A MULTI-AGENT REINFORCEMENT LEARNING APPROACH TO BUILDING ENERGY AND COMFORT MANAGEMENT

Mustafa Shaikh[1], Ji-Su Kim[1], Babatunde Giwa[1], Chi-Guhn Lee[1]*,
Simone Stancari[2], Davide Montanari[2], Giovanni Anceschi[2], Fabio Ferrari[2]

[1]Department of Mechanical and Industrial Engineering
5 King's College Road, M5S 3G8, Toronto, Ontario, Canada; cglee@mie.utoronto.ca
[2]Energy Way srl
Via Sant'Orsola, 37, 41121 Modena, Italy; info@energyway.it

Abstract

We present a case study in which an Italian retail store tries to reduce the energy cost of three connected heating, ventilating and air conditioning (HVAC) systems, while providing a comfortable shopping environment to its customers. The objective comprises two main terms: the total energy consumption and a penalty proportional to the aggregate deviation of comfort level among shoppers. The penalty is multiplied by a weight set by the management, and the aggregate deviation of comfort level is defined as the product of the number of shoppers and the degree of deviation of the comfort level from the ideal level. The optimization problem was formulated as a multi-agent Markov decision process, in which each agent controls its own HVAC such that the system-wide objective can be optimized. The solution is to provide a complete policy by which the store can determine the temperature set-point of the three HVAC systems in response to system state involving the external and internal temperatures, and the estimated number of shoppers. The complexity of the problem is very high and therefore we developed a reinforcement learning algorithm. To reduce the complexity further, we devised a threshold-based state space reduction method by partitioning the temperature, humidity and population range into intervals. We present numerical studies using the historical data to demonstrate the potential savings that can be achieved by the control policy found by the proposed algorithm.

Keywords

Energy and comfort management; Multi-agent system; HVAC control; Markov decision process; Reinforcement learning.

1 Introduction

We study a dynamic optimization problem in which an Italian retail store tries to optimally control three-connected heating, ventilation, and air conditioning (HVAC) systems. The objective is to minimize the weighted sum of energy cost and penalty due to discomfort among its customers. The store consists of three main sections and has sensors installed that measure external temperature, internal temperature, internal humidity, and the CO2 concentration level, every 30 minutes. The store wishes to provide a pleasant shopping

conditions to its customers while minimizing its energy cost. This lead us to develop a dynamic optimization model to determine optimal set-temperatures across the three sections.

Due to the high complexity of the dynamic optimization of decentralized HVACs, the existing literature focuses on heuristics or performance evaluation of priority rules (see Mo and Mahdavi (2003), Davisson and Boman (2005), and Zhang et al. (2010) for related studies). For example, Davidsson (2003) utilized a room agent to control the room temperature set point depending on the presence of occupants to reduce HVAC energy consumption. Also, a number of heuristic control strategies for different types of building energy systems can be found in American Society of Heating, Refrigerating and Air-Conditioning Engineers (American Society of Heating 2015, Chapter 43).

In this study, we formulated the distributed control problem as multiple Markov decision processes (MDP). To optimize the performance at the system level, the multiple MDPs are overlapped with a multi-agent reinforcement learning (MARL) approach to the energy and comfort management problem (ECMP). Reinforcement learning, a branch of machine learning, is inspired by behaviourist psychology and an adaptive method in which an agent learns via interaction with its environment an optimal control policy to choose an optimal action in each state in line with the objectives. In this study, we adopt the MARL approach for the Italian retail store's energy and comfort management problem since it has several advantages - simple to understand and interpret, possesses a high degree of scalability, ease of addition of possible new agents, etc. Interested readers are encouraged to check Kok and Nikos (2004), Buşoniu et al. (2010), and Yang and Wang (2013) for more details of the MARL approach.

The paper is organized as follows. In the next section, the problem with overall multi-agent system is described in more details. The building models and multi-agent reinforcement learning approach is explained in the third section, and the case study results are reported in the fourth section. The final section concludes the paper with a summary and discussion of future research.

2 System and problem descriptions

This section explains the system with three connected HVACs, and presents the problem in more detail.

2.1 System description

The system configuration of the system is schematically described in Figure 1, where a cooling system mechanism with the three-connected HVACs is conditioned by two chillers (Chiller 1 and Chiller 2) that produce cold water distributed to the three connected HVACs. The cold-water exchanges cooling power with the air, which is circulated in the areas (yellow, orange,

and green areas) by the ventilation systems to reach the desired internal set temperature point.

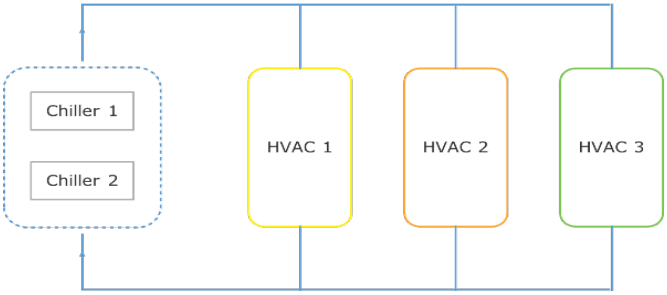


Figure 1: Cooling mechanism in retail store

The HVAC system for retail store, which is located at the Modena, Italy, considered in this case study is shown in Figure 2. The temperature control system operates its HVAC system in such a way that the area is divided into 3 sections, and two chillers, together with identical HVACs (HVAC 1, HVAC 2, and HVAC 3) assigned to each section. Currently, the HVACs are controlled identically in terms of temperature set points regardless of the different thermal and occupancy conditions of the respective areas. This leads to a sub-optimal performance in which holistic cognizance of the number of shoppers and weather conditions are ignored.

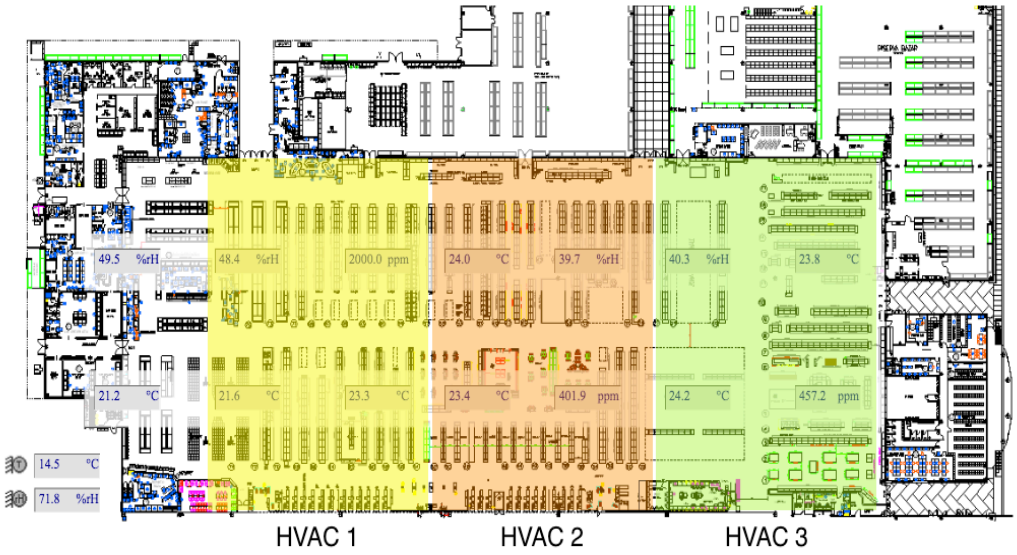


Figure 2: Three connected HVAC system

This gives rise to a need to have an adaptive optimization control of the temperature set points to allow for effective energy utilization while maintaining the desired comfort levels, especially during the summer period from 9am to 10pm daily. Note that only the cooling system will be considered in this study as the heating system is switched off.

2.2 Problem description

The problem is to determine the optimal temperature set points of three connected HVAC systems that minimize the sum of the energy consumption and a penalty proportional to the aggregate deviation of comfort level among shoppers currently present in the retail store. Sensors installed in the retail store monitor environmental states such as internal and external temperatures, internal and external humidity, energy consumption of the cooling systems, and CO_2 levels of the three regions. The system state comprises internal ($T_{internal}$), external ($T_{external}$) temperatures and population estimation (N) based on monitored CO_2 levels. The system state parameters were discretized in order to reduce the size of the state space and the complexity of the problem. We consider only store hours from 9am to 10pm daily over summer.

The problem is to determine the optimal temperature set point of each HVAC over time, based on system states. The objective function comprises two terms: the total energy consumption and a weighted discomfort penalty proportional to the deviation from an ideal value of comfort level among shoppers currently present in the store. The relative importance of the energy cost and discomfort penalty can be varied as per the requirements of the management. The objective function will be discussed in more detail in Section 3.1.

The optimal temperature control problem was formulated as a Markov Decision Processes (MDP). The control action influences the way the chain makes a transition to the next state. The MDP model includes a finite set of states (S), finite set of actions (A), transition function (T), cost function (C) and discount factor (γ). The states are combinations of internal temperature, external temperature, and occupancy level. The actions consist of all possible temperature set points. State space is defined as follows:

- $S = \{(t_{internal}, t_{external}, n)|\ t_{internal} \in T_{internal}, t_{external} \in T_{external}, n \in N\}$ where, $T_{internal} = \{22^oC, 23^oC, 24^oC, 25^oC, 26^oC\}$ is the set of possible internal temperatures; $T_{external} = \{below\ 21^oC, 22^oC \sim 26^oC, over\ 26^oC\}$ is the set of possible external temperatures; $N = \{0\sim6, 7\sim12, 13\sim18, 19\sim25, over\ 26\}$ is the set of possible occupancy levels.

Note that there are a total of 75 states in S. The set of actions are

- $A = \{22^oC, 23^oC, 24^oC, 25^oC, 26^oC\}$ temperature set points.

State transition functions and reward functions are as follows:

- $T(./s, a)$: probability distribution governing state transitions $s_{t+1} \sim T(./s_t, a_t)$.
- $q(./s, a)$: probability distribution for the rewards received $R(s_t, a_t) \sim q(./s_t, a_t)$.

3 Building models and solution algorithms

We developed a Q-learning algorithm in which three agents are responsible for optimizing the three HVACs in a coordinated manner by sharing certain information. The suggested algorithm is an extension of the standard Q-learning and the algorithm iteratively finds an equilibrium between a pair of agents. We begin with the single agent case and then extend the idea to multiple agents.

3.1 A single agent case

3.1.1 Comfort model

The comfort of customers is of significant importance to the retail store management. We introduce a penalty for the deviation of actual comfort level among present shoppers from a recommended value. To this end, we defined an index, which will be called comfort index throughout this paper, using a quantitative measure for human thermal comfort from the existing literature. Fanger et al. (1970) presented a model of determining human thermal comfort, in which the concept of the Predicted Mean Vote (PMV) and Predicted Population Dissatisfied (PPD) are used. The PMV represents a comfort level on a scale from +3 to -3: hot = +3, warm = +2, slightly warm = +1, neutral=0, slightly cool = -1, cool = -2, and cold = -3. The equation for PMV, which is called comfort index in this paper, is defined as

$$PMV = (0.303e^{-0.036M} + 0.028)\cdot L \quad (1)$$

where PMV = Predicted Mean Vote index
M = metabolic rate
L = thermal load - defined as the difference between the internal heat production and the heat loss to the actual environment - for a person at comfort skin temperature and evaporative heat loss by sweating at the actual activity level

The American Society of Heating, Refrigeration and Air Conditioning Engineers (2013) recommends that the PMV value in an indoor environment should be kept between -0.5 and +0.5. The PPD value estimates the percentage of population that will be dissatisfied with the conditions, although this parameter will not be used in this study.

The calculation of comfort index was not possible due to lack of internal relative humidity. To cope with the issue, a sensitivity analysis was performed in which PMV values were calculated for several values of relative humidity. The PMV turned out to be quite insensitive to the relative humidity: PMV varied between -0.4 and +0.4 as the relative humidity varied from 30% to 60%. However, relative humidity tends to be very stable in indoor setting: the minimum humidity was 35.5%, the mean was 54.5%, and the standard deviation 6.9% in the particular store used in our case study. Therefore, we assumed that the internal relative humidity remains constant at the mean value of 54.5%.

The values of comfort index for 40 combinations of internal and external temperatures are shown in Figure 3. Note that the index is -0.4 when T_{ext} = 34°C and T_{int} = 22°C. This means that shoppers would feel uncomfortable (too cold) if the internal temperature is loo low relative to the external temperature. This is because shoppers, who tend to wear thin clothing due to the high external temperature (34°C), are likely to feel uncomfortably cold in the store (22°C).

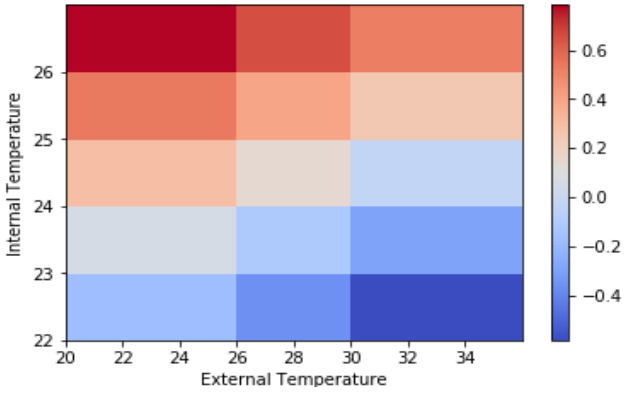


Figure 3: Comfort index for combinations of internal and external temperature

3.1.2 Occupancy estimation model

The level of occupancy in the store is an important input to the dynamic temperature control. The penalty for customer discomfort will be proportional to the number of shoppers in the store. Ito and Nishi (2012) presented an estimation model based on a mass balance on an enclosed space as follows:

$$n = \frac{Qvent}{k\left(1-e^{-\frac{Qvent}{v}(i-s)}\right)}\left(C_i - C_0 - (C_s - C_0)e^{-\frac{Qvent}{v}(i-s)}\right) \quad (2)$$

where n = number of people

Q_{vent} = ventilation rate

V = volume of the enclosed space

i = initial time

s = current time

k = CO_2 emissions per person

C_s = CO_2 concentration at time s

C_0 = CO_2 concentration in empty room

The value for *Qvent* was assumed to be 0.06 cfm/ft^2 according to the guidelines of the American Society of Heating Refrigeration and Air Conditioning Engineers (2013), and k was assumed to be 0.008L/s (Emmerich and Persily, 2001).

The estimated population size seemed to be prone to noise. We applied Savitsky-Golay filter (Savitsky and Golay, 1964) to remove the noise. Figure 4 shows the estimation with and without Savitsky-Golay filter.

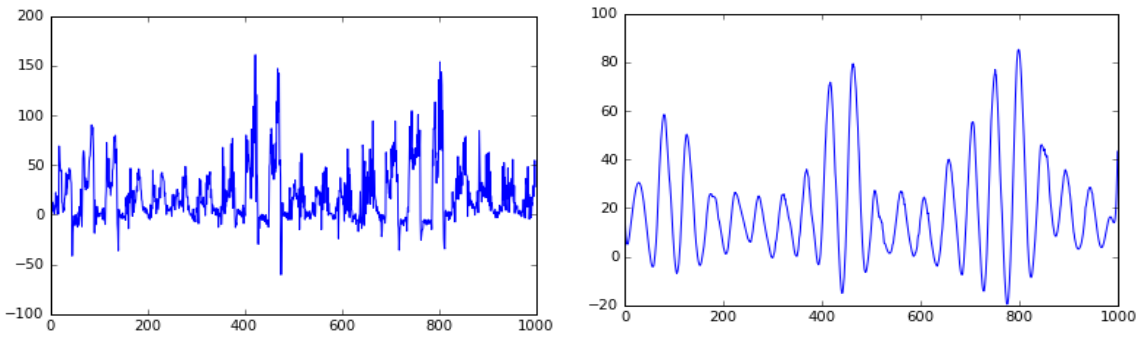


Figure 4: Unfiltered data vs. filtered data

3.1.3 Energy cost

We used a linear regression over the historical data to find the following function for energy cost:

$$\text{Cost} = 16.866T_{ext} + 1.4036Hu_{ext} - 23.1658T_{int} + 0.1032I_{co2} - 12.399T_{setpt} + 298.627 \quad (3)$$

where T_{ext} = external temperature (°C)
Hu_{ext} = external relative humidity (%)
T_{int} = internal temperature (°C)
I_{CO2} = CO_2 level (ppm)
T_{setpt} = setpoint temperature (°C)

3.1.4 Objective function

The objective is the sum of the energy cost and the penalty due to discomfort among shoppers, which is as follows:

$$C(s_t, a_t) = E_t + n \cdot w_c \cdot PMV_t \quad (4)$$

where E_t is the energy cost, PMV_t is the penalty for discomfort among shoppers, w_c is the weight set by the management, add n is the number of shoppers present.

3.15 Q-learning algorithm

A variation of the Q learning algorithm (Watkins, 1989) is used in this study. In this algorithm, the parameter i corresponds to the number of iterations. The "Initialize" function initializes the Q function with Q_0, which is initialized by choosing a random initial state and computing the Q value. The procedure of proposed Q-learning algorithm is presented in Figure 5. The algorithm begins with initialization of Q-matrix by setting all the entries to be 0. At the core of the Q-learning algorithm is a value iteration, which seeks to find the maximum reward for taking an action at the current state as well as a discounted reward for future actions. The Q-matrix is updated iteratively according to the following equation:

$$Q_i(s, a) = (1 - \alpha_i) \cdot Q_{i-1}(s, a) + \alpha_i [r(s, a) + \gamma \cdot max_a Q(s', a')] \quad (5)$$

Where $Q_t(s, a)$ = Q value associated with taking an action a in state s

$Q_t(s', a')$ = Q value associated with taking an action a' in the next state (s')

α_t = learning rate, learning rate was set to 1

r_t = immediate reward for taking action a in state s

γ = discount factor, discount factor was set to *0.99*

The particular Q-learning algorithm developed in this research is presented in Figure 5.

The Q-learning algorithm

Initialize Q_0

for $i = 0$ **to** $i = n$, **do**

$s_i \leftarrow$ Choose state

$a_i \leftarrow$ Choose action

{update Q_i}:

$$Q_i(s,a) = (1-\alpha_i)\big(Q_{i-1}(s,a)\big) + \alpha_i(R(s,a) + \gamma * max_a Q_i(s',a')$$

{initial smoothing of Q_i}:

if $i = 5000$ **and** $(Q_i(s) = 0 \text{ @ } i = 5000)$ **do**

$$Q_i(s) = \tfrac{1}{6}\big(Q_i(s^{*+}) + Q_i(s^{*-})\big) + \tfrac{1}{6}\big(Q_i(s^{**+}) + Q_i(s^{**-})\big) + \tfrac{1}{6}\big(Q_i(s^{***+}) + Q_i(s^{***-})\big)$$

{subsequent smoothing of Q_i}:

if $i > 5000$ **and** $i \,\%\, 1000 = 0$ **do**

for all $Q_i(s)$ such that $(Q_i(s) = 0 \text{ @ } i = 5000)$ **do**

$$Q_i(s) = 0.85 * \big(Q_i(s)\big) + 0.05 * \Big(\tfrac{1}{2}Q_i(s^{*+}) + \tfrac{1}{2}Q_i(s^{*-})\Big) + 0.05 * \Big(\tfrac{1}{2}Q_i(s^{**+}) + \tfrac{1}{2}Q_i(s^{**-})\Big) + 0.05 * \big(\tfrac{1}{2}Q_i(s^{***+}) + \tfrac{1}{2}Q_i(s^{***-})\big)$$

s^{*+} , s^{*-} : states with T_{int} 1 larger and 1 smaller than T_{int} @ Q_i(s)

s^{**+}, s^{**-}: states with n 1 larger and 1 smaller than n @ Q_i(s)

s^{***+} , s^{***-}: states with T_{ext} 1 larger and 1 smaller than n @ Q_i(s)

Figure 5: Procedure of proposed Q-learning algorithm

Note that the smoothing step is unique in our algorithm and is mainly to improve the convergence rate of the Q-matrix. We trained the Q-matrix for 5,000 iterations before smoothing. Beyond 5,000 iterations, we smoothed the Q-matrix every 1,000 iterations by taking an average of six neighbors plus the entry that was just updated.

3.2 Extension to Multi Agent System

An extension of the proposed Q-learning algorithm to multiple agents is based on the idea of pairwise equilibrium. The algorithm starts with a pair of adjacent agents, each responsible for a HVAC, and tries to find an equilibrium between the two agents. To this end, algorithm find an optimal policy for one agent as an optimal response to the optimal policy of the other agent. Once an equilibrium is found, we switch to a new pair consisting of an agent from the current pair and a new agent. When an equilibrium is found between the chosen pair, we repeat the procedure for a new pair in which one agent is from the current pair. The algorithm continues until equilibria are found for every pair.

In a two-agent Q-learning, the state definition should be augmented so that each agent should be able to take action conditional on the state of the two-HAVC system as well as action to be chosen by the other agent in the given state. Therefore, the state definition should be as follows:

- $S = \{(t_{internal,1}, t_{internal,2}, n_1, n_2, t_{external}, a) \mid t_{internal,i} \in T_{internal}, t_{external} \in T_{external}, n_i \in N, a \in A\}$

where $t_{internal,\,k}$ and n_k are internal temperature and the number of shoppers in the section controlled by agent k (k=1,2). Note that external temperature is common for both agents and denoted as $t_{texternal}$.

4 Test results

The test results can be presented in 5 parts: (a) convergence of Q-matrix, (b) visualization of the Q matrix, (c) visualization of the optimal policy, (d) cost comparisons of the optimal policy with current policy, and (e) system simulation. The proposed Q-learning algorithm was implemented in Python and run on a desktop with an Intel Core i5 at 2.00 GHz clock speed.

(a) Convergence of Q values: We used a stopping rule based on the sample variance of Q-matrix over iterations. When the sample variance becomes less than a given tolerance, the training ends. In most cases, the algorithm stopped shortly after 120,000 iterations. A few entries of the Q-matrix over iterations are shown in Figure 6. Note that the convergence of state 66 is significantly smoother than the convergence of the other states. This is because state 66 underwent the smoothing procedure outlined earlier. The oscillatory behavior seen for the other states can be mitigated by a number of techniques, some of which are currently being investigated.

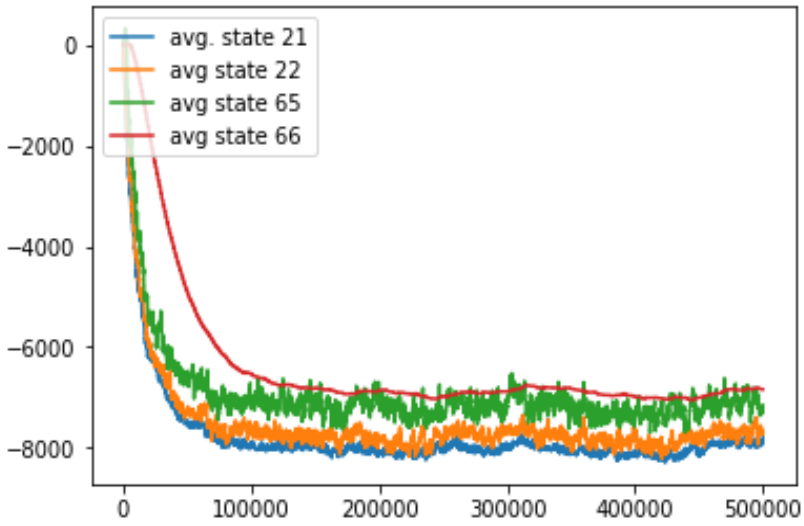


Figure 6: Q-value convergence for states 21,22,65,66

(b) The Q matrix: Figures 7 and 8 show the Q matrix at convergence. Note that lighter colour indicates a lower Q value and thus a better action. In Figure 7, the Q-value is unimodal along the x- and y-axis, as expected. However, the unimodal pattern is not found in Figure 8.

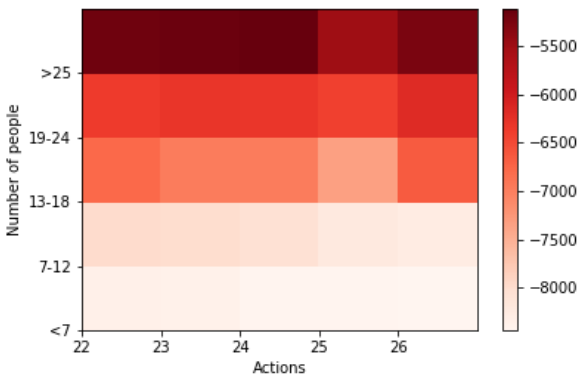


Figure 7: Q matrix for T_{int} = 26, Text = Low

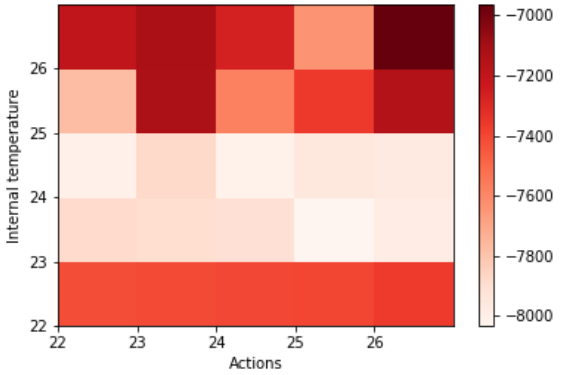
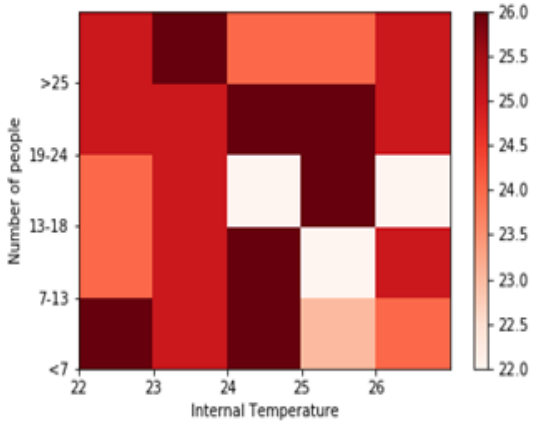


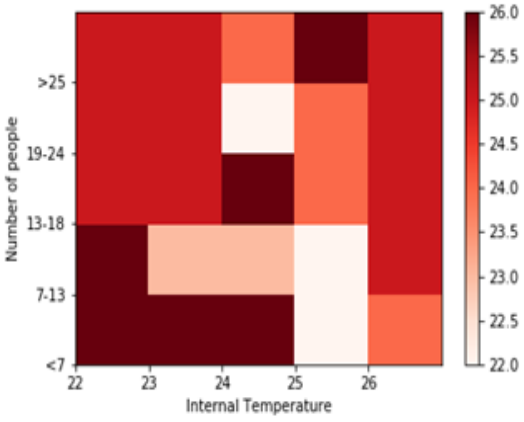
Figure 8: Q matrix for occupancy level = {13-18}

(c) Optimal policy: Figure 10 shows optimal set point temperature for various system state. Three graphs are for three external temperatures: high, medium and low. The set point temperature is represented by the colour scale, where each of the 5 colours correspond to a particular set point (see the color bar next to each map for detail). For example, looking at the map for T_{ext} = High, we see that the optimal action at an internal temperature of 25°C and >25 people, is to set the temperature to 24°C.

T_{ext} = High



T_{ext} = Medium



T_{ext} = Low

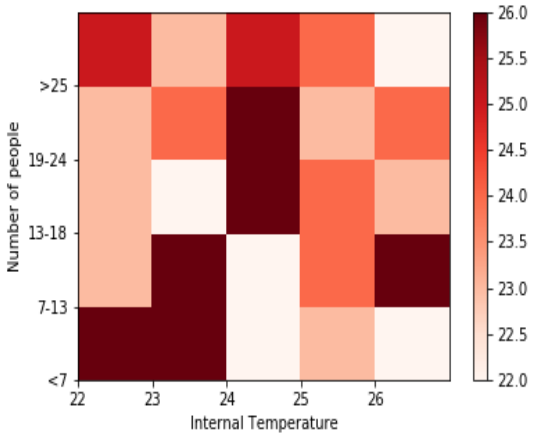


Figure 9: Optimal policy visualization

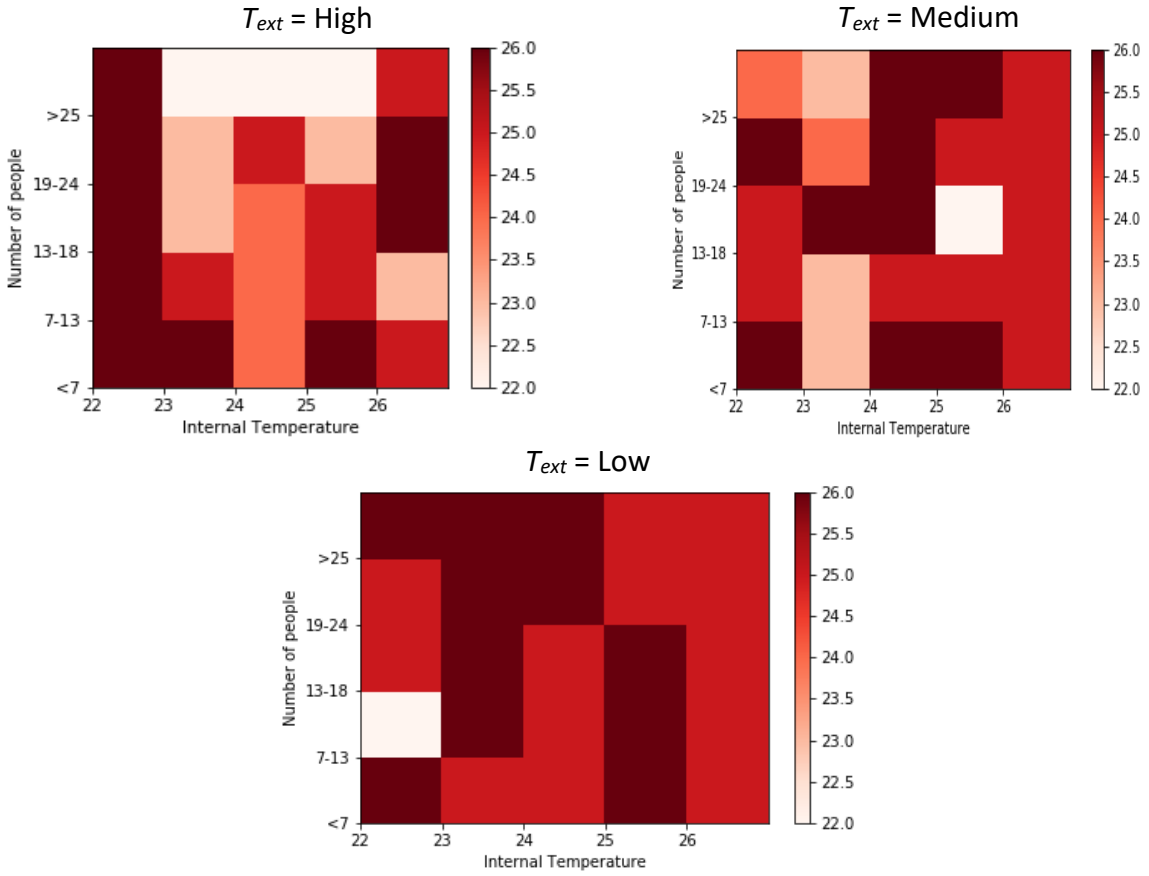


Figure 10: Optimal policy visualization for Q matrix without smoothing, epsilon stopping

We observe that the optimal policy does not show gradual patterns. This can be due to the Q-matrix being non-unimodal, as discussed earlier. It was shown earlier that some states has their Q-values oscillating at convergence.

We now compare the optimal policy dynamics of the smoothed Q matrix with epsilon stopping, to the Q matrix without smoothing or epsilon stopping. Figure 10 shows the optimal policy

without smoothing. We see that the optimal policy is difficult to understand, and is significantly less smooth than the optimal policy with smoothing.

(d) The optimal policy vs. the current policy: Following the determination of the optimal policy, we conducted cost comparisons between the optimal policy and the current policy being implemented at the retail store. Total cumulative cost, as well as energy and comfort cost separately, were determined using the data set. At each time step, the cost for taking the action that was historically taken, in the state that the system was in at the time, was computed. The procedure was then carried out for the optimal policy; the initial state was chosen as per the historical data, the optimal action selected from the Q matrix, and the next state chosen probabilistically using the state transition matrices. This entire process was repeated several times, with newly trained Q matrices. From figure 11, we see that the optimal cost is consistently lower than the current cost, but the absolute values of the optimal cost vary significantly. This again is a result of the oscillatory behaviour seen during convergence of the Q matrix. Consecutive training episodes yield slightly different Q values which, over time, affect the optimal policy and total incurred costs.

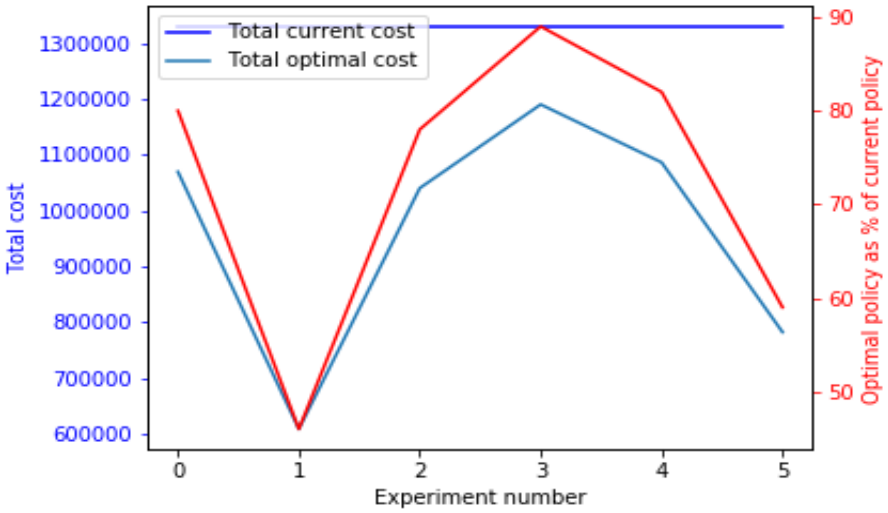


Figure 11: Cost comparison of optimal policy vs. current policy

Figure 12 below shows a decomposition of the cost term into its constituent components; energy and discomfort. We see that the discomfort cost resulting from following the optimal policy always remains below the current discomfort cost. The energy cost from the optimal policy, however, is always greater than the current energy cost, except at training episode 3. This is caused by the effect of the discomfort penalty weight on the relative importance of discomfort and energy; the current weight seems to slightly favour discomfort cost over energy cost. Adjusting this weight will change the relative magnitudes of the energy and discomfort costs – see figure 13 for details. Note that the energy costs are negative; this is simply a technicality resulting from the method of linear regression performed to develop a model for the energy cost. For energy cost, then, more negative is better as it represents higher 'reward'.

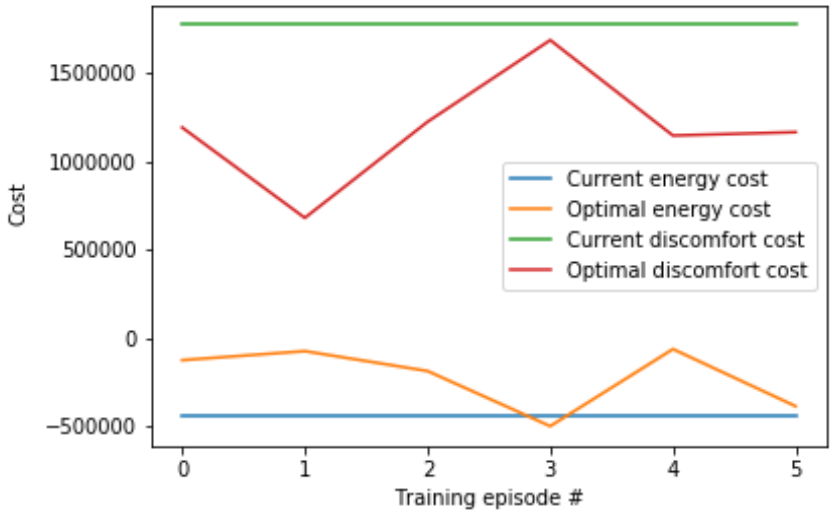


Figure 12: Decomposition of costs over several training episodes

Figure 13 below shows the effect of varying the discomfort penalty weight on the energy and discomfort costs. As we expect, the current energy cost will be unaffected, whilst the discomfort costs for both the current and optimal policy will increase with increasing weight. Again, we see that the discomfort cost from the optimal policy is always lower than the discomfort cost from the current policy. The energy cost also varies as expected; we see that the energy cost from the optimal policy decreases as the comfort weight decreases. This is expected as the optimal policy assigns greater importance to energy relative to discomfort and the optimal actions will lead to increased energy savings at the expense of higher discomfort costs. Therefore, altering the discomfort penalty weight affects both the discomfort cost as well as the energy cost.

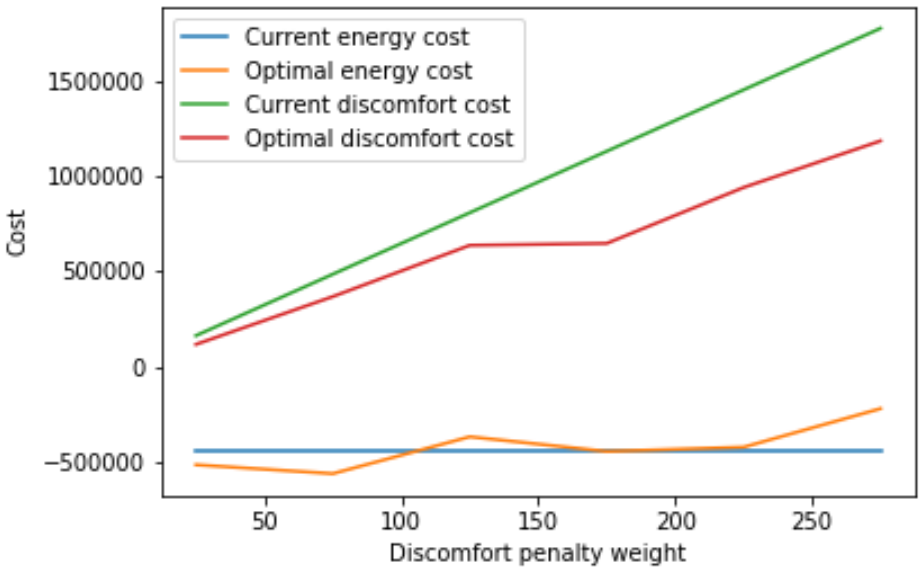


Figure 13: Variation of cost terms with discomfort penalty weight

(e) System simulation: Figure 14 shows a simulation over 1,000 time steps. A random starting state was chosen, the optimal policy was determined from the Q matrix, and the next state chosen probabilistically. For that next state, the optimal action was determined, and the process was repeated 1000 times. This simulation serves to illustrate the most common action selections seen by the optimal policy over 1000 decision steps. It can be observed that the most common optimal actions (i.e. red dots) were temperature set points of 23°C, 24°C and 25°C.

As stated earlier, we report the expected results for the three connected HVACs since the multi agent system is still under investigation. We intend to consider an extension to the multi agent system based on Q-learning, as outlined earlier.

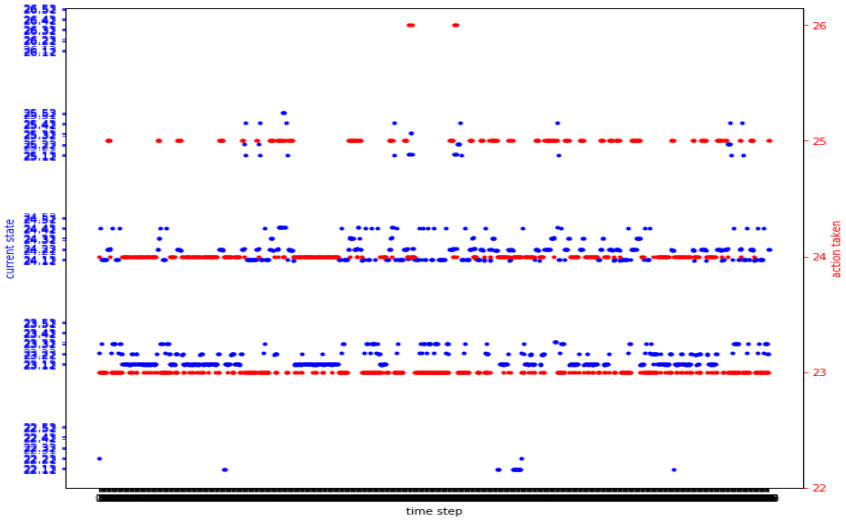


Figure 14: Simulation over 1000 time steps

5 Conclusions

In this study, we developed a variant of the Q-learning algorithm and applied the method to the energy and comfort management of an Italian retail store. The main decision is selecting the temperature set point for each HVAC at each time step. The multi agent reinforcement learning approach is used to select a temperature set point at each HVAC appropriate for a specific system state and hence the burdens required for developing simulation models and carrying out simulation runs can be reduced. The results for the single agent system were presented, with notes regarding further work.

This study can be extended in several directions. First, the single agent approach can be extended to the multi agent system, as outlined in section 3.3. Second, an improved solution could be attained with the inclusion of the heat transfer dynamics among the three HVAC systems. Another extension to the study involves using model free forms of temporal difference learning, and using bootstrapping techniques to learn the data and optimize post-learning. Finally, case studies for other applications are worth considering.

Acknowledgements

Dr. Ji-Su Kim was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education of Korea government (Grant number: NRF-2016R1A6A3A03010592).

References

[1] Mo, Z., An agent-based simulation-assisted approach to bi-lateral building systems control, Doctoral Dissertation, Carnegie Mellon University, Pittsburgh, PA, USA, 2003.

[2] Davidsson, P., *Distributed monitoring and control of office buildings by embedded agents*, *Information Sciences*, *Volume 171* (2005), Issue 4, pp. 293-307.

[3] Zhang, L., Building energy saving design based on multi-Agent system, 5th IEEE Conference on Industrial Electronics and Applications, 2010, pp. 840-844.

[4] American Society of Heating, Chapter 43. HVAC Commissioning in *Heating, ventilating, and air-conditioning applications* (Inch-Pound Edition), Refrigerating and Air-Conditioning Engineers, Inc. Atlanta, GA, USA, 2015, pp. 1-14.

[5] Kok, J. R., and Nikos, V., Sparse cooperative Q-learning. *Proceedings of the twenty-first international conference on Machine learning*, Banff, Canada, 2004, pp. 61-68.

[6] Yang, R., and Wang, L., *Development of multi-agent system for building energy and comfort management based on occupant behaviors*, *Energy and Buildings*, *Volume 56,* (2013), Issue 1, pp. 1-7.

[7] Buşoniu, L., Babuška, R., and De Schutter, B., Multi-agent reinforcement learning: An overview, *Innovations in multi-agent systems and applications-1*. Springer, Berlin, Germany, 2010, 183-221.

[8] Fanger, P., Thermal Comfort: Analysis and applications in environmental engineering, Danish Technical Press, Copenhagen, Denmark, 1970.

[9] American Society of Heating Refrigeration and Air Conditioning Engineers, *Thermal environmental conditions for human occupancy*, Inc. Atlanta, GA, USA, 2013.

[10] Ito, S., and Nishi, H., Estimation of the number of people under controlled ventilation using a CO2 concentration sensor, 38th Annual Conference on IEEE Industrial Electronics Society, Quebec, Canada, 2012, pp. 4843-4839.

[11] Emmerich, S., and Persily, A., State of the art review of CO2 demand controlled ventilation technology and application, National Institute of Standards and Technology, U.S., 2001.

[12] Savitsky, A., and Golay, M. J. E., Smoothing and differentiation of data by simplified least squares procedures, *Analytical Chemistry*, *Volume 36*, (1964), Issue 8, pp. 1627-1639.

[13] Watkins, C., Learning from delayed rewards, Imperial University, London, England, 1989