lastly, we conclude the paper in the final section.

2. Behaviours of window opening and closing

Occupant behaviour is a complex process and there are many drivers for an occupant to interact with building control. Apart from the contextual, psychological, physiological, and social factors of a building occupant, physical environmental factors have been considered as the most direct driver (Fabi et al., 2012). Since the behaviour of window opening and closing has a significant impact on both the indoor environment and energy consumption, understanding the underlying drivers and modeling methods will contribute to implementing efficient control techniques.

2.1. Drivers of window opening and closing

Investigation of window opening and closing behaviour can be conducted in many ways. One approach is to use surveys in the form of questionnaires. This makes it flexible for the investigator to raise desired questions (Jeong et al., 2016; Nunes de Freitas & Guedes, 2015). These surveys reveal that most of the time, such behaviour aims at improving physical feelings and that it varies for different seasons. For example, opening a window in winter is more explained by air quality, whereas closing a window in summer may be due to outside noise (Nunes de Freitas & Guedes, 2015). And a drop in indoor temperature can explain more the act of window closing than the outdoor temperature in winter (Jeong et al., 2016).

Another approach is to use statistical modeling to explain occupant behaviour (Haldi & Robinson, 2009; Pan et al., 2019). The probabilistic paradigm further allows us to monitor the distribution of behaviour in simulation studies. In an earlier work (Fritsch et al., 1990), it was stated that the probability of finding a window position depends on the preceding position of windows. A popular method for studying such behaviour is logistic regression analysis. It studies the binary dependent variables by fitting linearly independent variables and has been comprehensively used to model window opening and closing (Andersen et al., 2013; D'Oca & Hong, 2014; Fabi et al., 2013; Li et al., 2015; Pan et al., 2018; Rijal et al., 2008; Rijal et al., 2018; Yun & Steemers, 2008). It can both identify influential factors and predict window opening

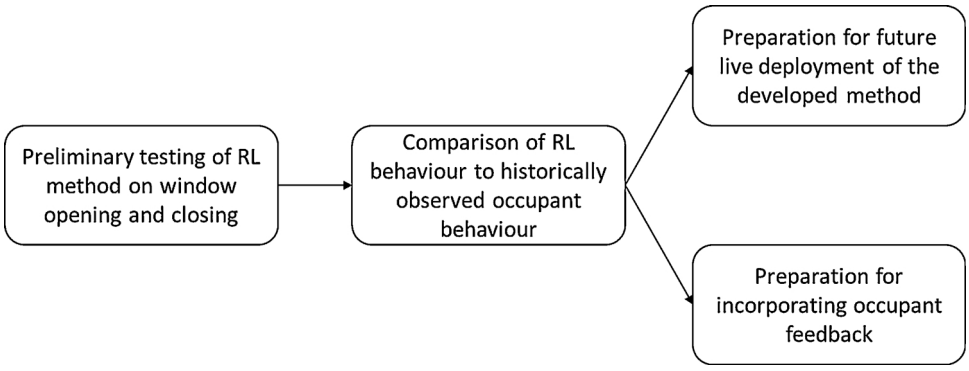


Fig. 1. Flowchart of the contributions.

probabilities. For example, in residential buildings in Denmark, the probability can be explained by indoor CO_2 concentration and outdoor temperature as the common patterns (Andersen et al., 2013). Large variations, however, are found between the patterns in naturally ventilated buildings and mechanically ventilated buildings (Fabi et al., 2013). For office buildings in a natural ventilation season, the outdoor temperature was identified as the primary driver among other environmental factors. Another study has found that the trigger point for occupants' window opening is to get better thermal comfort and air quality (Li et al., 2015). Simulation results indicate that during transition seasons, the probability of window opening in office buildings follows a normal distribution and increases linearly with the outdoor temperature growth. Further studies have revealed that factors such as indoor temperature, occupant arrival and leaving time, presence, window positions, solar radiation, wind speed, seasons and time of the day also contribute to explain window operation for office buildings (D'Oca & Hong, 2014; Pan et al., 2018). In a study of two general hospital wards conducted in Nanjing, China (Shi et al., 2018), the effects of air quality (i.e. indoor CO2 concentration and outdoor PM2.5 concentration) and climatic parameters (i.e. indoor/outdoor temperature, relative humidity, outdoor wind speed, wind direction, and rainfall) on window opening/closing behaviour were analysed. Indoor air temperature and relative humidity were found to be dominant factors for window opening behaviour. Outdoor temperature was found to be associated with the probability of window opening negatively during the cooling season, but positively during the transition and heating seasons. Indoor relative humidity positively affects the probability of window opening during the transition season while a negative impact appears during the cooling and heating seasons.

Logistic regression models have been successfully developed to predict window opening/closing and have been verified to be promisingly adaptable for an accurate result. Similarly, a Probit analysis also models the probability of window operations (Yun & Steemers, 2008). Yun & Steemers monitored data and gave evidence that there is a statistically significant relationship between window-opening behaviour patterns and indoor stimulus in summer. For example, a window in an office that featured a night cooling strategy was always open upon departure whenever the room temperature was above 23.6 °C.

2.2. Occupant comfort and intelligent controllers

Having understood the drivers, we can single out the control targets and focus on specific control or automation methods.

Thermal comfort and IAQ have been considered as the most pertinent objectives (Andersen et al., 2013; Jin et al., 2015; Li et al., 2015; Stazi et al., 2017; Tanner & Henze, 2014). The measurement of these largely depends on the operation of windows via changing the airflow rate. Thus, an advanced automatic window control method leads to a consequential change in the indoor environment and hence optimises the occupant's overall comfort level in terms of these aforementioned objectives.

Rijal et al. demonstrated that the adaptive Humphreys algorithm could assist in achieving more comfortable, lower energy buildings while avoiding overheating (Rijal et al., 2008) This algorithm can also be used to adjust CO_2 concentration to the desired level while keeping the operative temperature at a constant (Stazi et al., 2017).

Four algorithms were compared for reducing energy consumption and improving comfort with regard to smart windows in commercial buildings (Dussault et al., 2016). The ruled-based controller and the quasi-optimal controller obtained by the genetic algorithm showed the best real-time control. It was also pointed out that genetic algorithms and model predictive control (MPC) are powerful tools that can easily accept more complex objective functions or scenarios.

When the occupant is involved in the control system, a stochastic process for occupant behaviour can be modeled following a known distribution (Tanner & Henze, 2014). Tanner & Henze demonstrated this by implementing a stochastic MPC. Compared to deterministic optimal control, stochastic optimisation is more conservative but offers better performance. In a survey work by Han et al., the benefits of model-free control in such settings are illustrated from the methodology point of view (Han et al., 2019).

2.3. Building environment

Both descriptive and analytical methods for finding drivers for window opening and closing aim to efficiently operate windows so that the occupants are satisfied. By defining an objective function, we can design a controller for solving this sequential decision-making problem. A key issue for a computational agent to develop a control method is the ability to sense the change in the state of the environment. A common strategy for achieving this is to use building simulation programs (Li et al., 2015; Wang & Greenberg, 2015). This approach is fast and flexible for obtaining data, but cannot guarantee accuracy as occupant presence presents a significant influence on building performance. In this paper, we propose a data-driven approach to predict the change in the indoor environment due to the operation of windows. Distinct from building simulation programs, a data-driven learning process can gradually improve accuracy. As the model-free RL control method for improving indoor comfort has not been studied in this way, a prototype of the prediction-based implementation and its achievement is demonstrated.

3. RL and algorithms

Reinforcement learning (RL) essentially looks for best policies in the process of decision-making over time. The RL agent optimises its actions through interacting with and learning from the environment. It learns how to map situations to actions so as to maximise a numerical delayed reward signal. It doesn't need to have a "teacher" telling it how to take an action, rather, it makes decisions via implementing a trial-and-error search and recognising the delayed reward from the environment that the agent interacts with (Sutton & Barto, 2018).

The environment gives stochastic feedback to the agent. In most cases, the environment cannot be modeled accurately and thus model-free RL techniques such as Q-learning and SARSA are employed as learning algorithms. Richard Bellman came up with the concept of Markov decision processes (MDPs) or finite MDPs, a fundamental theory of RL, to formulate the underlying framework for solving such problems (Bellman, 1957a).

3.1. Markov decision processes

In a dynamic sequential decision-making process, the state $S_t \in \mathscr{S}$ refers to a specific condition of the environment at discrete time steps $t = 0,1, \ldots$. By realising and responding to the environment, the agent chooses a deterministic or stochastic action $A_t \in \mathscr{A}$ that tries to maximise future returns and receives an immediate reward $R_{t+1} \in \mathscr{R}$ as the agent transfers to the new state S_{t+1}. The reward is usually represented by a quantitative measurement. Fig. 2 (Sutton & Barto, 2018) shows how a sequence of state, action, and reward are generated to form an MDP.

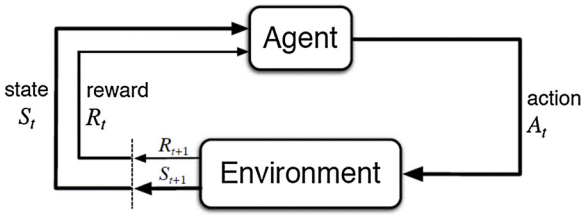


Fig. 2. The interaction between agent and environment in an MDP.

The Markov property tells us that the future is independent of the past and depends only on the present. In Fig. 2, S_t and R_t are the outcomes after taking an action and are considered as random variables. Thus, the joint probability density function for S_t and R_t is defined by,

$$p(s^{'}, r|s, a) = \mathbb{P}[S_t = s^{'}, R_t = r|S_{t-1} = s, A_{t-1} = a], \tag{1}$$

where $s, s' \in \mathscr{S}$, $r \in \mathscr{R}$ and $a \in \mathscr{A}$. It can be seen from Eq. (1) that the distribution of state and reward at time t depends only on the state and action one step before. Eq. (1) implies the basic rule (or dynamics) of how the MDP works and one can easily determine the marginal transition probabilities $p(s^{'}s, a)$,

$$p(s^{'}|s, a) = \mathbb{P}[S_t = s^{'}|S_{t-1} = s, A_{t-1} = a] = \sum_{r\in\mathscr{R}} p(s^{'}, r|s, a). \tag{2}$$

Eq. (3) gives the expected reward by using the marginal distribution of R_t:

$$r(s, a) = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = \sum_{r\in\mathscr{R}} r\sum_{s'\in\mathscr{S}} p(s^{'}, r|s, a). \tag{3}$$

Both Eq. (2) and Eq. (3) are used for solving the optimal value functions presented in Section 3.2.

3.2. Policies and value functions

A *policy* π is a distribution over actions given states. It fully defines the behaviour of an agent by telling the agent how to act when it is in different states. The policy itself is either deterministic or stochastic (Sutton & Barto, 2018) and the probability of taking an action, a, in state s is:

$$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]. \tag{4}$$

The policy can be considered as a function of actions. The selection of actions can be achieved by either creating a look-up table (see Section 3.3) or building an approximation model. The overall goal of RL is to find the optimal policy given a state.

An optimal policy tries to maximise the expected future return from time t: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...$, where $0 \le \gamma \le 1$ is the discount parameter. The *state-value function,* $v_\pi(s)$, and the *action-value function,* $q_\pi(s, a)$, are two useful measures in RL that can be estimated from the data. We define $v_\pi(s)$, of an MDP, under policy π, as the expectation of the return starting from state s:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s], \ \textit{for all } s \in \mathscr{S}. \tag{5}$$

In practical applications, $v_\pi(s)$ is more applicable for model-based problems, that is, problems where a model of the dynamics is known a priori. Whereas the action-value function, $q_\pi(s, a)$, is more useful in the model-free context when the dynamics is not known. Episodic simulations are often used to estimate $q_\pi(s, a)$, where,

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}_\pi[G_t|S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a], \ \textit{for all } s \in \mathscr{S} \textit{ and } a \in \mathscr{A}. \end{aligned} \tag{6}$$

The task of finding the optimal policy, π_*, is achieved by evaluating either the optimal state-value function

$$v_*(s) = \max_\pi v_\pi(s), \tag{7}$$

or the optimal action-value function

$$q_*(s, a) = \max_\pi q_\pi(s, a). \tag{8}$$

The way to optimise Eqs. (7) or (8) is to make use of the recursive relationships between two states in sequential order, known as the *Bellman optimality equation* for $q_*(s, a)$ (Bellman, 1957b), which is obtained by summating the following,

$$q_*(s, a) = r(s, a) + \gamma\sum_{s'\in\mathscr{S}} p(s^{'}|s, a)\max_{a^{'}} q_*(s', a'). \tag{9}$$

3.3. Q-learning and SARSA

A straightforward method to find $q_*(s, a)$ given a policy π is to iteratively update the values of $q_\pi(s, a)$ by maximising the sum of the discounted future returns and the immediate reward, known as the learning target. In a general iteration process, the new estimate of a target is updated by summing the old estimate with an error induced by the incremental observation, namely:

$$NewEstimate \longleftarrow OldEstimate + StepSize(Target - OldEstimate).$$

Different strategies regarding Q-value updates and action selections yield off-policy and on-policy algorithms. Among those theoretically convergent-guaranteed algorithms, the off-policy Q-learning and on-policy SARSA algorithms learn policies efficiently when the state-action pairs are discrete and the size is moderate. Thus, in this paper, we consider Q-learning and SARS*A* as our testing algorithms and examine their performances in adaptive window controls.

3.3.1. Q-learning

Q-learning (Watkins, 1989) is a value-based tabular method. A look-up table is built to store all state-action pairs and the corresponding action-values. When the agent is in a specific state and an action is selected, i.e. (s, a), the update for this state-action pair evaluates the transited state-action pair, i.e. (s', a'). The subsequent action, a', is taken such that $q(s', a')$ is maximised. As seen in Algorithm 1, the update to a new action-value is achieved by adding a so-called TD-error, $\alpha[R + \gamma\max_{a^{'}} Q(S^{'}, a') - Q(S, A)]$, to the old action-values. The value function $Q(S, A)$ asymptotically converges to $q_*(s, a)$. An ε -greedy exploration indicates that the agent chooses an action that has maximal estimated action-value with probability $1 - \varepsilon$, but with probability ε the agent selects an action at random with equal probability.

Algorithm 1. Tabular Q-learning

Input: discount parameter γ; step size parameter α; $\{s, a\} \in \{\mathscr{S}, \mathscr{A}\}$; $\varepsilon > 0$ initialised $Q(s, a)$.

1: **Loop** for each episode
2: Initialise S
3: **Loop** for each time step
4: Choose A from S by e.g., ε-greedy policy
5: Take action A and observe R and $S^{'}$
6: $Q(S, A) \longleftarrow Q(S, A) + \alpha[R + \gamma\max_{a'} Q(S^{'}, a') - Q(S, A)]$
7: $S \longleftarrow S^{'}$
8: **Until** S is terminal

Output: Q-table

3.3.2. SARSA

Compared to Q-learning, SARSA is more conservative and sensitive to errors. When the agent of SARSA updates its Q-table, it observes the successor state and takes an action according to, for example, an ε -greedy policy (or another exploration method) derived from Q, whereas Q-learning always looks for the maximum Q-value by evaluating those possible successor actions. Moreover, Q-learning re-selects the successor actions after updating the Q-table (due to exploration), which makes the policy of the learning agent distinct from the policy for updating the Q-table and thus behaves off-policy. The incremental update in the SARSA algorithm uses all of the elements in $(S, A, R, S^{'}, A')$ to obtain the action-value,

$$Q(S, A) \longleftarrow Q(S, A) + \alpha[R + \gamma Q(S', A') - Q(S, A)],$$

where A' is derived from an ε -greedy or another exploration method. Algorithm 2 gives the implementation details of SARSA.

Algorithm 2. Tabular SARSA

Input: discount parameter γ; step size parameter α; $\{s,\ a\} \in \{\mathscr{S},\ \mathscr{A}\}$; $\varepsilon > 0$; initialised $Q(s, a)$.

1: **Loop** for each episode
2: Initialise S
3: Choose A from S by e.g. ε-greedy policy
4: **Loop** for each time step
5: Take action A and observe R and S'
6: Choose A' from S' by e.g. ε-greedy policy
7: $Q(S, A) \longleftarrow Q(S, A) + \alpha[R + \gamma Q(S', A') - Q(S, A)]$
8: $S \longleftarrow S'$ and $A \longleftarrow A'$
9: **Until** S is terminal

Output: Q-table

4. Data and methods

4.1. Data

In this study, data from an office building constructed of reinforced concrete and brick at a university in Beijing are used. The construction material of the building is composed of 370 mm common brick with thermal conductivity 0.6 W/(m·K) and 200 mm polystyrene foam with thermal conductivity 0.033 W/(m·K). All the offices, shown in Fig. 3(a), are located on the second floor in the building, whereas the floor spaces are used for laboratories (Pan et al., 2018). There is one door and one push-pull south pointing window for the experimental room. As shown in Fig. 3(b), the window size is 1.5 m × 1.6 m and can be half-open. One out of ten offices is selected for our experiment. As illustrated in Fig. 4, the experimental room (No.8 according to the serial number of the building) is located at the southeast corner with a size of 3.29 m × 3.11 m.

The data collection took place between March 16, 2015 and May 15, 2015. This period in Beijing is the transition season with moderate outdoor temperature and so natural ventilation is highly preferred. To facilitate the comparison, the same occupant following the working routine in the university was in the room during the data collection period. An earlier work (Pan et al., 2018) gives a detailed description of the variables and the settings of the sensors. The devices that were used for collecting the data are highlighted in Fig. 5. An indoor air temperature sensor TR (v1.2) was placed inside the room to avoid direct sunshine and local heating sources. A portable outdoor meteorological weather station was put over the roof where outdoor temperature, solar radiation, AQI, and wind speed and directions are measured. Moreover, an intelligent human body inductor P100 was used to detect the wavelength of the human body. For the days when the room was occupied for at least 30 min, the daily occupied time ranges from 50 min to 11 h and has a mean value of 5.5 h. To monitor the window, a displacement tester was applied to detect and record the position of the window. This was achieved by the magnetic induction of two dry spring pipes positioned on the window. The action of opening was recorded when the window was opened more than 3 cm and the opening period exceeded 3 s.

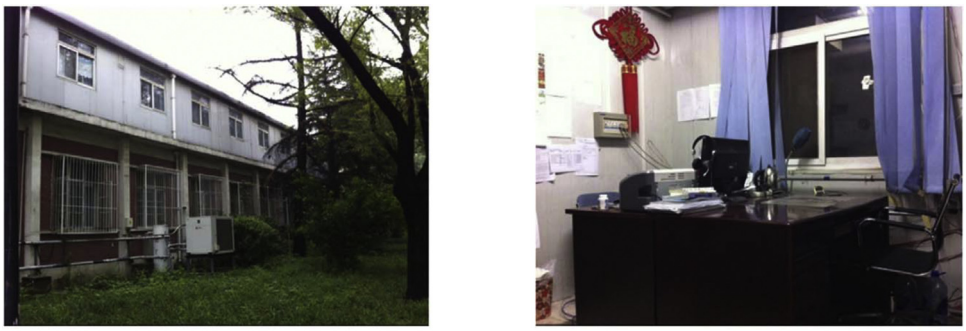
Indoor temperature (T^{in}), outdoor temperature (T^{out}), solar radiation (SR), wind speed (WS), wind direction (WD), and outdoor air quality index (AQI) were selected as our environment variables. The position of the window (P_{window}, open/closed) and the occupancy information of the room (occupied/unoccupied) were also tracked. Given that the comfort factors T^{in} and AQI do not change drastically, each data record was collected at a time resolution of 10 min, which are accessible from a data logger. Indoor temperature is the main component for thermal comfort. Due to the limitation of the devices and time, we did not measure other thermal comfort factors, for example, metabolic rate, clothing insulation, radiant temperature, air speed and humidity. Given this, we simplify the factors of thermal comfort by only considering the most representative factor, temperature. For full factors, the adjustment of Eq. (10) is straightforward according to the method from ASHRAE standard 55 (*ASHRAE Standard 55*, 2017). Since simplified thermal comfort also generates an interval or zone to indicate discomfort that affects reward in Eq. (10), the simplification has no impact on our training method compared to comfort with full factors. Hence, thermal comfort mentioned in this study refers to the concept with simplified factor.

4.2. Methods

The RL agent optimises its behaviour by exploring a number of different trajectories as time goes. The observed state at a time point has a strong correlation to the following states, because the change of the environment is time-dependent. Instead of simulating the trajectories, we conduct a data-driven approach to mimic the impact on the environment when an action is made. Among the variables in $\mathscr{S}$, the indoor temperature is susceptible to the actions taken and so the prediction of it helps to evaluate the future return.

The recurrent neural network (RNN) (Mandic & Chambers, 2001) makes use of sequential data to make predictions. An RNN has a memory that stores previous information about what has been calculated. This is achieved by including the hidden layer that is obtained from one step earlier as input to the current hidden layer. In Fig. 6, the information flows of the input I and the output O have been stored and passed into the hidden layer a. An additional weight matrix W_a connects the hidden layers between two time points by computing the function $a_t = f(W_1 I_t + W_a a_{t-1})$ given a nonlinear activation function f. The unrolled RNNs share the same weight parameters W_1, W_2 and W_a across the entire prediction steps.

Training an RNN is similar to training an ordinary neural network. By using the backpropagation Through Time (BPTT) method (Werbos, 1990), the gradient for an RNN at each output depends on both the current input and previous output. The gradient at time t needs to sum up all previous $t - 1$ gradients. In practice, the long-term dependency



(a) (b)

Fig. 3. Office building (a) and room (b).

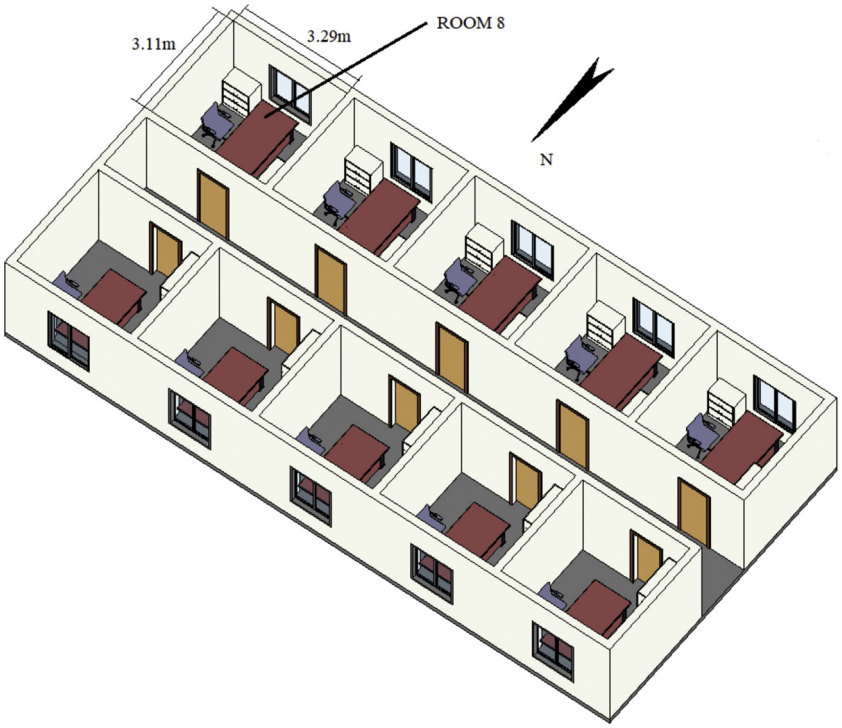


Fig. 4. The selected office room.

makes BPTT unable to work due to the problem of vanishing or exploding gradients (Pascanu et al., 2013). A special case of an RNN, known as Long Short-Term Memory (LSTM), can perfectly avoid this problem by adding gates to open and close access to the previous information (Hochreiter & Schmidhuber, 1997). We analyse the results of our RNN predictive model in Section 5.

Predicting the change in the environment enables the agent to start learning. To train an RL agent that acts optimally given a certain state, the state S has to be representative of the environment, thereby giving the agent the potential to figure out the transition probabilities and estimated future reward. Hence, both the indoor and outdoor temperatures as well as the AQI have been identified as direct environmental factors that have impacts on the position change of the window. Wind speed and solar radiation have also been included as factors since they affect both the airflow rate and ambient temperature. Since the current position of the window forms the baseline for the agent, this too

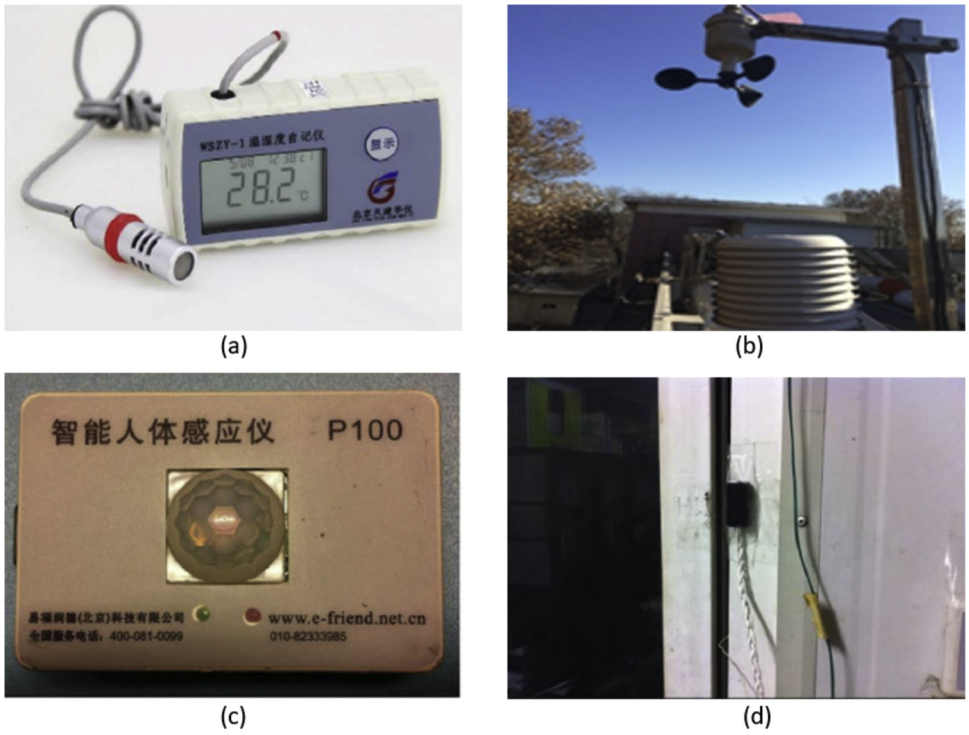


Fig. 5. Data collection devices: (a) indoor temperature sensor; (b) portable outdoor meteorological weather station; (c) human body inductor; (d) window tester.

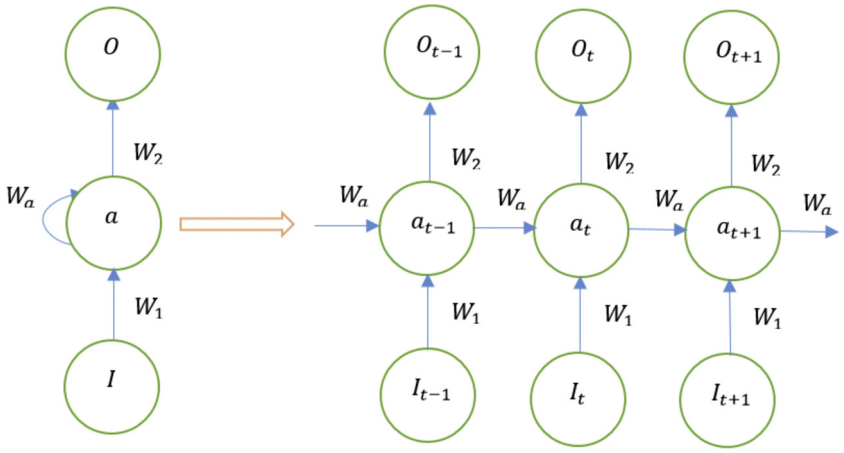


Fig. 6. Structure of RNN.

has been included as part of the state components. Furthermore, Gauss distribution models and logistic regressions have also shown that the variables T^{in}, T^{out}, SR, WS, AQI significantly influence window operations (Pan et al., 2018, 2019). We therefore formulate a single state as a sextuple: S= $\langle T^{in}, T^{out}, SR, WS, AQI, P_{window}\rangle$. Dynamic variables such as airflow rate or variables that have indirect effects to comfort parameters such as wind direction are excluded from S. As some of the components of the state are continuous and imperative for making policies, we have thus discretised these so that the tabular RL algorithms can work.

Given the observed state, the action set, $\mathscr{A} = \{Switch, Inaction\}$, consists of only two elements, since we do not measure the degree of opening. *Switch* refers to either closing or opening the window depending on the current state. *Inaction* means keeping the position of the window unchanged under either scenario. The rationality to formulate the actions in this way is so that we are able to keep an eye on the current position of the window as well as track the actual position change. Given an open window, for example, it is more natural to say "keep it open" than to say "open the window".

The reward R reflects the comfort and is composed of both thermal comfort (as measured by the indoor temperature) and air quality (as measured by the proxy variable, AQI). We first define the thermal discomfort, τ, at time t by evaluating the squared difference between T^{in} and given thresholds,

$$\tau_t = \begin{cases} [\min(|T_t^{in} - T_{UB}|,\ |T_t^{in} - T_{LB}|)]^2, & T_t^{in} \notin [T_{LB},\ T_{UB}] \\ 0, & otherwise \end{cases}, \tag{10}$$

where T_{UB} and T_{LB} are, respectively, the upper and lower thresholds of the comfort temperature. Higher outdoor AQI can also bring discomfort to the occupant. A survey shows that people have an incentive to close windows for better indoor air quality when high outdoor AQI is detected (Pan et al., 2018). Hence, an additional component of discomfort, σ, is considered when the window is open and the AQI is higher than a given threshold, where,

$$\sigma_t = \begin{cases} AQI_t - AQI_{LB}, & AQI_t \in (AQI_{LB}, \infty) \\ 0, & otherwise \end{cases}. \tag{11}$$

A combination of the normalised thermal and air quality discomfort components yields the following reward for discomfort,

$$R_t = \omega_1 \frac{\tau_t - \min\tau_t}{\max\tau_t - \min\tau_t} + \omega_2 P_{window} \frac{\sigma_t - \min\sigma_t}{\max\sigma_t - \min\sigma_t}, \tag{12}$$

where the indicator variable, P_{window}, takes zero for a closed window and one for an open window, and the weight parameters, $0 \leq \omega_i \leq 1$, allocate the importance between the components. An inverse transformation, $R_t^{'} = 1/(R_t + \xi)$, where $\xi > 0$ is some small real number, allows us to solve the maximisation problem defined in Eq. (9). It should be noted that unlike the discretised components making up a state, T_t^{in} and AQI_t in Eqs. (11) and (12) are, respectively, the RNN predictions and the numerical values from the observations.

5. Results

The results for both indoor temperature modeling given by a trained RNN, as well as the control performance of the trained RL agents are discussed in this section. All experiments are conducted in python 3.6.5 using TensorFlow's (v1.12.0) high-level API, Keras (v2.2.4). These were implemented on a Single Intel(R) 64-Bit Core(TM) i5-7300U 2.70 GHz CPU with 16GB RAM.

5.1. RNN predictions

An RNN-LSTM with a single hidden layer of 50 LSTM units and a dense output layer for predicting inside temperature was trained for the experimental room. Since opening and closing the window has a direct effect on the airflow rate and ventilation speed, the rate of change of indoor temperature may increase after an immediate switch of the window position. This effect of a change in window position will gradually wear off until the next switch occurs. Therefore, we incorporate the following lagged - by one time-step – as input variables: T^{in}, T^{out}, SR, WS, P_{window}, WD, time of day, time since a switch to the window was made, the presence of the occupant, and outdoor humidity.

We use the 70%–30% rule to divide the data into training and validation sets under the period between March 20 and May 7. Since the sequential order matters in RNNs, we strictly follow the time series observations and do not shuffle the data. We further select observations in a continuous 3-day period (May 12 to May 15) as the testing set. The training stops when the average losses are not significantly reduced. As can be seen in Fig. 7, the number of epochs needed for getting a stable loss is about 40. In our experiments we do not shuffle the data and so the sequential feature may produce higher loss for unusual observations in the early stage of training. Nevertheless, this unusual high loss diminishes as the number of epochs increases. Moreover, the predicted T^{in} s of the 3-day period are compared against the actual values in the validation set in Fig. 8(a). For almost all time points, the predicted values are close to the actual values.

Once the RNN-LSTM is trained, we further test its accuracy on a test set distinct from both the training or validation datasets. The comparisons between the actual vs the predicted values are given in Fig. 8(b), where it can be seen that no significant deviations are found, and hence the RNN-LSTM predictive model thus trained generalises well to new inputs. The root mean squared error (RMSE) for the experimental room is 0.2 °C that is too trivial for the occupant's sensory-receptors to sense.

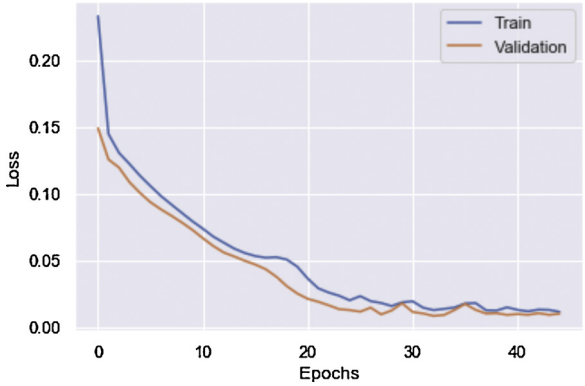


Fig. 7. Losses for RNN training.

Thus, the trained RNN-LSTM is verified as a suitable predictive model for simulating the environment.

5.2. Performance of RL agents

Due to a lack of computational resources, the learning outcomes are illustrated for a single day (April 8) as a prototype. The occupant spent eight hours in the room on this day, which is a typical office routine in China. This same procedure can be applied to any other day, but we do not see any reason why our agents would behave differently.

Discretisation to the continuous states for our experimental room and date are shown in Table 1. We restrict $T^{in} \in [20°C, 27°C]$ and $T^{out} \in [6°C, 15°C]$ to be within the bounds of the actual extreme values. Considering the lengths between the bounds in the intervals, a step-size of 1°C is set for both T^{in} and T^{out}. In running the tabular algorithms, both the actual and predicted states are numerically rounded to the nearest endpoints. The distribution of SR is skewed and so we consider ordinal indicators for representing uneven intervals. The AQI and P_{window} are binary variables where we only distinguish if AQI is greater than AQI_{LB} or not. We take $AQI_{LB} = 150$ as vulnerable groups of people start to have aggravation of symptoms of heart and respiratory diseases when $AQI_{LB} > 150$, furthermore, outdoor activities are not recommended when values of AQI are greater than the given threshold (Jassim & Coskuner, 2017; Jin et al., 2015; Pu et al., 2017). Further, in Eq. (10) we take $T_{LB} = 21°C$ and $T_{UB} = 27°C$ (*ASHRAE Standard 55*, 2017; Chen et al., 2018; Zhang et al., 2011). There is no obvious best temperature within this interval. Since we try to establish a general framework of agent training, we consider the most representative thresholds for the majority of occupants. Although individual comfort preference can vary, the consequence of the occupant's behaviour is emphasized in this study.

Table 1
Discretization of states.

Variables	Minimum	Maximum	Interval
T^{in}	20°C	27°C	1°C
T^{out}	6°C	15°C	1°C
WS	$0m/s$	$2.5m/s$	$0.5m/s$

Table 2
Penalty actions.

	$AQI_t > 150$	$AQI_t \leq 150$
$T_t^{in} < 21°C$	open	open
$21° C \leq T_t^{in} \leq 27°C$	open	–
$T_t^{in} > 27°C$	close, when $\frac{\tau_t - \min\tau_t}{\max\tau_t - \min\tau_t} > \frac{\sigma_t - \min\sigma_t}{\max\sigma_t - \min\sigma_t}$	close
	open, when $\frac{\tau_t - \min\tau_t}{\max\tau_t - \min\tau_t} < \frac{\sigma_t - \min\sigma_t}{\max\sigma_t - \min\sigma_t}$	

We take 144 time steps as one episode for both Q-learning and SARSA. To evaluate the learning performances of our agents, we not only monitor the reward function but we also examine the accumulative number of penalty actions (defined in Table 2) for each epoch. When the agent is in a specific state, we consider six different scenarios made by the values of AQI and T_t^{in}. For example, the agents should have learnt to close the window when $AQI_t > 150$ and $T_t^{in} < 21°C$; opening under this scenario would result in a penalty. Clearly, the reward function and the accumulative number of penalties are inversely related – an increase in the former will result in a decrease in the latter. By the end of the 20th epoch in Fig. 9, both Q-learning and SARSA are able to improve their reward functions and reduce their penalties. As stated in theory, Q-learning has higher variance than SARSA due to following a different policy to its behaviour policy. As we can see, by the end of training, SARSA slightly outperforms Q-learning.

We further evaluate the performances of the agents against the actual occupant. Specifically, we compare the average reward and penalty per time step for the whole day and three periods when the occupant

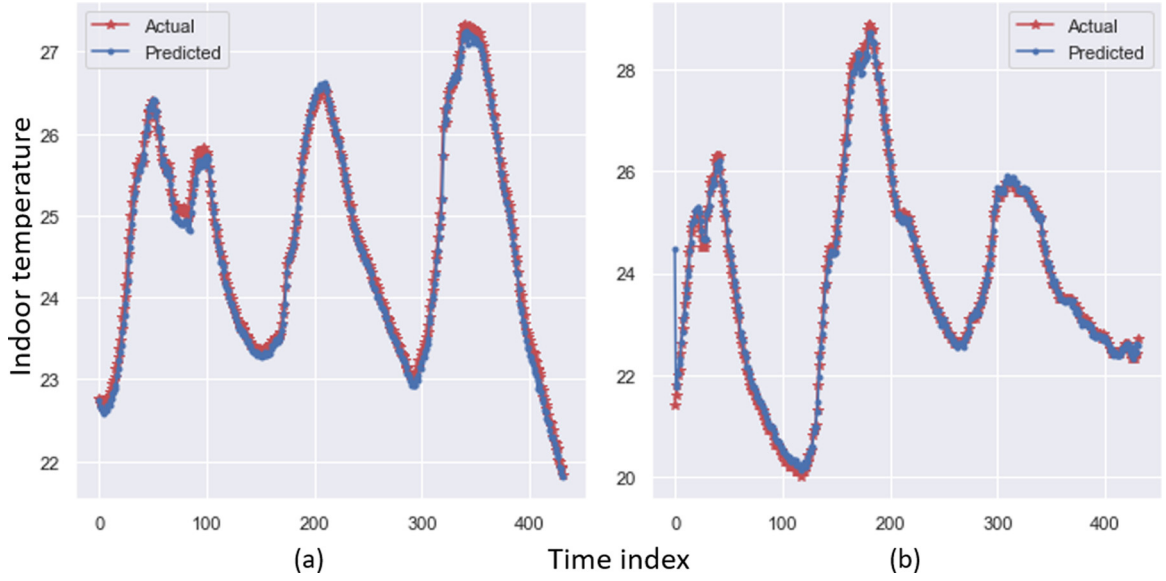


Fig. 8. 3-day period comparisons: (a) predicted validation sets and actual values; (b) predicted test sets and actual values.

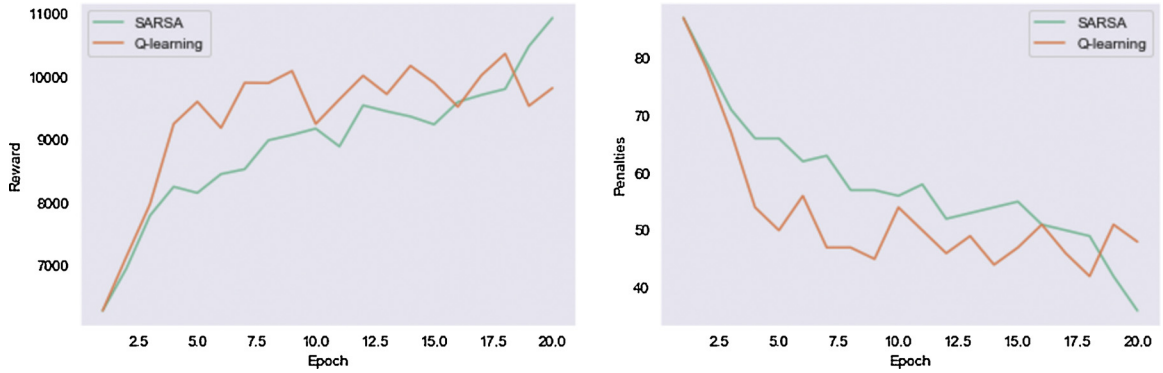


Fig. 9. Performances of Q-learning and SARSA.

Table 3
Average reward and penalty for the agents and occupant per time step.

		24-h	8:25 – 12:45	13:25 – 15:15	15:45 – 17:35
Q-learning	Avg. reward	70.12	7.60	55.26	47.20
	Avg. penalty	0.31	0.96	0.27	0.45
SARSA	Avg. reward	76.07	7.60	64.54	73.88
	Avg. penalty	0.26	0.96	0.27	0.18
Occupant	Avg. reward	23.57	7.82	28.89	3.10
	Avg. penalty	0.85	0.96	0.55	0.90

was in the room, namely, 8:25 – 12:45, 13:25 – 15:15, and 15:45 – 17:35. As shown in Table 3, the trained agents give on average more than 70 in reward for a 24-h period, whereas they reduce to 7.60 in the morning and increase in the afternoon. For the three occupied periods, the agents give lower average rewards than the 24-h period. This is because the agents have to compromise the gain when $AQI_t > 150$ and $T_t^{in} > 27°C$.

For the morning period, negligible differences in reward can be seen between the agents and the occupant, indicating that the agents can perform at least as good as the occupant. This tiny difference in reward is due to the prediction of indoor temperatures. If we scrutinize the actual actions given by both the agents and the occupant in Fig. 10, we see that the actions in the morning coincide with each other. In the afternoon, however, the longer time steps with an open window for the occupant make the reward higher for the agents. An explanation for the occupant's irrational behaviour is the inertial thinking from the morning, failing to sense the gradually increased AQI. The occupant may have been concentrating on his work and so easily forgot to close the window. The RL agent, however, is always able to learn from the environment and keep a level that is close to the 24-h average.

6. Conclusions

The control of windows in naturally ventilated buildings have a large impact on occupant comfort. Among the comfort factors, better thermal comfort and IAQ are of most concern for occupants. In typical Chinese office buildings, occupants may not behave optimally due to the complex climate and weather, and therefore intelligent control methods aiming at improving thermal comfort and IAQ become indispensable for smart and sustainable buildings. Previous intelligent control methods applied to adaptive window control have been based on models and their performances are therefore heavily dependent on the accuracy of these models. Furthermore, these models need to be corrected and re-identified as a consequence of a change in the dynamics of the building caused by, for example, retrofits. Thus, as an alternative solution which addresses such challenges, we have developed an automatic control prototype based on reinforcement learning for improving occupant comfort and tested it in a data-driven simulated environment.

An RNN-LSTM predictive model was used for predicting the indoor temperature given environmental variables and was verified by a test set. The high accuracy of the predictive model enabled us to simulate the actions of an agent in a flexible setting. Two tabular algorithms, Q-learning and SARSA, were used to train two RL agents whose learned behaviours were evaluated against the occupant's historically observed behaviour. The agents achieved much better policies than that of the historically observed occupant's policy measured in terms of both accumulative reward and penalties. An RL agent aims to maximise cumulative future return instead of the immediate reward at a single time point. Even though the performance of our trained agents failed to surpass the average level of the complete learning period for some specific sub-periods when the room was occupied, the agents still behaved close to the average level. This therefore means that tabular algorithms can inherently reduce the variance.

The prototype established in this paper leads to a large number of novel and valuable topics that are recommended for future works. We are still at the early stage of understanding the behaviours of window opening and closing. Human behaviour is indispensable for controlling the level of comfort in the indoor environment, and with regard to occupant-centric RL, we believe that occupant feedback will not only continuously correct the reward function in the process of learning, but will also increase the actual learning experience. Human effects for different occupants should be individually treated and they are highly related to psychological, physiological and social factors of the occupants. To this end, algorithms built for multi-agent cooperative systems (where agents have a joint action-value function in which the exploration of new states becomes complex as the number of agents increases) are valuable to explore and adapt accordingly in order to make them feasible in practice. The comfort level of an occupant is made up of the four factors, thermal comfort, indoor air quality, lighting, and noise, and therefore holistic approaches to measuring the comfort level of occupants should be explored and intergrated in the design of an intelligent agent. While discretising the state space allows for the application of tabular RL methods, in so doing errors may arise. Hence, solutions for training an agent with a continuous state space are therefore promising and thus approximation techniques need to be developed. With advanced computational power, as well as advances in algorithm design that address the problem of sample efficiency in RL (Botvinick et al., 2019), high performance is expected.

Declaration of Competing Interest

All authors have read and understood Elsevier's guidelines for defining authors. The funding bodies have been acknowledged. There are no conflicts of interest to disclose.

Acknowledgment

We are thankful for the support from the "13th Five-Year" National

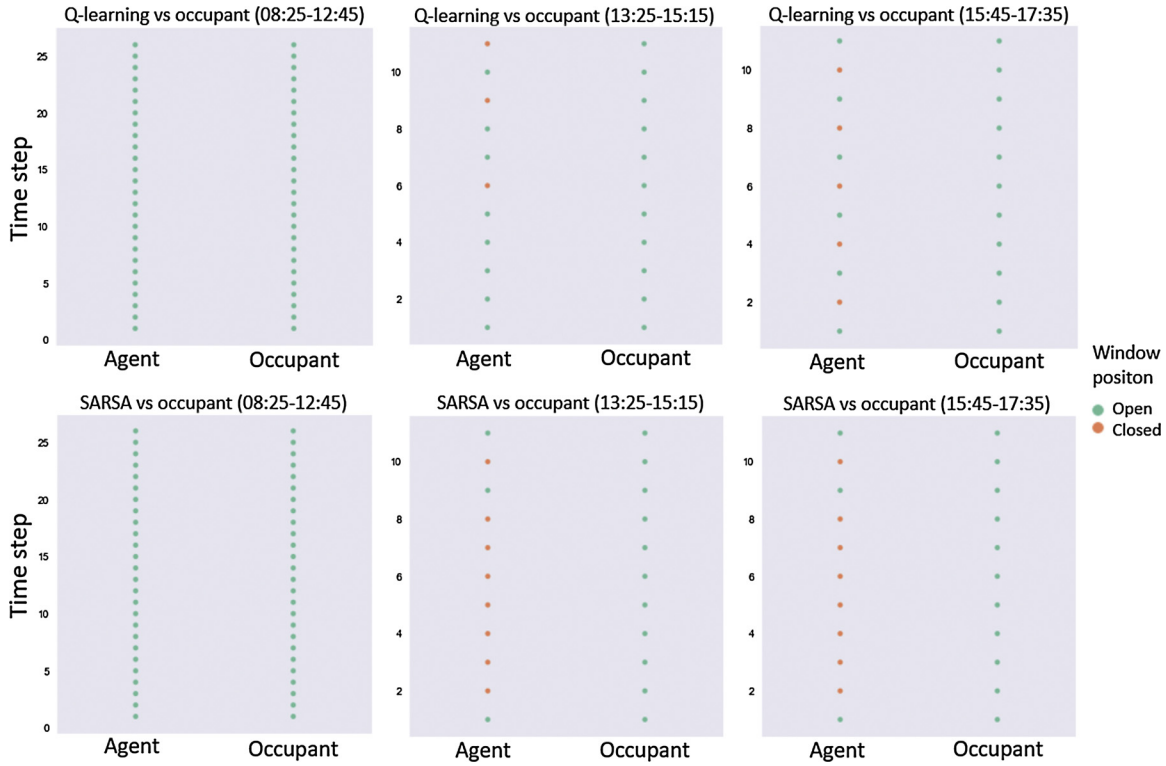


Fig. 10. Comparison of actions for different time periods.

Science and Technology Major Project of China (Grant No. 2017YFC0702202) and the IMMA project of research network, Dalarna University, Sweden.

References

Andersen, R., Fabi, V., Toftum, J., Corgnati, S. P., & Olesen, B. W. (2013). Window opening behaviour modelled from measurements in Danish dwellings. *Building and Environment, 69*, 101–113. https://doi.org/10.1016/j.buildenv.2013.07.005.

ASHRAE Standard 55–Thermal environmental conditions for human occupancy. ASHRAE Inc.

Bellman, R. (1957a). A markovian decision process. *Indiana University Mathematics Journal, 6*(4), 679–684. https://doi.org/10.1512/iumj.1957.6.56038.

Bellman, R. (1957b). *Dynamic programming.* Princeton Univ. Pr.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences, 23*(5), 408–422. https://doi.org/10.1016/j.tics.2019.02.006.

Chen, B., Cai, Z., & Berges, M. (2019). *Gnu-RL: A precocial reinforcement learning solution for building HVAC control using a differentiable MPC policy.* 316–325. https://doi.org/10.1145/3360322.3360849.

Chen, Y., Norford, L. K., Samuelson, H. W., & Malkawi, A. (2018). Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. *Energy and Buildings, 169*, 195–205. https://doi.org/10.1016/j.enbuild.2018.03.051.

Cheng, W.-L., Chen, Y.-S., Zhang, J., Lyons, T. J., Pai, J.-L., & Chang, S.-H. (2007). Comparison of the revised air quality index with the PSI and AQI indices. *The Science of the Total Environment, 382*(2–3), 191–198. https://doi.org/10.1016/j.scitotenv.2007.04.036.

D'Oca, S., & Hong, T. (2014). A data-mining approach to discover patterns of window opening and closing behavior in offices. *Building and Environment, 82*, 726–739. https://doi.org/10.1016/j.buildenv.2014.10.021.

Dalamagkidis, K., Kolokotsa, D., Kalaitzakis, K., & Stavrakakis, G. S. (2007). Reinforcement learning for energy conservation and comfort in buildings. *Building and Environment, 42*(7), 2686–2698. https://doi.org/10.1016/j.buildenv.2006.07.010.

Ding, X., Du, W., & Cerpa, A. (2019). *OCTOPUS: Deep reinforcement learning for holistic smart building control.* 326–335. https://doi.org/10.1145/3360322.3360857.

Dussault, J.-M., Sourbron, M., & Gosselin, L. (2016). Reduced energy consumption and enhanced comfort with smart windows: Comparison between quasi-optimal, predictive and rule-based control strategies. *Energy and Buildings, 127*, 680–691. https://doi.org/10.1016/j.enbuild.2016.06.024.

Enescu, D. (2017). A review of thermal comfort models and indicators for indoor environments. *Renewable and Sustainable Energy Reviews, 79*, 1353–1379. https://doi.org/10.1016/j.rser.2017.05.175.

Fabi, V., Andersen, R. V., Corgnati, S., & Olesen, B. W. (2012). Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models. *Building and Environment, 58*, 188–198. https://doi.org/10.1016/j.buildenv.2012.07.009.

Fabi, V., Andersen, R. V., Corgnati, S. P., & Olesen, B. W. (2013). A methodology for modelling energy-related human behaviour: Application to window opening behaviour in residential buildings. *Building Simulation, 6*(4), 415–427. https://doi.org/10.1007/s12273-013-0119-6.

Fazenda, P., Veeramachaneni, K., Lima, P., & O'Reilly, U.-M. (2014). Using reinforcement learning to optimize occupant comfort and energy usage in HVAC systems. *Journal of Ambient Intelligence and Smart Environments, 6*(6), 675–690. https://doi.org/10.3233/AIS-140288.

Fritsch, R., Kohler, A., Nygård-Ferguson, M., & Scartezzini, J.-L. (1990). A stochastic model of user behaviour regarding ventilation. *Building and Environment, 25*(2), 173–181. https://doi.org/10.1016/0360-1323(90)90030-U.

Frontczak, M., Andersen, R. V., & Wargocki, P. (2012). Questionnaire survey on factors influencing comfort with indoor environmental quality in Danish housing. *Building and Environment, 50*, 56–64. https://doi.org/10.1016/j.buildenv.2011.10.012.

Haldi, F., & Robinson, D. (2009). Interactions with window openings by office occupants. *Building and Environment, 44*(12), 2378–2395. https://doi.org/10.1016/j.buildenv.2009.03.025.

Han, M., May, R., Zhang, X., Wang, X., Pan, S., Yan, D., Jin, Y., & Xu, L. (2019). A review of reinforcement learning methodologies for controlling occupant comfort in buildings. *Sustainable Cities and Society, 51*, Article 101748. https://doi.org/10.1016/j.scs.2019.101748.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hong, T., Wang, Z., Luo, X., & Zhang, W. (2020). State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings, 212*(109831), 1–15.

Huizenga, C., Abbaszadeh, S., Zagreus, L., & Arens, E. A. (2006). Air quality and thermal comfort in office buildings: Results of a large indoor environmental quality survey. *Healthy Buildings, 3*, 393–397.

Jassim, M. S., & Coskuner, G. (2017). Assessment of spatial variations of particulate matter (PM10 and PM2.5) in Bahrain identified by air quality index (AQI). *Arabian Journal of Geosciences, 10*(19), https://doi.org/10.1007/s12517-016-2808-9.

Jeong, B., Jeong, J.-W., & Park, J. S. (2016). Occupant behavior regarding the manual control of windows in residential buildings. *Energy and Buildings, 127*, 206–216. https://doi.org/10.1016/j.enbuild.2016.05.097.

Jin, W., Zhang, N., & He, J. (2015). Experimental study on the influence of a ventilated window for indoor air quality and indoor thermal environment. *Procedia Engineering, 121*, 217–224. https://doi.org/10.1016/j.proeng.2015.08.1058.

Kyrkilis, G., Chaloulakou, A., & Kassomenos, P. A. (2007). Development of an aggregate Air Quality Index for an urban Mediterranean agglomeration: Relation to potential health effects. *Environment International, 33*(5), 670–676. https://doi.org/10.1016/j.envint.2007.01.010.

Li, N., Li, J., Fan, R., & Jia, H. (2015). Probability of occupant operation of windows during transition seasons in office buildings. *Renewable Energy, 73*, 84–91. https://doi.org/10.1016/j.renene.2014.05.065.

Mandic, D. P., & Chambers, J. A. (2001). *Recurrent neural networks for prediction: Learning*

algorithms, architectures, and stability. John Wiley.
Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). *Playing atari with deep reinforcement learning.* ArXiv:1312.5602 [Cs]http://arxiv.org/abs/1312.5602.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236.
Mozer, M. C. (1998). The neural network house: An environment that adapts to its inhabitants. *AAAI Spring Symposium Intelligent Environments, 58*, 110–114.
Nagy, A., Kazmi, H., Cheaib, F., & Driesen, J. (2018). *Deep reinforcement learning for optimal control of space heating*.
Nunes de Freitas, P., & Guedes, M. C. (2015). The use of windows as environmental control in "Baixa Pombalina's" heritage buildings. *Renewable Energy, 73*, 92–98. https://doi.org/10.1016/j.renene.2014.08.029.
Pan, S., Han, Y., Wei, S., Wei, Y., Xia, L., Xie, L., Kong, X., & Yu, W. (2019). A model based on Gauss Distribution for predicting window behavior in building. *Building and Environment, 149*, 210–219. https://doi.org/10.1016/j.buildenv.2018.12.008.
Pan, S., Xiong, Y., Han, Y., Zhang, X., Xia, L., Wei, S., Wu, J., & Han, M. (2018). A study on influential factors of occupant window-opening behavior in an office building in China. *Building and Environment, 133*, 41–50. https://doi.org/10.1016/j.buildenv.2018.02.008.
Park, J. Y., Dougherty, T., Fritz, H., & Nagy, Z. (2019). LightLearn: An adaptive and occupant centered controller for lighting based on reinforcement learning. *Building and Environment, 147*, 397–414. https://doi.org/10.1016/j.buildenv.2018.10.028.
Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *International Conference on Machine Learning,* 1310–1318 arXiv:1211.5063.
Pu, H., Luo, K., Wang, P., Wang, S., & Kang, S. (2017). Spatial variation of air quality index and urban driving factors linkages: Evidence from Chinese cities. *Environmental Science and Pollution Research, 24*(5), 4457–4468. https://doi.org/10.1007/s11356-016-8181-0.
Rijal, H. B., Tuohy, P., Nicol, F., Humphreys, M. A., Samuel, A., & Clarke, J. (2008). Development of an adaptive window-opening algorithm to predict the thermal comfort, energy use and overheating in buildings. *Journal of Building Performance Simulation, 1*(1), 17–30. https://doi.org/10.1080/19401490701868448.
Rijal, H. B., Humphreys, M. A., & Nicol, J. F. (2018). Development of a window opening algorithm based on adaptive thermal comfort to predict occupant behavior in Japanese dwellings. *Japan Architectural Review, 1*(3), 310–321. https://doi.org/10.1002/2475-8876.12043.
Roulet, C.-A., Johner, N., Foradini, F., Bluyssen, P., Cox, C., De Oliveira Fernandes, E., Müller, B., & Aizlewood, C. (2006). Perceived health and comfort in relation to energy use and building characteristics. *Building Research & Information, 34*(5), 467–474. https://doi.org/10.1080/09613210600822279.
Ruelens, F., Claessens, B. J., Vandael, S., Iacovella, S., Vingerhoets, P., & Belmans, R. (2014). *Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning.* 1–7.
Ruelens, F., Iacovella, S., Claessens, B. J., & Belmans, R. (2015). Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning. *Energies, 8*, 8300–8318. https://doi.org/10.3390/en8088300.
Shaikh, P. H., Nor, N. B. M., Nallagownden, P., Elamvazuthi, I., & Ibrahim, T. (2013). Robust stochastic control model for energy and comfort management of buildings. *Australian Journal of Basic and Applied Sciences, 7*(10), 137–144.
Shi, G., Liu, D., & Wei, Q. (2017). Echo state network-based Q-learning method for optimal battery control of offices combined with renewable energy. *IET Control Theory and Applications, 11*(7), 915–922.
Shi, Z., Qian, H., Zheng, X., Lv, Z., Li, Y., Liu, L., & Nielsen, P. V. (2018). Seasonal variation of window opening behaviors in two naturally ventilated hospital wards. *Building and Environment, 130*, 85–93. https://doi.org/10.1016/j.buildenv.2017.12.019.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature, 529*(7587), 484–489. https://doi.org/10.1038/nature16961.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature, 550*(7676), 354–359. https://doi.org/10.1038/nature24270.
Singh, J. (1996). Review: Health, comfort and productivity in the indoor environment. *Indoor and Built Environment, 5*(1), 22–33. https://doi.org/10.1177/1420326X9600500105.
Stazi, F., Naspi, F., Ulpiani, G., & Di Perna, C. (2017). Indoor air quality and thermal comfort optimization in classrooms developing an automatic system for windows opening and closing. *Energy and Buildings, 139*, 732–746. https://doi.org/10.1016/j.enbuild.2017.01.017.
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (second edition). The MIT Press.
Tanner, R. A., & Henze, G. P. (2014). Stochastic control optimization for a mixed mode building considering occupant window opening behaviour. *Journal of Building Performance Simulation, 7*(6), 427–444. https://doi.org/10.1080/19401493.2013.863384.
Wang, L., & Greenberg, S. (2015). Window operation and impacts on building energy consumption. *Energy and Buildings, 92*, 313–321. https://doi.org/10.1016/j.enbuild.2015.01.060.
Watkins, C. J. C. H. (1989). *Learning from delayed rewards. Ph.D. Thesis.* University of Cambridge.
Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE, 78*(10), 1550–1560. https://doi.org/10.1109/5.58337.
Yun, G. Y., & Steemers, K. (2008). Time-dependent occupant behaviour models of window control in summer. *Building and Environment, 43*(9), 1471–1482. https://doi.org/10.1016/j.buildenv.2007.08.001.
Zhang, H., Arens, E., & Pasut, W. (2011). Air temperature thresholds for indoor comfort and perceived air quality. *Building Research & Information, 39*(2), 134–144. https://doi.org/10.1080/09613218.2011.552703.