Master's Thesis of Engineering

Occupant Diversity Analysis for Reliable Modeling of Occupant Behavior

재실자 행동 모델링을 위한 재실자 다양성 분석

August 2023

Department of Architecture and Architectural Engineering
College of Engineering
Seoul National University

Sunghyun Kim

Occupant Diversity Analysis for Reliable Modeling of Occupant Behavior

Advised by Prof. Cheol-Soo Park

Submitting a master's thesis of Engineering
August 2023

Department of architecture and Architectural Engineering
College of Engineering
Seoul National University

Sunghyun Kim

Confirming the master's thesis written by
Sunghyun Kim

August 2023

Chair ___Myoung-Souk Yeo___(Seal)
Vice Chair ___Cheol-Soo Park___(Seal)
Examiner ___Sun Sook Kim___(Seal)

Abstract

Occupant behavior (OB) plays a crucial role in building performance simulations, but its complexity and variability pose challenges for accurate modeling. This thesis emphasizes the importance of reproducibility and replicability in OB models and explores the impact of occupant diversity on building energy control and prediction. Three types of occupant diversity are identified: temporal, spatial, and behavioral. In this thesis, in-situ experiments were conducted in three residential buildings in Seoul, South Korea, involving 31 households to investigate occupant diversity. Various aspects of occupant behavior, including occupant presence, window state, light switch, AC switch, and Boiler switch, as well as indoor and outdoor environmental data were collected.

The results showed significant temporal diversity in occupant presence, highlighting the need for considering the temporal variability of behavior in OB models. The analysis of window adjustment behavior revealed individual preferences and the influence of multiple factors. Furthermore, variations in behavior types among households demonstrated diverse perspectives on indoor environment control and energy conservation. To address the performance gap in building simulations resulting from occupant behavior modeling, this research underscores the importance of considering occupant diversity to improve the accuracy and effectiveness of building performance simulations. Future research should focus on developing more reliable and reproducible occupant models incorporating occupant diversity, bridging the gap between actual and simulated building energy use.

Keywords: Occupant behavior, Occupant diversity, Residential buildings, Occupant modeling, Building simulation

Student Number: 2021-27940

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Background

Occupant behavior (OB), which describes occupant interactions with buildings, is one of the main sources of uncertainty in building performance simulations (Yan et al., 2017; O'Brian et al., 2020; Dong et al., 2022). Understanding and accurately modeling occupant behavior is crucial to achieving occupant-centric building energy control and predicting energy demand (Yang et al., 2022; O'Brian et al., 2017; Ahn et al., 2017; Carlucci et al., 2020; D'Oca et al., 2018; Wagner et al., 2018; Norouziasl et al., 2021).

Many attempts have been made to develop OB models, for example, utilizing occupant presence, window operation, shading operation, lighting control, thermostat adjustment, appliance use, and clothing (Page et al., 2008; Wang et al., 2011; Salimi et al., 2019; Langevin et al., 2015; Park et al., 2019; Yilmaz et al., 2017; Qu et al., 2021). Page et al. (Page et al., 2008) developed an occupant presence model with Markov chain transition probabilities to generate a time-series for each occupancy in a single zone. The model can reproduce key occupancy properties, such as arrival and departure times. Wang et al. (2011) modeled the occupant movement occurring in the spaces inside and outside a building. The Markov chain approach was used to simulate the stochastic movement of the occupants. Salimi et al. (2019) enhanced occupancy modeling using an inhomogeneous Markov chain prediction model based on real occupancy data. Further, the agent-based thermal adjustment has been simulated (Langevin et al., 2015), focusing on unconstrained adaptive behaviors to maintain thermal sensation, for example, the occupants' fan, heater, and window use. Park et al. (2019) developed a lighting control model based on reinforcement learning (RL) and trained on individual occupant behavior and indoor environmental conditions to determine personalized set points. Yilmaz et al. (2017) simulated three appliance operations using stochastic processes to capture daily variations in appliance occupant behavior. Qu et al. (2021) modeled a logistic outdoor clothing adjustment based on the assumption that local past temperatures influenced it. A four-parameter logistic function was used for the logistic regression.

However, despite the scientific evidence indicating that OB models contribute to the performance gap, they are still occasionally used in building energy control and predicting energy usage. Many simulation programs like EnergyPlus default to deterministic OB schedules, which simplify considering occupant preferences in indoor environments and energy control processes. These simplified approaches often rely on predefined schedules or simplified occupancy models that do not fully capture the complexity and variability of occupant behavior. This oversimplification of occupant preferences and behavior can lead to discrepancies between simulated and actual energy consumption, as well as suboptimal indoor comfort and energy efficiency (Azar et al., 2012). It highlights the need for more sophisticated and realistic modeling approaches that better represent the diverse range of occupant behaviors and preferences. Previous research attributes this phenomenon to the lack of OB model standardization and clear documentation (Dong et al., 2018; Luo et al., 2021), which results in models' limited reproducibility and replicability (Dong et al., 2022).

In this thesis, the author emphasizes the importance of reproducibility and replicability, concepts that are often overlooked in current research practices. Reproducibility refers to the ability to obtain consistent results using the same input data, while replicability refers to obtaining consistent results across studies that aim to address the same scientific questions but employ different data (Dong et al., 2022). Through the pursuit of reproducibility and replicability, current researches can address the performance gap in OB models and establish a more reliable and trustworthy foundation for building energy control and prediction.

In addition, it is widely believed that OB patterns in residential buildings would follow a regular pattern (Aragon et al., 2019; Richardson et al., 2008). However, other studies have emphasized that because occupants have diverse occupant profiles, describing OB in residential buildings is complicated (Balvedi et al., 2018; Li and Dong, 2017; Carlucci et al., 2016). In such residential buildings, whether occupant behavior is predictable and whether similar results are found for each household remains indeterminate.

1.2. Main objectives

In this thesis, the author argues that occupant diversity hinders reproducibility and replicability, resulting in a performance gap in OB models. For example, assuming an average occupant or using deterministic patterns can be problematic. The author categorizes occupant diversity into tree types as follows (Figure 1.1):

- Temporal diversity: Occupants exhibit different behaviors in the same environment at different times. This emphasizes the need to consider the temporal variability of occupant behavior and capture its dynamics in OB models.

- Spatial diversity: Occupants in different spaces, such as different households or buildings, demonstrate varying behavior patterns. Recognizing spatial diversity is crucial for developing accurate OB models that reflect occupants' specific characteristics and preferences in different contexts.
- Behavioral diversity: Different types of behaviors, such as occupancy, window states, light switches, etc., require distinct modeling approaches. Considering the unique characteristics and dependencies among various behavior types is essential for improving the fidelity and performance of OB models.

In this thesis, the author aims to demonstrate the existence and quantify the three types of occupant diversity mentioned above. By doing so, the thesis highlights the importance of considering occupant diversity in OB models and its impact on reproducibility and replicability. Through empirical analysis and data-driven approaches, the research provides evidence for the variability and heterogeneity of occupant behavior across different temporal, spatial, and behavioral contexts.

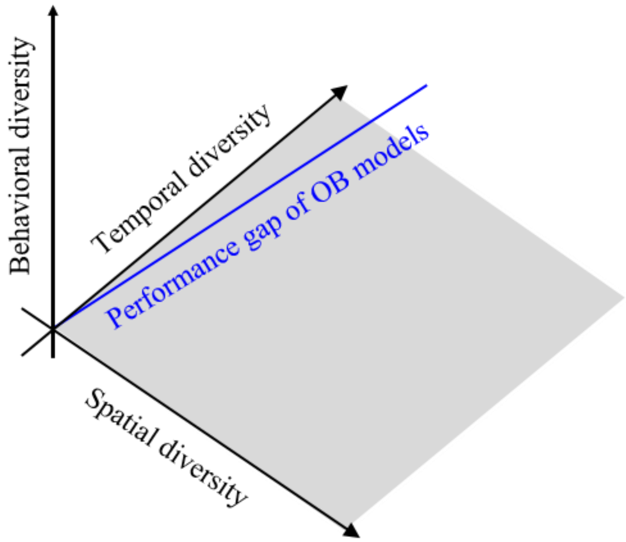


Figure 1.1 Three types of occupant diversity and performance gap of OB models

To implement the OB model and investigate occupant diversity, the author conducted in-situ experiments in three residential buildings in Seoul, South Korea. A total of 31 households were selected for data collection.

The collected data encompassed various aspects of occupant behavior, including occupant presence, window state, light switch, AC switch, and Boiler switch. In addition to occupant behavior data, environmental data deemed influential on occupant behavior were also collected (Wei et al., 2014). This included indoor and outdoor temperature, indoor and outdoor humidity, indoor CO_2 concentration, indoor PM2.5 concentration, and indoor illuminance.

1.3 Thesis organization

- **Chapter 1**: Provides an introduction to the background and objectives of the thesis. Defines occupant diversity into three types: temporal, spatial, and behavioral.
- **Chapter 2**: Describes the data collection process, focusing on the target residential buildings and sensor information utilized in the research.
- **Chapter 3**: Examines temporal diversity by employing random walk theory to identify the presence of occupancy patterns within each household over time. Demonstrates the existence and quantification of temporal diversity and presents guidelines for mitigating its impact.
- **Chapter 4**: Investigates spatial diversity using Explainable AI (XAI) techniques to model occupants' window adjustment behavior (WAB). Quantifies the feature influence of various environmental and occupant factors, highlighting the variation in environmental perception across different households.
- **Chapter 5**: Develops a multinomial OB model using LSTM (Long Short-Term Memory) to explore behavioral diversity. Analyzes the diversity of occupant behavior types and quantifies the degree of dependence between different behaviors. Addresses potential issues in multinomial OB modeling.
- **Chapter 6**: Concludes the thesis by addressing the findings of this thesis and presenting future research directions.

Table 1.1 Thesis organization

Chapter	**Title**	**Problem suggestion**	**Objectives**	**Solution (Methodology)**
1	Introduction	Why does the performance gap occur in building simulations due to occupant behavior (OB) models?	Problem suggestion, specification of objectives	Literature review
2	Target buildings	-	Data collection for the experiments	Sensor installation, digital mapping
3	Predictability quantification in occupant presence	Do occupants exhibit the same behavior in the same environment at different times?	Quantification of predictability of occupant presence in terms of temporal diversity	Random walk approach
4	Feature influence quantification in window adjustment behavior using XAI	Do different occupants have the same perspective on indoor environmental control?	Investigating of individual preferences in window adjustment behavior in terms of spatial diversity	XAI (Logistic regression, XGBoost, SHAP)
5	Multinomial occupant behavior model	Can different types of behaviors be modeled identically based on the same environmental data?	Implementing multinomial occupant behavior model and analyzing degree of dependence between different behaviors	Multinomial OB modeling using LSTM, mutual information
6	Conclusion	-	Summary of findings regarding occupant diversity, future work	-

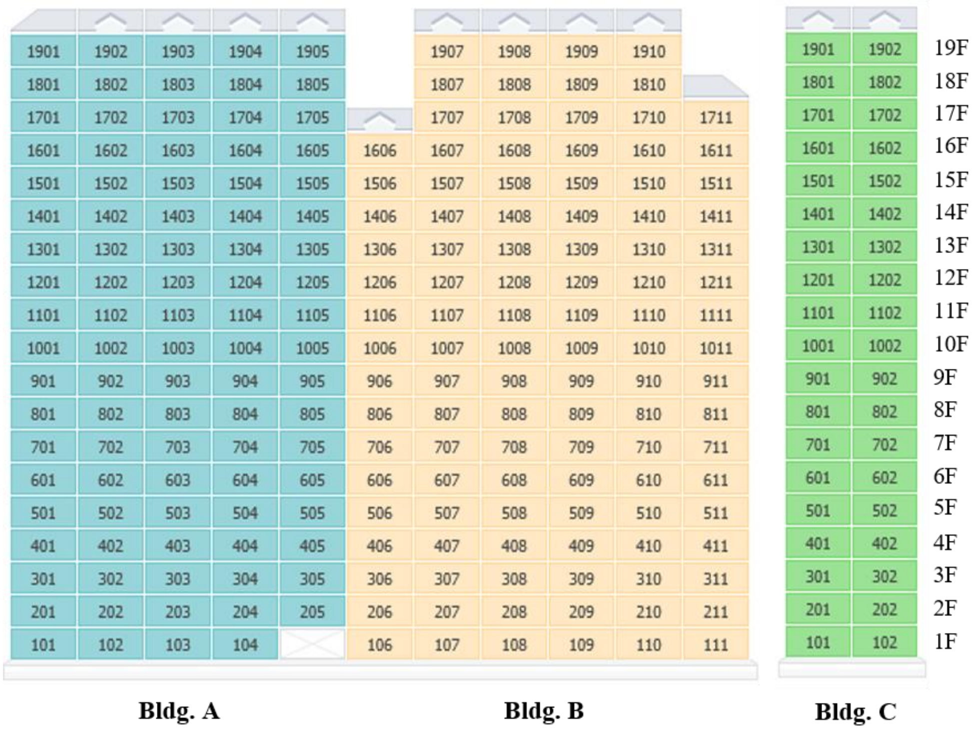
Chapter 2. Experiments

2.1 Target buildings

The author conducted a pre-survey and selected 31 households who agreed to participate in this chapter. Those 31 households in three residential apartment buildings in Seoul, South Korea, were selected (Figure 2.1). The number of residences in each household was collected through a survey. All households were naturally ventilated, apart from the temporarily running exhaust ventilation from the kitchen hood and bathroom vents. The outside noise problem was negligible with low traffic on nearby roads and a large height of five or more floors of households. In addition, the entire building was non-smoking, so no constraints on window opening were imposed. In each household, fluorescent lights are installed on the ceilings of all rooms, and depending on the household, separate lighting fixtures are installed as well. The AC is based on the device installed in the living room of each household and is also autonomously controlled by the occupants in terms of on/off, set point, and operation mode. The boiler operates as a water heating and floor heating system, and the occupants in each household autonomously control the on/off, set point, and operating modes. It is unknown whether separate heating/cooling devices such as fans or electric heating mats exist in the households, as this information was not collected.

(a) Exterior view (Bldg. A, B, and C from right to left)

(b) Site plan of the three apartment buildings



(c) Elevation of the three apartment buildings

Figure 2.1 Target buildings

2.2 Sensor installation for measuring occupant and environmental data

USM-300-ZB multi-sensors (Figure 2.2(g), developed by Shinasys) were installed in all living spaces (Figure 2.4, three(four) bedrooms and a living room) to measure indoor temperature, humidity, illuminance, and occupant presence. The indoor environmental data were recorded using sensors in the living room adjacent to the balcony. The USM sensor employs a PIR sensor that detects occupant presence. The USM sensors were installed in all living spaces, e.g., three(four) bedrooms and a living room (Figures 2.4(a)–(c)). Because the kitchen can be regarded as part of the living room, therefore one USM sensor was installed in the living room. Based on the measured data, it was found that the PIR sensor could detect occupant presence in the nocturnal periods. If occupant movement was detected by any USM sensor in the living spaces of the corresponding household, it was recorded that occupants were in their homes. Finally, it was assumed that occupants were present in the space when any movement was recorded at least once during the sampling interval of temporal resolution.

The window states (0: closed, 1: open) of each household were recorded using DSM-300-ZB window sensors (Figure 2.2(d), developed by Shinasys). The DSM sensor was installed only on the openable window outside the main balcony (Figures 2.4 and 2.5). Each balcony space is isolated from the interior space with inner walls and glass doors with weak insulation and airtightness performance. Therefore, the opening and closing of the external window considerably affect the indoor environment.

The states of the living room ceiling light (0: off, 1: on) in each household were recorded using the STM-300-W smart lighting controller (Figure 2.2(e), developed by Shinasys). Although occupants control the ceiling lights in all rooms through the controller, only the living room lighting was considered to simplify the prediction model.

The CCM-300-W (Figure 2.2(c), developed by Shinasys) measures the power consumption of each electrical outlet, and based on the power consumption of the outlet connected to the AC, it determines the control states (0: off, 1: on) of the AC. If the power consumption of that outlet exceeded 30W, it was recorded as the air conditioner in operation. Notably, the power consumption difference between when the AC is in operation and when it is not is distinct, rendering the threshold value less meaningful.

The boiler is used for both supplying hot water and floor heating within each household. The BCM-300-W records the operating mode set by occupants (only hot water supply, only floor heating, both hot water supply and floor heating). The author extracted the state when the switch is turned on, and floor heating is in progress (0: off, 1: on).

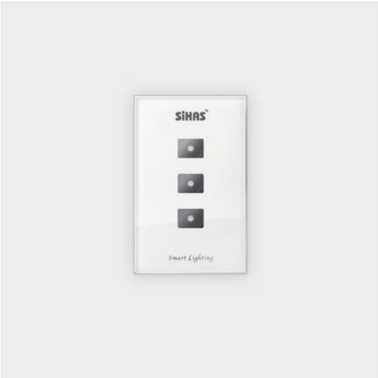
The indoor CO_2 and $PM_{2.5}$ concentrations were recorded using AQM-300-W air quality sensors (Figure 2.2(a), developed by Shinasys), installed next to each USM sensor. Figure 2.3 illustrates the data collection process of the sensors. Figure 2.4 shows the specific locations of each sensor type. Figure 2.5 shows sensor installation and location in an experimental space of household #4. Outdoor temperature and humidity data were provided by the Korea Meteorological Administration weather data service. Each outdoor environmental sensor works at the ground level in Seoul, South Korea, within 15 km of the target buildings. Table 2.1 specifies sensors used for measuring occupant and environmental data.

(a) AQM-300-W

(b) BCM-300-W

(c) CCM-300-W

(d) DSM-300-ZB



(e) STM-300-W

(g) USM-300-ZB

Figure 2.2 Photos of sensors used for measuring occupant and environmental data

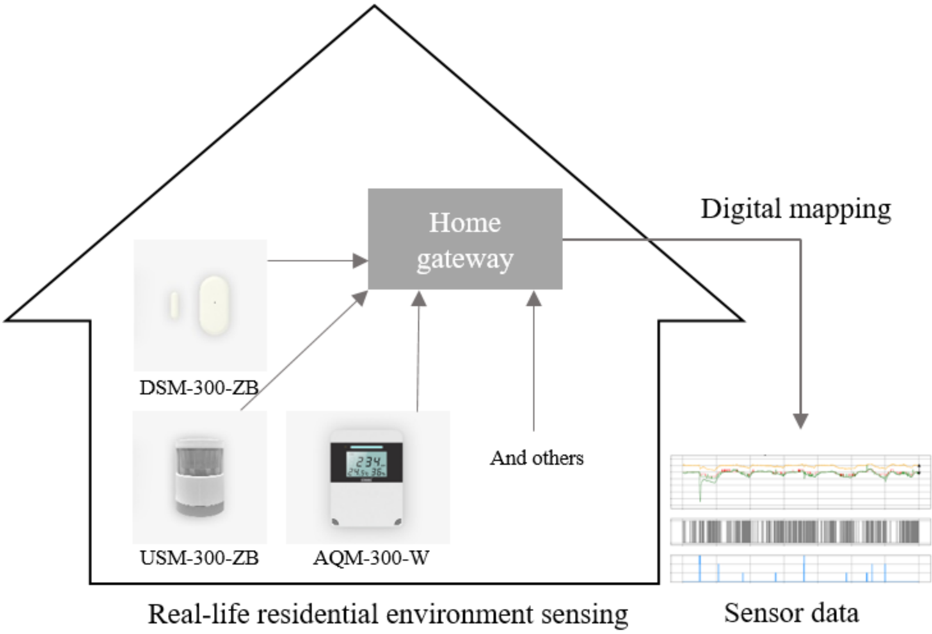
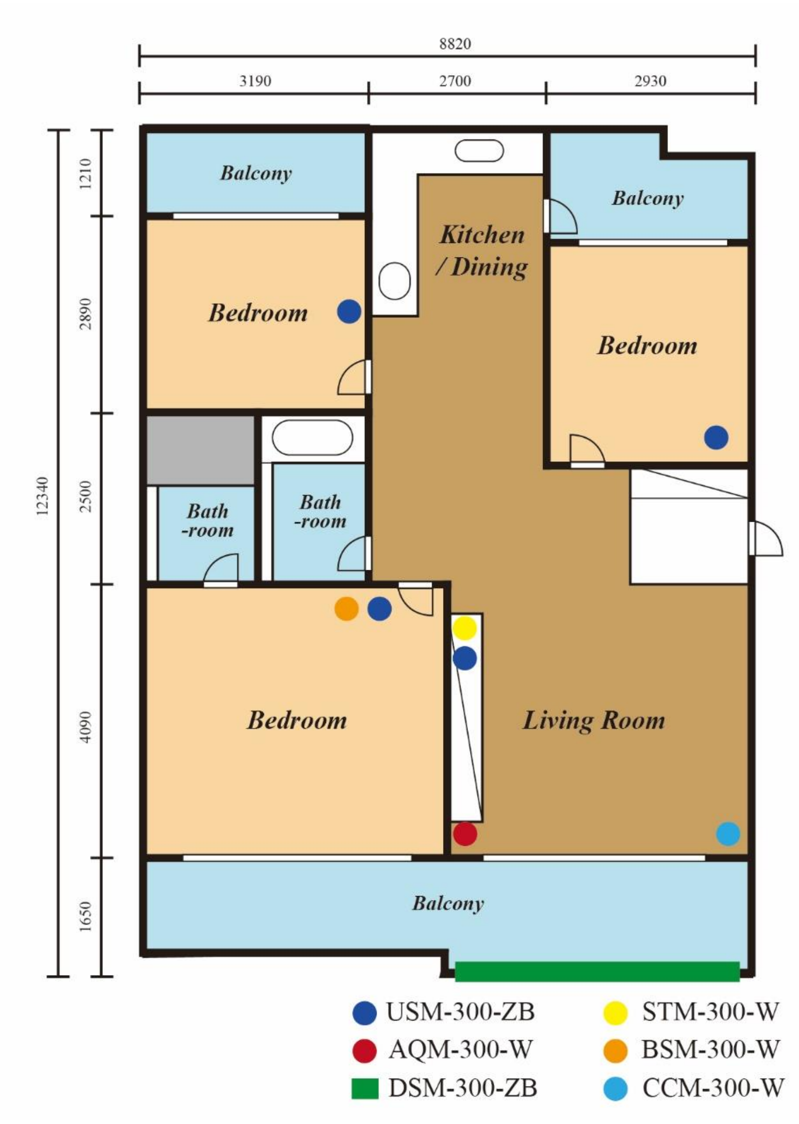
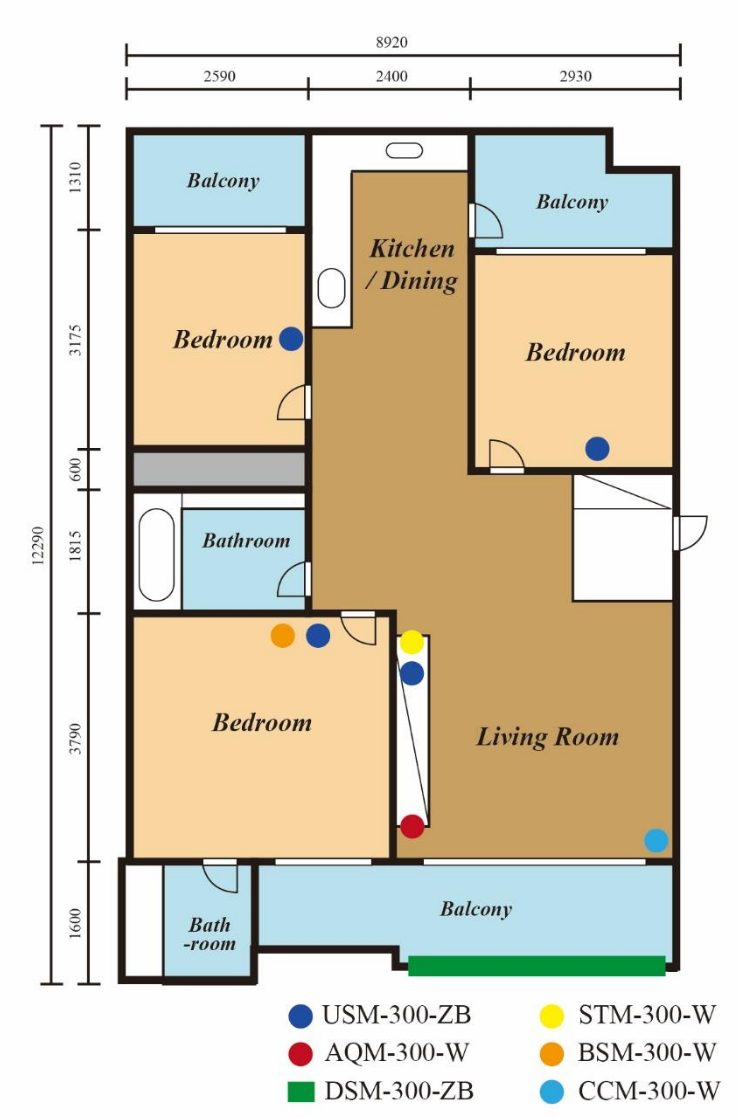


Figure 2.3 Data collection process from installed sensors



(a) Plan of Bldg. A



(b) Plan of Bldg. B

(c) Plan of Bldg. C

Figure 2.4 Building plans and location of each sensor (the unit of measure shown in mm)

Figure 2.5 Sensor installation locations and the experimental space

Table 2.1 Specification of sensors used for measuring occupant and environmental data

Sensor	Variable	Sampling time	Range	Resolution	Accuracy
BCM-300-W	Boiler switch	1 min	0 or 1	-	≥99%
CCM-300-W	AC switch		0 or 1	-	≥99%
DSM-300-ZB	Window state		0 or 1	-	≥99%
STM-300-W	Light switch		0 or 1	-	≥99%
USM-300-ZB	Indoor temperature		0–50 ℃	0.1 ℃	±0.5 ℃
	Indoor humidity		0–100% RH	1% RH	±2% RH
	Indoor illuminance		1–65,528 lx	1 lx	-
	Occupant presence (PIR)		0 or 1	-	Detecting occupant movement ≥99.5%
AQM-300-W	Indoor CO_2 concentration		0–10,000 ppm	1 ppm	±30 ppm
	Indoor $PM_{2.5}$ concentration		0–500 $\mu g/m^3$	1 $\mu g/m^3$	-
Metallic temperature sensor	Outdoor temperature		−40 to 60 ℃	0.1 ℃	±0.3 ℃
Capacitive humidity sensor	Outdoor humidity		0–100% RH	0.1% RH	±3% RH

Table 2.2 The statistical description of the monitored outdoor environmental variables

Environmental variables	**Mean**	**Standard deviation**	**Max**	**Min**
Outdoor temperature (℃)	13.5	11.1	36.1	−15.5
Outdoor relative humidity (%)	64.8	18.4	100.0	15.0

Chapter 3. Predictability quantification of occupant presence

3.1 Introduction

Current modeling approaches utilize rule-based, stochastic, data-driven, or agent-based approaches. The rule-based approach includes but is not limited to, time-dependent user profiles, as defined by Lee and Kim (2017). Stochastic models probabilistically define OB and are the result of multiple contextual factors, such as habitual behaviors and adaptive triggers that evolve over time (Frontczak et al., 2012; Li and Dong, 2017; Altomeonte and Schiavon, 2013; Carlucci et al., 2020). The data-driven approach is described as a black-box model derived from relevant input and output data. Using machine learning (ML) methods, a data-driven model is implemented without in-depth domain knowledge or an understanding of OB (Carlucci et al., 2020; Hong et al., 2017; Brager et al., 2004). Finally, the agent-based approach models individual “agent” behavior. While other approaches assume an “average occupant” in the space of multiple persons, agent-based models aim to describe the interaction between each occupant (Robinson et al., 2011).

These approaches are based on the hypothesis that sufficient data and knowledge can provide reliable prediction models. In other words, most of the previous OB studies have been conducted based on the premise that OB can be predictable. In contrast, studies based on the random walk approach (Ahn and Park, 2016; Ahn et al., 2017; Ahn and Park, 2019; Kim and Park, 2022) reported that in certain types of buildings/spaces, occupant behavior follows a “random walk” pattern, which is difficult to predict. The random walk hypothesis was utilized to investigate the predictability of the time-series data, with the degree of randomness being determined by the normalized cumulative periodogram (NCP) and Bartlett’s test. In this thesis, *predictability* means quantifying the possibility of whether the next state of occupancy can be predicted from the present and past states of occupancy.

In Ahn and Park (2016), the authors observed the occupancy and behavior in a university laboratory occupied by seven people. It was shown that occupancy in the university laboratory was random, and the variance of their behavior had no particular frequency. In Ahn et al. (2017), the predictability of occupancy in laboratories and reading rooms was investigated. It was found that it is difficult to apply a stochastic occupancy model in random walk-driven buildings and can result in a significant performance gap. In Ahn and Park (2019), occupancy data were observed in six rooms of a university library building for 16 days. It was investigated whether temporal and spatial resolutions influence the predictability of occupancy. In addition, it was shown that the number of occupants dominantly drove such predictability.

In previous studies (Ahn and Park, 2016; Ahn et al., 2017; Ahn and Park, 2019), the authors analyzed predictability using the random walk approach focused on university buildings, which have been classified as random walk-driven building types (Ahn and Park, 2016). In contrast, Ahn and Park (2016) hypothesized that residential buildings would have process-driven occupancy patterns that could be easily predicted.

In this chapter, the author delves into the concept of temporal diversity by employing the random walk theory to quantify the consistency of occupant behavior through the analysis of autocorrelation in time-series data. The primary focus is investigating the predictability, or autocorrelation, of occupant presence to demonstrate and quantify temporal diversity.

The main objectives of this chapter are as follows: (1) examining whether occupant presence in specific households within residential buildings follows a random walk pattern, which implies unpredictability; (2) quantifying the predictability of occupant presence at different temporal and spatial resolutions; and (3) assessing the degree of variation in predictability across households.

To achieve these objectives, collected occupant presence data for 147 days from 31 households in residential apartment buildings in Seoul, South Korea, were used. The author adopted methodologies utilized in previous studies, such as the random-walk hypothesis, Normalized Cumulative Periodogram (NCP), and Bartlett's test (Ahn and Park, 2016; Ahn et al., 2017; Ahn and Park, 2019) to analyze the data.

By investigating the predictability of occupant presence in real-life scenarios, this chapter aims to provide insights into the temporal variations and patterns of occupant behavior, thus establishing empirical evidence for temporal diversity (Kim and Park, 2022).

3.2 Random walk approach

This section presents the methodology of predictability analysis depicted in Figure 3.1. In this chapter, it was hypothesized that the predictability of occupant presence could be influenced by temporal resolution, spatial resolution, and the length of measurement periods. Thus, the author quantified the predictability of occupant presence in the residential buildings in terms of temporal resolution (from one minute to 120 minutes), measurement periods (from one day to 147 days), day types (weekdays, weekends), spatial resolution (from one household to 31 households). The predictability was analyzed with the two popular tests: NCP and Bartlett's test.

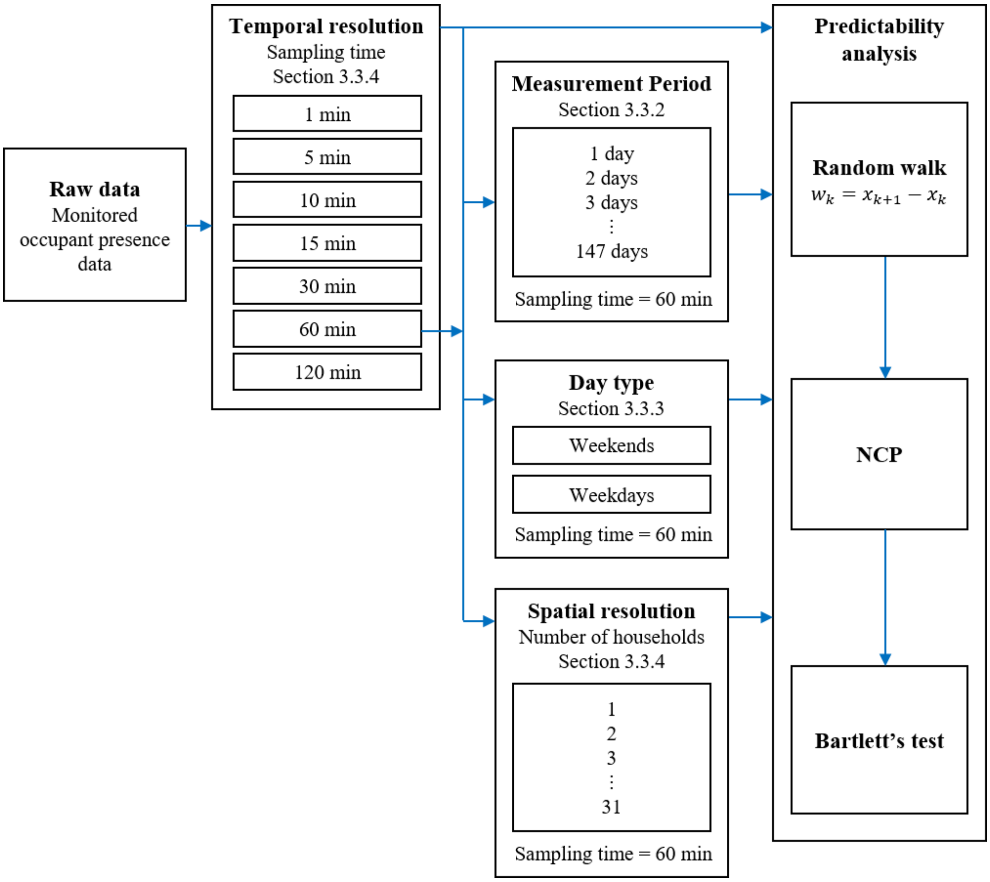


Figure 3.1 Predictability analysis process (Kim and Park, 2022)

Thirty-one households in three residential apartment buildings in Seoul, South Korea, were selected (Figure 3.2). Occupancy data (0: absence, 1: presence) measured for 147 days (2021.08.03-2021.12.27) were used.

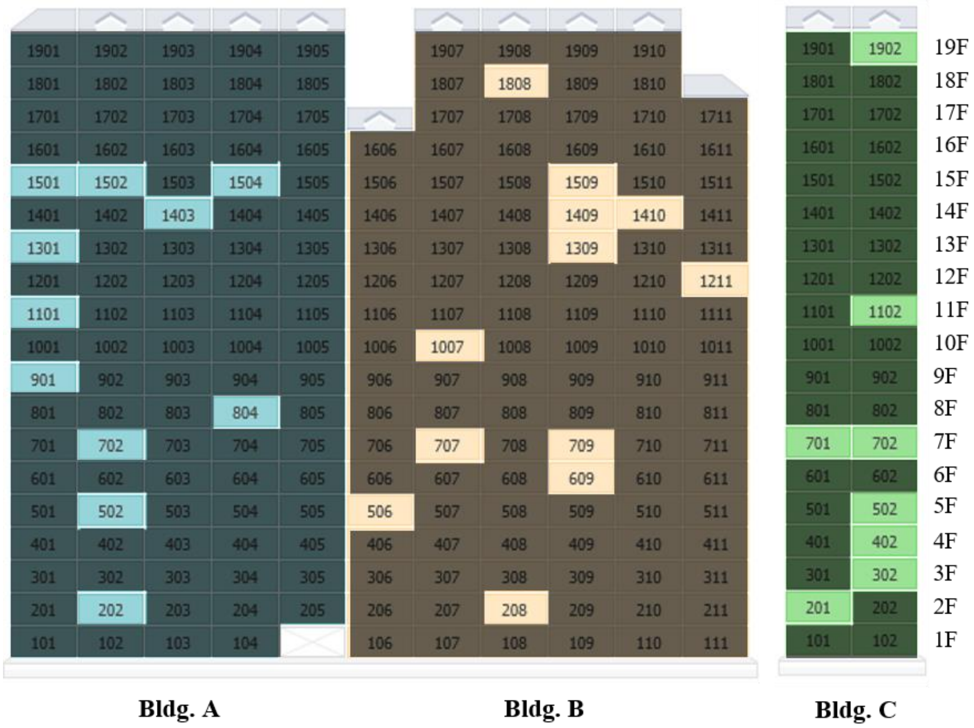


Figure 3.2 The 31 households selected

3.2.1 Mathematical form

A random walk is the mathematical formalization of a path consisting of a succession of random steps. The term, first introduced by Pearson (1905), has been used in many fields (e.g., ecology, economics, and psychology) to explain the observed behavior of time-series data. Figure 3.3 shows an example of 10 random walk-driven time series in one dimension. The mathematical form of a random walk for time-series data can be expressed as shown in Equations 3.1 and 3.2 (Ahn and

Park, 2019; Kim and Park, 2022).

$$x_{k+1} = x_k + w_k \qquad \text{Equation 3.1}$$

$$w_k = x_{k+1} - x_k \qquad \text{Equation 3.2}$$

where x_k is the state of the kth time step; x_{k+1} is the state of the (k+1)th time step; and w_k is the difference between x_k and x_{k+1}, representing the state fluctuation over time. According to the random walk hypothesis, if the change (w_k) in presence data is a random value with a uniform probability distribution, occupant presence (x_k) is deemed unpredictable (Ahn and Park, 2019).

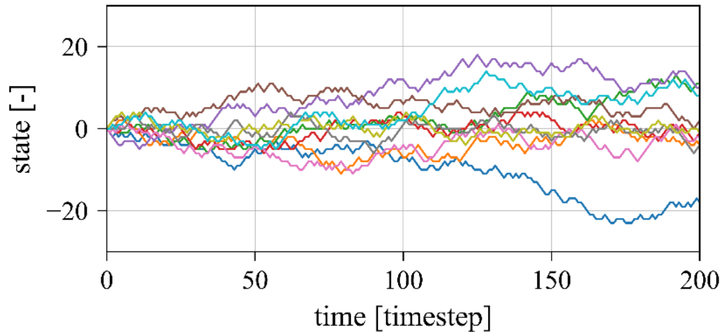


Figure 3.3 Example of ten random walks (Ahn and Park, 2019)

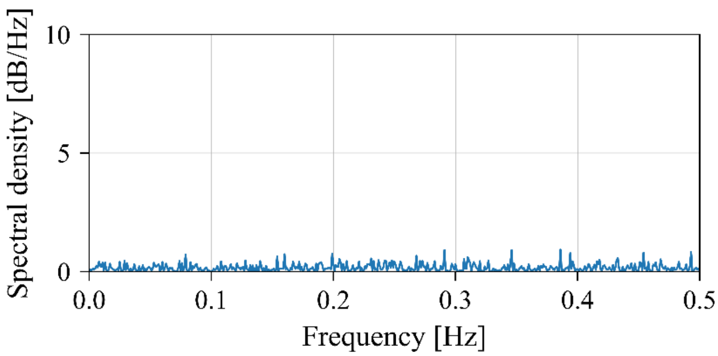
3.2.2 Normalized Cumulative Periodogram (NCP)

The NCP is a common method for identifying a given time series's periodicity (randomness) in the frequency domain (Newton, 1988). For a given n stationary time series (x_k), the periodogram function ($\hat{f}(\omega_j)$), which shows the spectral density of the time series at each frequency, is calculated using Equation 3.3 (Ahn and Park, 2019):

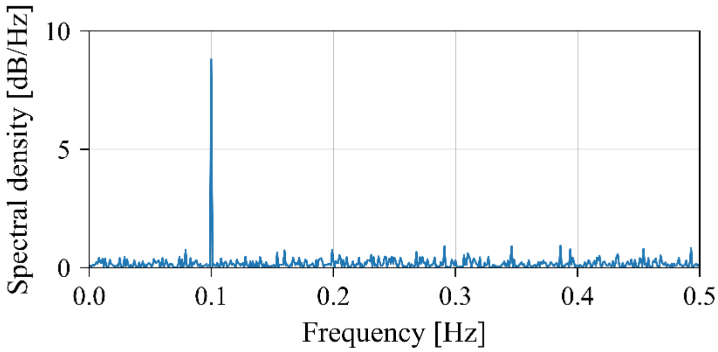
$$\hat{f}(\omega_j) = \frac{1}{n}\left|\sum_{l=1}^{n} x(t)e^{2\pi i(t-1)\omega_j}\right|^2 \qquad \text{Equation 3.3}$$

where $\omega_j = (j-1)/n$ is the jth frequency ($j = 1, \ldots, \mathrm{q}$), n is the length of the time series, $\mathrm{q} = \left[\frac{\mathrm{n}}{2}\right] + 1$, $\hat{f}(\omega_j)$ is the periodicity spectrum at a frequency of ω_j, and $x(t)$ is the time series data at time t. The periodogram provides a graphical representation of the frequency distribution of the time-series data.

Two thousand random numbers were generated by the random module in Python and recorded to represent random time series on the periodogram. The periodograms of the periodic data and random time series are shown in Figure. 3.4(a). The random time series are not concentrated at a few specific frequencies but are uniformly distributed over the entire frequency domain. Therefore, a random time series is understood as white noise or a random signal having equal intensity at different frequencies (Diggle and Fisher, 1991). On the other hand, for the periodic data, because the peak frequency is 0.1 Hz, the representative period is verified as 10 s (Figure 3.4(b)) (Kim and Park, 2022).



(a) Random numbers



(b) Random numbers + 0.1 Hz sine wave

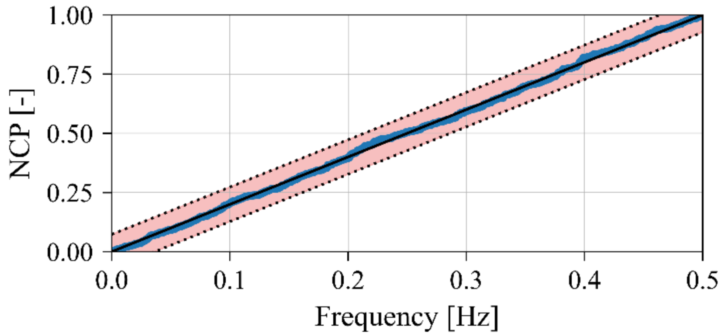
Figure 3.4 Example of periodogram of time series for 1000 s (sampling time: 1 s) (Kim and Park, 2022)

NCP is a cumulative form based on the periodogram $\hat{f}(\omega_j)$ as follows (Newton, 1988):

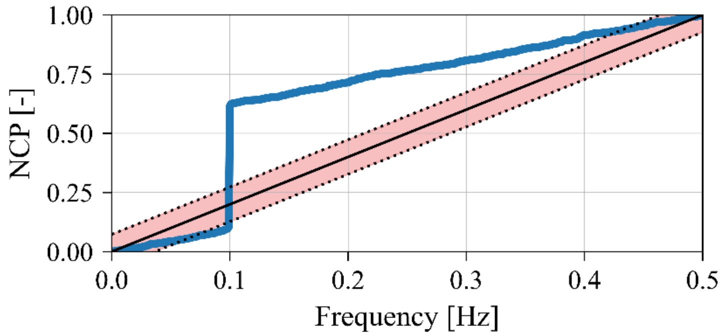
$$\hat{F}(\omega_k) = \frac{\sum_{j=1}^{k} \hat{f}(\omega_j)}{\sum_{j=1}^{q} \hat{f}(\omega_j)}, \qquad \mathrm{k} = 1, \ldots, \mathrm{q} \qquad \text{Equation 3.4}$$

where $\hat{F}(\omega_k)$ is the NCP at the frequency ω_k. Note that $\hat{F}(0) = 0$ and $\hat{F}(\omega_q) = 1$.

Figure 3.5 shows the NCP for a random time series (bold blue line), where the red area indicates the 99% confidence intervals for testing the random walk. Presumably, the time series data follows a random walk if the bold blue line is drawn within a confidence interval with a straight line joining (0, 0) and (0.5, 1) in the NCP (Figure 3.5(a)). The confidence interval lines (dotted lines) are drawn at vertical distances $\pm \frac{K_e}{[\frac{n-1}{2}]}$ above and below the straight line joining (0, 0) and (0.5, 1), where $[\frac{n-1}{2}]$ denotes taking only the integer portion of the number of brackets and K_e, a parameter for determining confidence limits in the cumulative periodogram, is set to 1.63 with a 99% confidence interval (Hipel and McLeod, 1994). The NCP can be used to qualitatively evaluate whether the time-series data have periodicity (predictable) or not (not predictable) (Ahn and Park, 2019).



(a) Random numbers



(b) Random numbers + 0.1 Hz sine wave

Figure 3.5 Example of NCP of time series for 1000 s (sampling time: 1 s) (Ahn and Park, 2019)

3.2.3 Bartlett's test

Bartlett's test (Bartlett, 1967) is a common method for testing the null hypothesis that data are derived from white noise. Bartlett's test statistic B and the p-value were calculated, as shown in Equations 3.5 and 3.6 (Ahn and Park, 2019; Nason and Savchev, 2014).

$$B = \sqrt{q} \max_{1 \le k \le q} \left| \hat{F}(\omega_k) - \frac{k}{q} \right|, \qquad k = 1, \ldots, q \qquad \text{Equation 3.5}$$

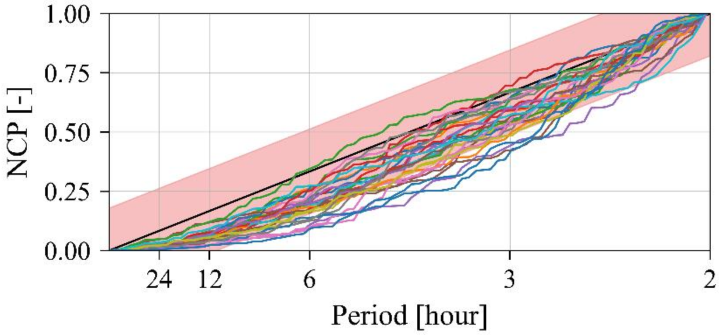
$$p - \text{value} = 1 - \sum_{j=-\infty}^{\infty} (-1)^j e^{-2B^2 j^2} \qquad \text{Equation 3.6}$$

As mentioned above, $\hat{F}(\omega_k)$ is the NCP at the frequency ω_k, where $\omega_k = (k - 1)/n$ is the k^{th} frequency ($k = 1, \ldots, q$), and $q = \left[\frac{n}{2}\right] + 1$. Bartlett's test statistic B is defined as the deviation of $\hat{F}(\omega_k)$ from a straight line in the NCP. The null hypothesis of white noise is rejected if the p-value calculated from Bartlett's test statistic B is less than a specified significance level α (Kim and Park, 2022). In this chapter, the calculation was performed using the Bartlett B. test function in R.

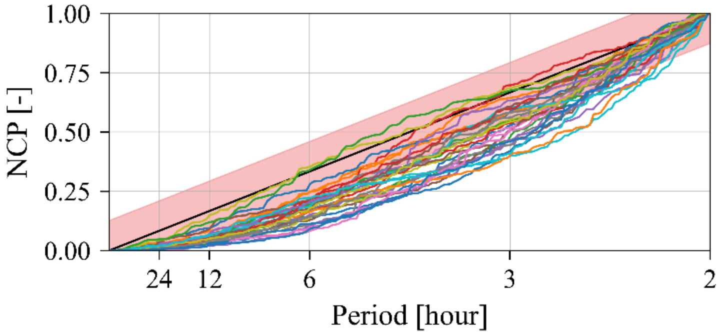
3.3 Results

3.3.1 NCPs of 31 households

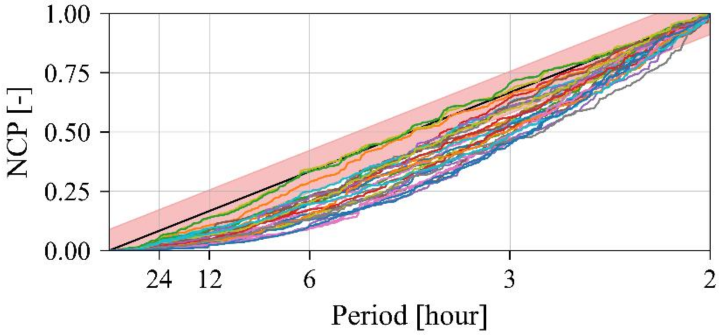
Figure 3.6 shows the NCPs of 31 households. Each line represents the NCP of the occupant presence in each household. If the presence corresponding to a household was located within the red-colored band, it was considered unpredictable at a significance level of 0.01 (Hipel and McLeod, 1994). In the relatively short measurement period (7 days), occupant presence in half of the households proved to be unpredictable. In contrast, in the NCP of 147 days (Figure 3.6(d)), the occupant presence of all households was indicated as predictable. This was substantiated by Bartlett's test (Section 3.3.2). In addition, the difference in the NCPs between households indicates a difference in the predictability of the occupant presence.



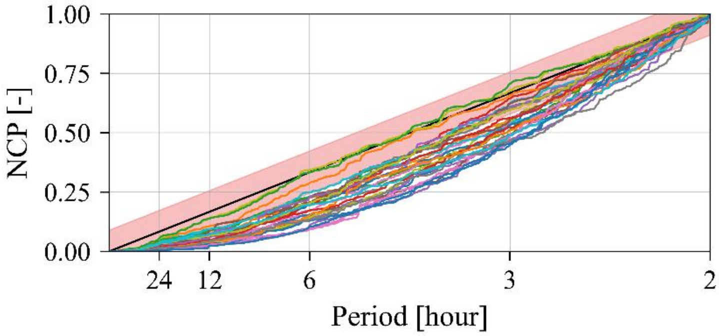
(a) 7 days



(b) 14 days



(c) 28 days



(d) 147 days

Figure 3.6 NCPs of 31 households according to measurement periods (each line represents each household's measured occupant presence) (sampling time = 60 min)

3.3.2 Predictability with varying measurement periods

Table 3.1 shows the results of Bartlett's test for the NCPs in Figure 3.6. By comparing Bartlett's test statistic for each household with the reference value (1.63, significance level of 0.01), it is possible to identify the predictability of occupant presence. Alternatively, the presence is deemed unpredictable when the *p*-value is greater than the significance level of 0.01 (red-colored), indicating a 1% risk of concluding that a difference from white noise exists when there is none (Hipel and McLeod, 1994). Please note that there is no set-in-stone rule or universal rule for determining the significance level. However, the significance level of 0.01 has been widely used, as addressed in (Hipel and McLeod, 1994). Accordingly, 17 households whose presence was unpredictable within 7 days of the measurement period were identified. Approximately five households were identified as unpredictable within 14 days, and three were identified as unpredictable within 28 days. No households were identified as unpredictable within 147 days. For a long measurement period of nearly five months, all households were found to be predictable. In addition, it can be inferred that longer measurement periods increased the predictability of occupant presence.

Figure 3.7 shows the results of the correlation analysis between the presence predictability and measurement periods. Each line represents Bartlett's test results for each household. When Bartlett's test statistic was less than 1.63, the presence of the household was determined to be unpredictable at a significance level of 0.01 (Nason and Savchev, 2014). Bartlett's test statistics tended to be proportional to the measurement period in most households. Even households whose presence was unpredictable could be changed to predictable by extending the measurement period. In general, the results exhibited a monotonic increase. However, for cases of surge/decrease, an alteration in the presence pattern was estimated.

Moreover, the graphs of each household show a significant difference in their gradients (Figure 3.7). The last household (household #23) could be predicted with 86 days of data. While households with high gradients can secure high predictability with relatively short measurement periods, households with low gradients require longer periods to acquire the same level of predictability.

Table 3.1 Bartlett's test statistic and p-value of 31 households according to measurement periods (text in red indicating unpredictable)

Household #	7 days		14 days		28 days		147 days	
	statistic	p-value	statistic	p-value	statistic	p-value	statistic	p-value
1	1.443	0.035	3.284	0.000	4.199	0.000	9.865	0.000
2	1.262	0.100	1.619	0.011	2.584	0.000	8.151	0.000
3	1.080	0.202	2.893	0.000	3.171	0.000	7.845	0.000
4	1.390	0.045	1.821	0.003	2.812	0.000	5.956	0.000
5	1.492	0.029	3.133	0.000	3.992	0.000	8.038	0.000
6	2.226	0.000	3.422	0.000	4.638	0.000	9.433	0.000
7	1.130	0.166	2.071	0.000	3.149	0.000	8.713	0.000
8	2.073	0.000	3.298	0.000	4.056	0.000	10.203	0.000
9	1.852	0.002	2.827	0.000	3.921	0.000	9.822	0.000
10	0.627	0.840	3.829	0.000	2.665	0.000	4.045	0.000
11	1.400	0.047	1.279	0.086	3.137	0.000	10.999	0.000
12	0.627	0.840	1.407	0.039	1.556	0.016	8.423	0.000
13	1.458	0.031	2.030	0.001	2.305	0.000	6.690	0.000
14	1.653	0.009	2.380	0.000	3.575	0.000	7.298	0.000
15	2.840	0.000	3.455	0.000	4.938	0.000	10.374	0.000
16	2.353	0.000	2.546	0.000	2.389	0.000	6.088	0.000
17	2.361	0.000	3.317	0.000	4.671	0.000	10.278	0.000
18	1.550	0.017	2.571	0.000	2.647	0.000	4.222	0.000
19	0.627	0.840	0.680	0.749	0.754	0.624	5.054	0.000
20	1.691	0.009	2.396	0.000	1.900	0.002	9.513	0.000
21	2.921	0.000	3.774	0.000	4.764	0.000	8.279	0.000
22	1.949	0.001	3.670	0.000	3.474	0.000	7.714	0.000
23	0.627	0.840	0.858	0.463	0.842	0.486	2.525	0.000
24	1.227	0.103	2.054	0.000	2.740	0.000	8.300	0.000
25	1.532	0.020	2.148	0.000	2.457	0.000	4.400	0.000
26	1.608	0.013	2.105	0.000	2.464	0.000	7.702	0.000
27	2.636	0.000	3.635	0.000	4.993	0.000	11.185	0.000
28	1.949	0.001	2.718	0.000	4.139	0.000	10.039	0.000
29	1.870	0.002	2.281	0.000	2.568	0.000	8.713	0.000
30	1.311	0.075	3.459	0.000	3.721	0.000	5.996	0.000
31	2.559	0.000	3.468	0.000	4.952	0.000	10.066	0.000

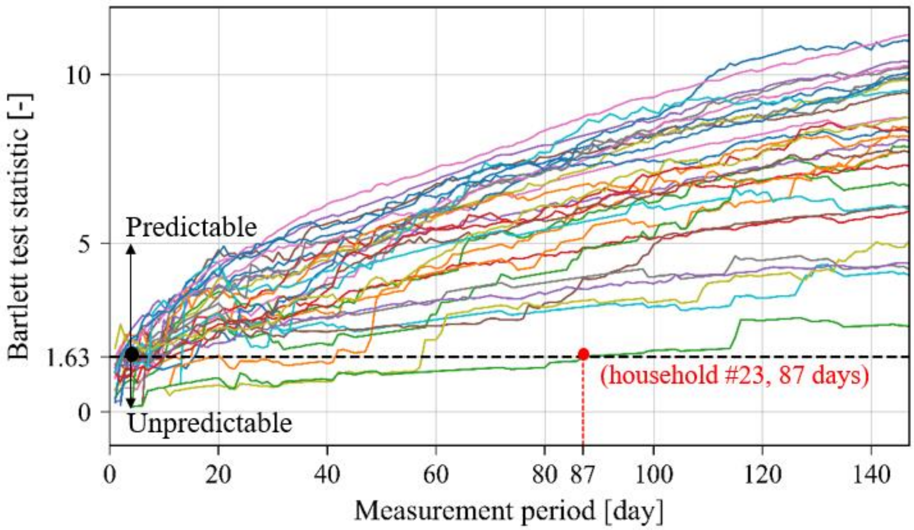


Figure 3.7 Predictability of occupant presence with varying measurement periods (each line represents each household's measured occupant presence) (sampling time = 60 min)

3.3.3 Predictability between weekdays/weekends

Figure 3.8 compares each household's predictability (Bartlett's test statistic) on weekdays and weekends (including holidays). For a fair comparison, both data sets were analyzed for the same number of days (30 days). Both cases were almost predictable, with a significance level of 0.01. On weekdays, the occupant presence of four households (#10, #12, #19, #23) was proven to be unpredictable. One household was unpredictable on weekends and holidays (household #23). The predictability between weekdays and weekends is similar in most households, but there are significant differences between several households. For households #3, #10, #12, #20, #24, and #30, occupant presence patterns differed between weekdays and weekends. In other words, it is essential to develop a separate presence-prediction model.

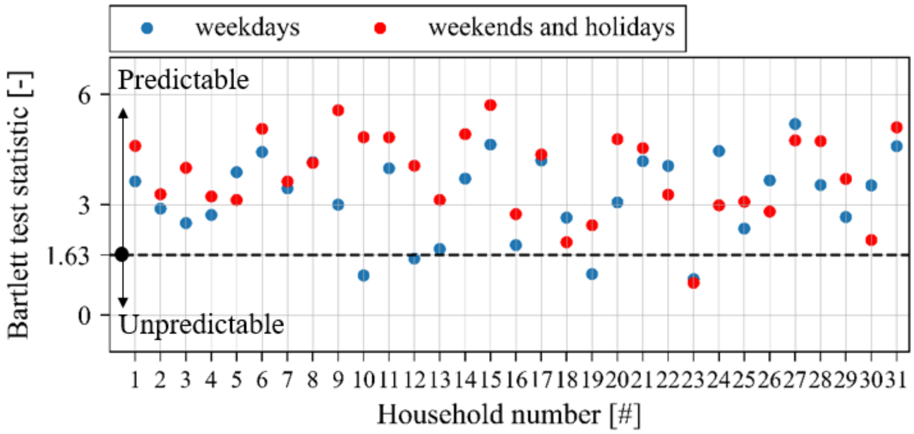
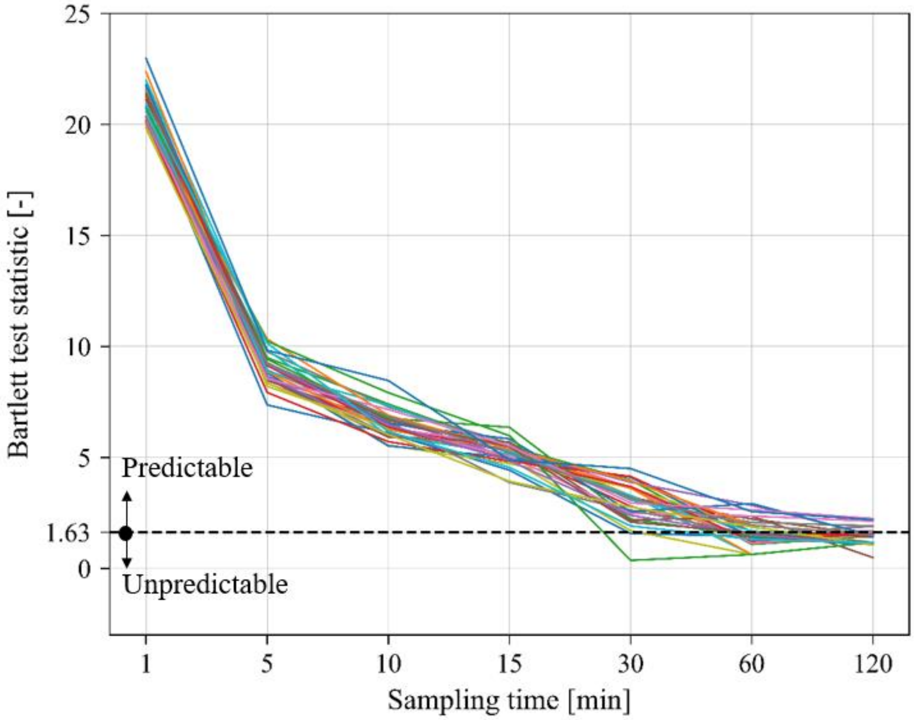


Figure 3.8. Predictability of occupant presence on weekdays/weekends for 30 days (sampling time = 60 min)

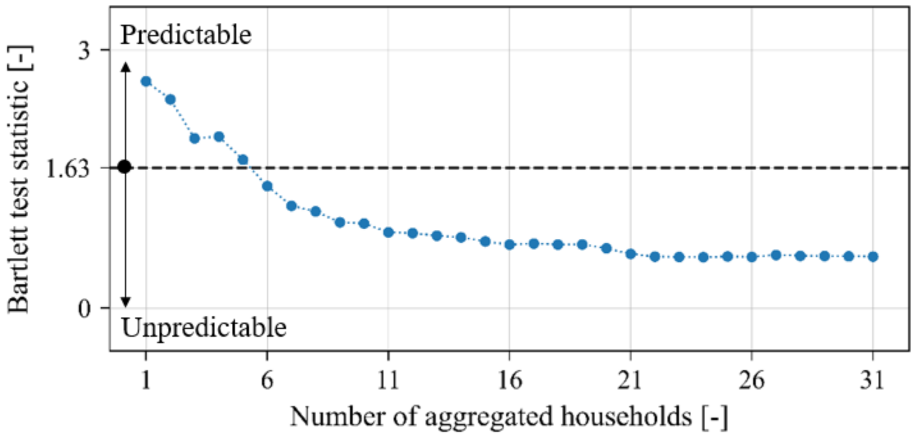
3.3.4 Predictability with varying temporal and spatial resolutions

Figure 3.9 shows Bartlett's test results with varying sampling times and numbers of households. As shown in Figure 3.9(a), the shorter the sampling time, the greater the predictability of occupant presence. Figure 3.9(b) shows the changes in predictability as the number of households aggregates. Python's random module was used in the households' aggregation process. In other words, no special grouping was applied, e.g., in terms of buildings (A, B, C) or plan type (A, B, C) (Figure 2.4). In addition, 7 days of the presence data of selected multiple households were summed. Presence data with a value of 0 or 1 have a value between 0 and N (the number of aggregated households) after aggregation.

Figure 3.9(b) suggests that predicting the occupant presence of multiple (aggregated) households is more difficult than individual households. In other words, it may be easier to predict occupant presence in small spaces, for example, several rooms or a single floor, than to predict the occupant presence of the whole building.



(a) Predictability of occupant presence in terms of sampling time (each line represents the measured occupant presence of each household)



(b) Predictability of occupant presence in terms of the number of aggregated households (sampling time = 60 min)

Figure 3.9 Predictability of occupant presence with varying temporal and spatial resolutions (7 days' measured data)

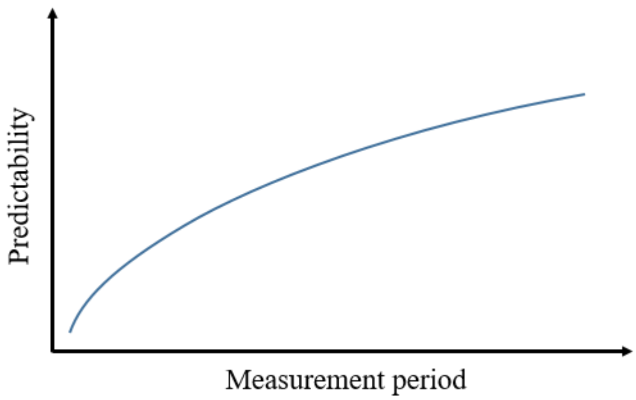
3.4 Summary

This chapter investigates the predictability of occupant presence in residential buildings regarding temporal diversity. Presence data from 31 households were collected for 147 days, and NCP analysis and Bartlett's test were used to examine the predictability of occupant presence.

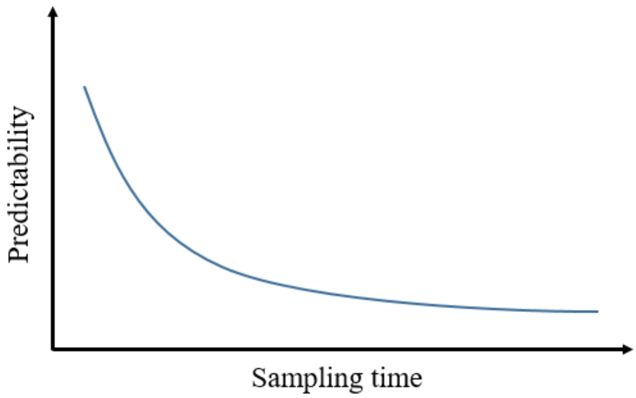
The findings are summarized graphically in Figure 3.10. Data features, such as the measurement period and temporal/spatial resolution, significantly influence the predictability of occupant behavior. Three analyses regarding the variation in the predictability of occupant presence are presented as follows: (1) with different measurement periods (Figure 3.10(a)), (2) with different sampling times (Figure 3.10(b)), and (3) individual vs. aggregated households (Figure 3.10(c)).

The measurement period significantly influenced the predictability of the occupant presence. In general, the longer the presence data are collected, the higher the predictability. Therefore, securing a sufficient measurement period is recommended to predict occupant presence better. Notably, the degree of predictability increase varies according to occupant characteristics. Second, it was found that the shorter the sampling time is, the greater the predictability of the occupant presence. Finally, predicting the occupant presence of multiple or aggregated households would be much more difficult than that of an individual household. Additionally, the predictability between weekdays and weekends is similar in most households but differs in multiple households. Thus, developing a separate presence prediction model for weekdays and weekends is essential.

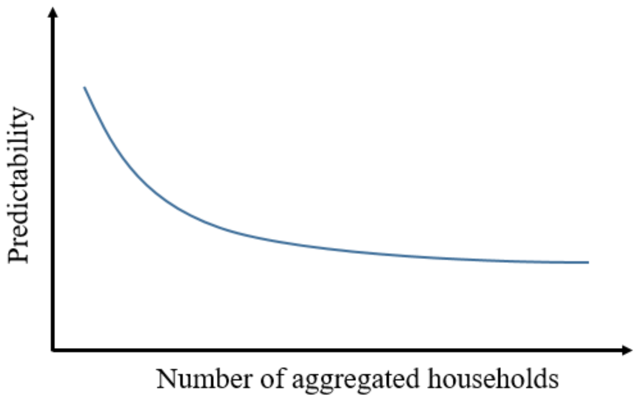
In summary, the occupant presence of specific households followed a random walk pattern for short measurement periods, which means there is significant temporal diversity of occupant presence. The predictability of occupant presence differs significantly between households, and the degree of variation is shown in Figure 3.7. Notably, the predictability of occupant presence in residential buildings cannot be defined as a single state and varies widely according to occupant features (e.g., number of family members and occupancy patterns). Therefore, to implement a reliable OB model, it is necessary to check whether it applies to each occupant and building/space. In other words, the findings of this chapter are limited by the information of the householders (Kim and Park, 2022).



(a) Measurement period vs. predictability of occupant presence



(b) Sampling time vs. predictability of occupant presence



(c) Number of aggregated households versus predictability of occupant presence

Figure 3.10 Variation of predictability of occupant presence with varying data features (Kim and Park, 2022)

Chapter 4. Feature influence quantification in window adjustment behavior using XAI

4.1 Introduction

Window adjustment is one of the most common ways employed by occupants to control the indoor environment, and window adjustment behavior (WAB) is known to be a crucial factor for predicting building energy consumption. WAB in a dynamic manner is triggered by various influencing factors, such as indoor/outdoor temperature, CO_2 concentration, and time of day (Plieninger et al., 2016; Wei el al., 2014; Hong et al., 2015; Stazi et al., 2017; Fabi et al., 2012). In recent decades, many attempts have been made to develop reliable WAB models by finding correlations between environmental and non-environmental factors and WAB. For instance, Andersen et al. (2013) developed a WAB model for Danish dwellings and proposed four models of the window opening and closing behavior patterns based on measured environmental data. A probabilistic approach using logistic regression was applied. Cali et al. (2016) investigated the time of day as the most common driver to open a window by comparing German households using logistic regression. An artificial neural network (ANN) model with higher accuracy than traditional stochastic approaches was proposed by Wei et al. (2019). Better interpretability of influencing factors was also demonstrated compared to the logistic regression and Markov models. Zhou et al. (2021) proposed other machine-learning models. The random forest algorithm was compared with two other machine learning models: support vector machine (SVM) and extreme gradient boost (XGBoost) algorithms. For stochastic modeling, a Bayesian network was suggested with its applicability to capture the complicated underlying relationships between various influencing factors and WAB (Barthelmes et al., 2017).

Similarly, a large number of WAB models have been developed based on occupant responses to environmental and non-environmental factors. However, crucial topics still need to be considered to achieve reliable WAB modeling, but many studies ignore them (Liu et al., 2022; Kim and Park, 2023). This chapter proposes a novel approach to WAB modeling to address the following three issues.

First, most of the studies have ignored the variability of individual preferences (spatial diversity) and treated it in an “average occupant” fashion. It was reported that the average occupant approach is detrimental to understanding the differences in people’s behaviors and can result in a performance gap between the actual and predicted building energy consumption (Liu et al., 2022; Liu et al., 2022). Customized models that reflect individual preferences can be applied as a solution rather than a universal WAB model. On the other hand, several studies have focused on behavioral diversity and attempted to characterize occupant WAB patterns (Haldi and Robinson, 2009; Yun et al., 2009; D’Oca and Hong, 2014). Haldi and Robinson (2009) classified their sample into “active” and “passive” types based on the proportion of window opening time. D’Oca and Hong (2014) clustered patterns of WAB in 16 offices along four dimensions (motivation, opening duration, interactivity, and position) based on association rule mining techniques. The studies mentioned above provide an initial understanding to explain the spatial diversity of WAB better. However, further evaluation is necessary if it is possible to apply it to external data and define specific criteria (Kim and Park, 2023).

The second issue is analyzing the impact of each factor, which is a fundamental consideration for making the prediction model more realistic. Many studies have analyzed the feature impact based on the interaction between environmental factors and WAB, but they have different opinions on influencing factors. For instance, in (Liu et al., 2022), the outdoor temperature was considered the most dominant factor for WAB. However, other factors, such as indoor temperature (Anderson et al., 2013; Yun and Steemers, 2008; Yun and Steemers, 2010), humidity (Sun et al., 2018), outdoor $PM_{2.5}$ concentration (Gu et al., 2021), and time of day (Cali et al., 2016), were also shown to have significant influences in some instances. Therefore, WAB cannot be described by only one specific environmental factor because it is a response to the interaction of multiple factors. Moreover, occupants behave differently even in the same environment, depending on their individual perspectives of the environment and energy demand. Previous studies quantified randomness as the influence of non-environmental factors such as occupancy, time of day, building characteristics, and personal preferences (Stazi et al., 2017; Fabi et al., 2012; Pan et al., 2018). Therefore, it is worth discussing how to reveal different responses of occupants to multiple factors (including unknown factors), select the appropriate factors for different cases, and improve the model's reliability (Kim and Park, 2023).

Finally, diverse modeling approaches have been used to calculate the correlation between WAB and its influencing factors accurately. In recent decades, studies have modeled the state probability of WAB rather than using fixed schedules. Logistic regression is the most widely used model for predicting the probability of a window state (Liu et al., 2022). Nicol (2001) first presented a coherent probability distribution for the predicted window state as logit functions of outdoor and indoor temperatures. Logistic regression was also used to quantify the feature impact with the correlation coefficients of its formula and categorize the patterns of occupant responses (Anderson et al., 2013; Cali et al., 2016; Pan et al., 2018). Several machine-learning algorithms have also been introduced to obtain a model with a relatively higher performance than logistic regression models (Wei et al., 2019; Zhou et al., 2021; Mo et al., 2019; Pan et al., 2019; Han et al., 2020; Niu et al., 2022; Park et al., 2021). Park et al. (2021) compared six machine learning algorithms (KNN, RF, ANN, CART, CHAID, and SVM) with a logistic regression model to predict the window state. In addition to machine learning algorithms, deep learning algorithms have been proposed for WAB (Markovic et al., 2018; Markovic et al., 2019). However, such data-driven models depend highly on their datasets, and their applicability to external data is unknown. In the case of black-box models, the influence of each factor cannot be identified, and sufficient explanation, such as feature selection, is not provided for model validation. Therefore, the current data-driven approaches must overcome the lack of explainability and reliability. In summary, the aforementioned modeling approaches have their advantages, but their disadvantages are also obvious (Kim and Park, 2023).

To address the aforementioned three issues, this chapter focuses on explainable artificial intelligence (XAI), which adds explainability to existing machine learning models. In terms of feature influence, the XAI technique explains how each variable affects the prediction results of the model. Consequently, it quantifies the individual's perception and the spatial diversity that causes the occupant responses. This chapter applies the XAI to develop a reliable WAB model that considers individual differences. The three main objectives are as follows: (1) to quantify the diversity of preferences of individual occupants concerning WAB; (2) to reveal relevant information on occupant perceptions and behavioral patterns regarding indoor environment control; and (3) to present a practical approach for developing a reliable WAB model based on multiple influencing factors.

The author used occupant data (occupant presence and window state) and environmental data (temperature, humidity, CO_2 concentration, $PM_{2.5}$ concentration, illuminance, and time of day) from 12 households for one year. Three XAI approaches, logistic regression, XGBoost classifier, and Shapley additive explanations (SHAP), were introduced to analyze the interaction between WAB influencing factors and occupant behavioral patterns.

4.2 Feature influence analysis

This section introduces three XAI approaches, namely, logistic regression, XGBoost classifier, and SHAP, to analyze the interaction between environmental factors and occupant behavioral patterns (Figure 4.1).

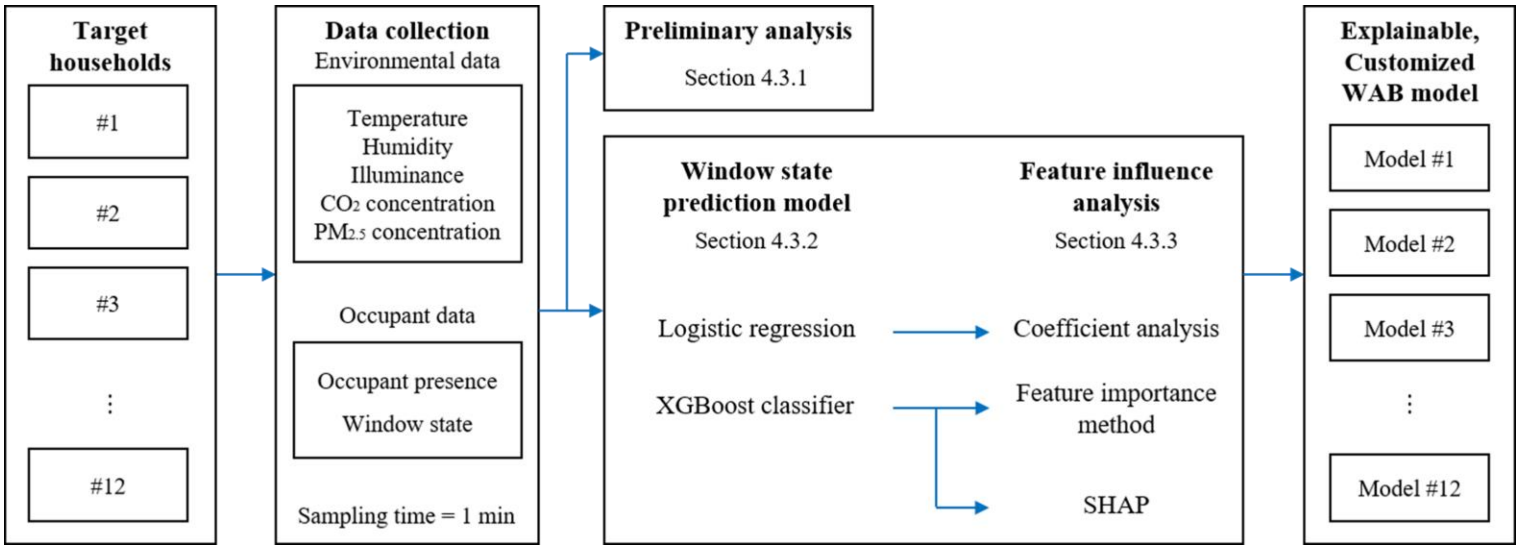


Figure 4.1 Feature influence analysis process (Kim and Park, 2023)

Those 12 households were selected for the experiment (Figure 4.2). Occupant data (occupant presence and window state) and environmental data (indoor/outdoor temperature, indoor/outdoor humidity, indoor CO2 concentration, indoor PM2.5 concentration, indoor illuminance, and time of day) from 12 households for one year were used.

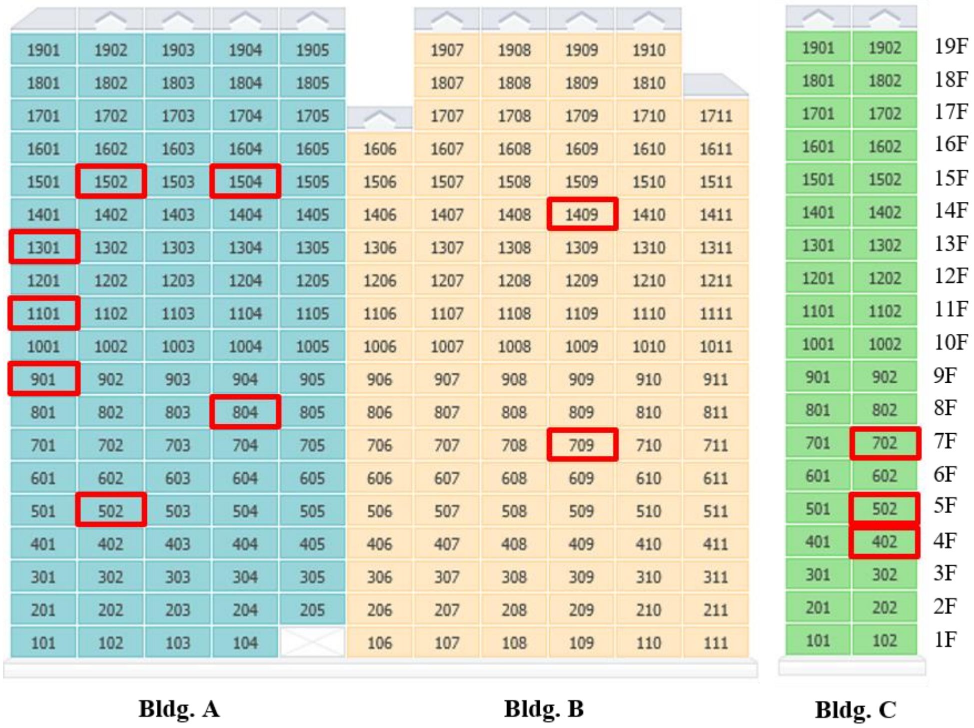


Figure 4.2 The 12 households selected

4.2.1 XAI

The term XAI was first defined by Van Lent et al. (2004), who pointed out that while computing systems are becoming more complex, their self-explanatory functions are not evolving. The evaluation functions of existing machine learning algorithms only present generalized results for the entire dataset and do not provide intuitive evidence of how the model can be improved. However, XAI interprets how the model accepts the data after implementation. Therefore, XAI can present various reasonable perspectives by providing new information about the existing models, as indicated in Figure 4.3. Table 4.1 provides examples of renowned XAI algorithms. Each shows the relationship between the input and output data in different ways. In particular, XAI, in the case of modeling occupant behavior, can provide a better understanding of the interaction between the environment and occupants by quantifying the impacts of influencing factors (Kim and Park, 2023).

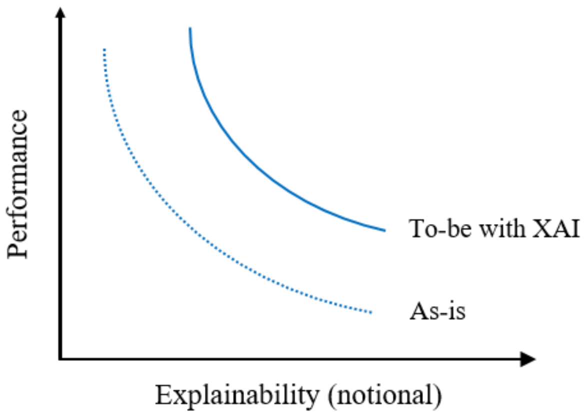


Figure 4.3 Model performance versus explainability (modified from Gunning (2017))

Table 4.1 Renowned XAI algorithms and their characteristics (Gorissen et al., 2009)

Algorithm	Linearity	Monotone	Objectives
Linear regression	O	O	Regression
Logistic regression		O	Classification
Decision tree		Partial	Classification, regression
Naïve Bayes			Classification
K-nearest neighbor (KNN)			Classification, regression

Outdoor/indoor temperature, outdoor/indoor humidity, indoor illuminance, indoor CO_2 concentration, indoor $PM_{2.5}$ concentration, and occupant presence were measured as influencing factors of WAB for 12 households. The influence of each feature on the WAB of each household was analyzed. In addition, two window state models were implemented using the logistic regression and XGBoost classifier.

4.2.2 Logistic regression

Logistic regression (Hastie, 2017) is the most popular stochastic method for analyzing and modeling binary variables (e.g., the state of a window and window-opening/closing behavior). Logistic regression is based on a sigmoid function (Equation 4.1), normalizing multiple parameters into a binary result as 0 or 1. Using the definition of logistic regression, Equation 4.1 can be transformed into Equation 4.2.

$$p = \frac{1}{1 + e^{-(\alpha+\beta x)}} \qquad \text{Equation 4.1}$$

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \qquad \text{Equation 4.2}$$

where

p is the probability that the state of the window is open (0: closed, 1: open)

α is the intercept

β_i are coefficients

x_i are explanatory variables (e.g., indoor/outdoor temperature and occupant presence)

n is the number of explanatory variables

When p is greater than 0.5, the window state is assumed to be open; otherwise, it is assumed to be closed. The scales of each explanatory variable were normalized in the range of 0 to 1 using a min-max scaler. The intercept and coefficients were then estimated using the maximum likelihood method. The magnitude of each coefficient represents the impact of determining the window state of each feature. Therefore, logistic regression effectively reveals the correlation between the environmental factors and WAB and provides quantitative evidence (Wei et al., 2019).

4.2.3 XGBoost classifier

XGBoost is a scalable machine-learning system for tree boosting (Chen and Guestrin, 2016). It has been widely used in many applications, such as window behavior modeling (Mo et al., 2019), prediction of building cooling loads (Fan et al., 2017), and fault detection for HVAC systems (Chakraborty and Elzarka, 2019). The boosting method predicts accuracy by integrating the predictions of weak classifiers into a robust classifier via a serial training process. The computational process of XGBoost is given by Equation 4.3.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \qquad \text{Equation 4.3}$$

where $\hat{y}_i$ is the final tree model, $f_k(x_i)$ are weak classifiers (base tree models) organizing model F, x_i is the ith dataset, and K is the total number of weak classifiers. The objective function of model F is expressed as Equation 4.4.

$$\mathrm{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad \text{Equation 4.4}$$

where $l(y_i, \hat{y}_i)$ is a loss function that calculates errors between the actual (y_i) and predicted $(\hat{y}_i)$ values, and $\Omega(f_k)$ is a complexity function that controls the weights of each tree (f_k). The parameters of the XGBoost model are specified in Table 2. XGBoost is a type of tree algorithm; hence, it supports the feature importance method, a technique for analyzing the degree of each data feature's effect on the exact classification of algorithms. Feature importance determines the variable with the most significant influence on the prediction by permuting variables. This is a powerful method for measuring the feature influence; however, the importance may vary due to the degree of permutation and error-based estimation limitations. In addition, feature importance neglects the dependence between features (Kim and Park, 2023).

Table 4.2 Parameters of implemented XGBoost model

Parameter	Value
booster	gbtree
max depth	9
gamma	0
objective	binary: logistic

4.2.4 Shapley additive explanations

SHAP, first introduced by Lundberg et al. (2018), uses the Shapley value (Shapley, 1953) to numerically express the contribution of each feature to predict the overall results. The formula for the Shapley value is given by Equation 4.5.

$$\phi_i(v) = \sum_{S \in N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left(v(S \cup \{i\}) - v(S) \right) \qquad \text{Equation 4.5}$$

where

ϕ_i is the Shapley value of the ith feature

n is the number of features

S is the subset of the entire set (N), excluding the ith feature

$v(S)$ is the contribution of the subset S

The Shapley value is calculated for each time step. The average Shapley value during the entire period is calculated in calculating the feature influence. Notably, SHAP does not form a model by itself but decomposes the output of the existing model into contributions of each feature (Figure 4.4). In this chapter, the XGBoost classifier model was used as the target model for the feature influence analysis. The Shapley value can be negative, meaning that the specific feature negatively affects the prediction (Kim and Park, 2023).

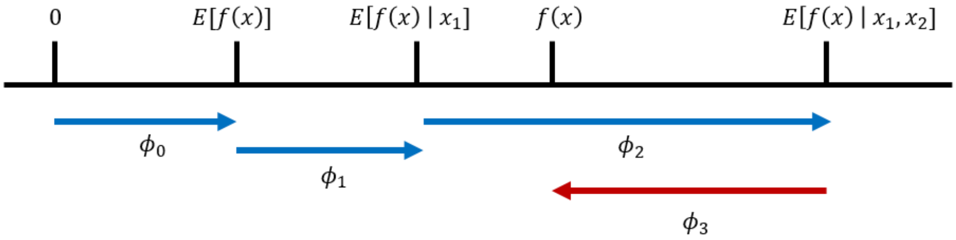


Figure 4.4 Decomposition of the result of model f by features (ϕ_i) (Lundberg et al., 2018)

4.3 Results

4.3.1 Analysis of measured window state by households

This section presents an overview of the measured window data and simple visualizations of the correlation with environmental data. Figure 4.5 displays the proportion of windows in the open or closed state for the 12 households. The household with the highest open ratio (household #6) kept the window open for 93% of the year, whereas the lowest (household #2) kept the window open for only 1%. There is a significant variation in the proportion between households relative to the perception of residents on indoor environment control and energy saving.

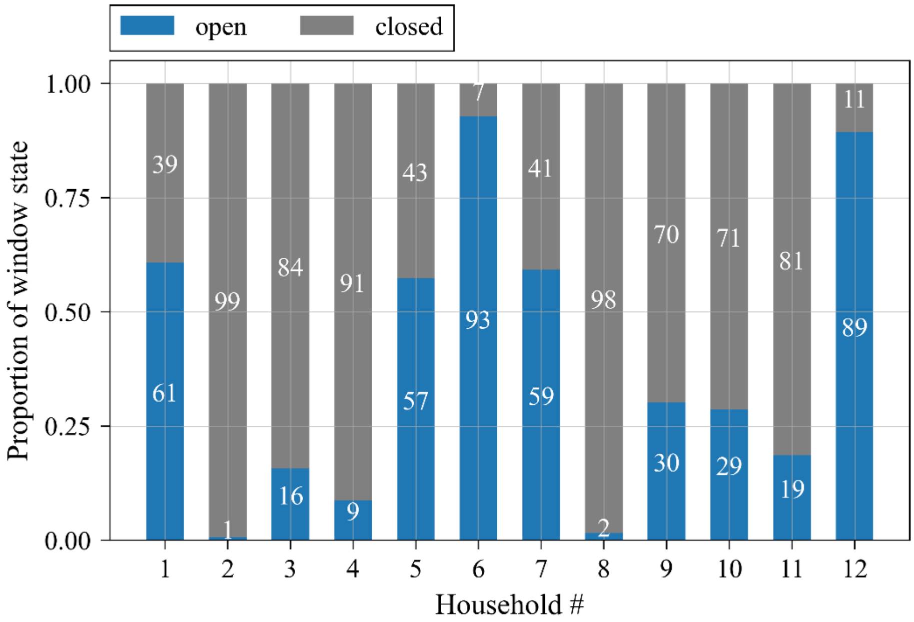


Figure 4.5 Proportion of windows in open/closed state of the 12 households for one year

Figure 4.6 illustrates the correlation between the probability of windows in the open state and outdoor temperature. In most households, the probability varies significantly depending on the outdoor temperature. The probability increases as the outdoor temperature increases from −10 ℃ to 30 ℃. However, at high temperatures over 30 ℃, the residents in several households tended to close their windows (households #2, #4, #8, and #11). In addition, all households kept their windows closed at low outdoor temperatures below −10 ℃.

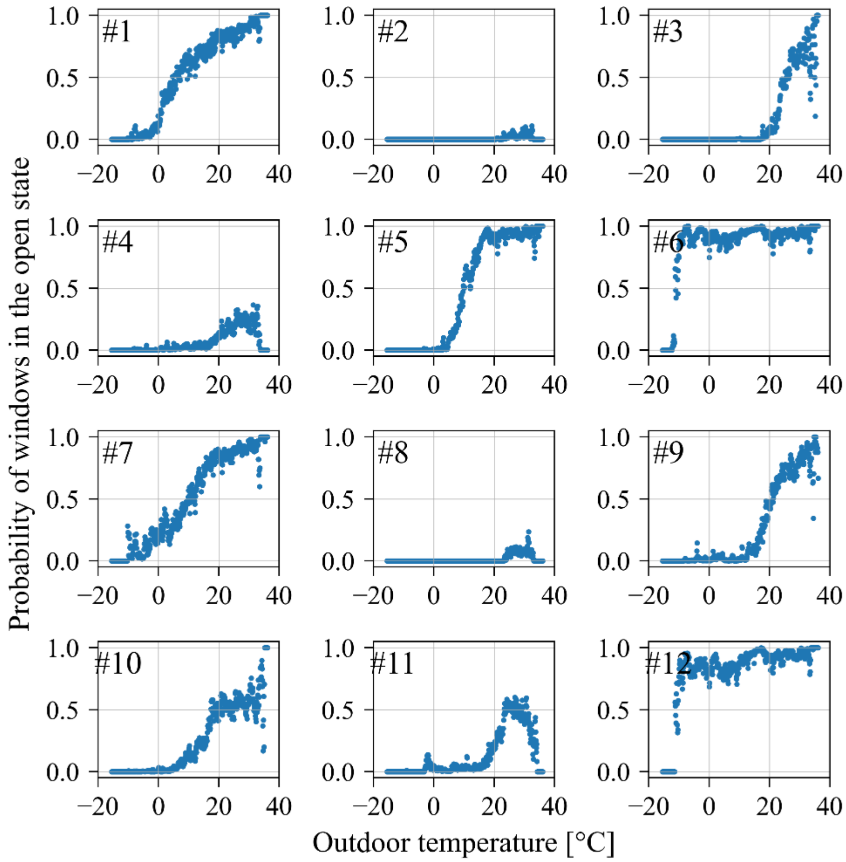


Figure 4.6 Outdoor temperature versus probability of windows in the open state of the 12 households for one year

Figure 4.7 depicts the correlation between the probability of windows in the open state and indoor temperature. Compared to the outdoor temperature, the tendency between households differs considerably. The probability monotonically increases as the room temperature increases in four households (households #3, #5, #10, and #11). Other households tend to decrease and then increase at a specific temperature. For this reason, different from the outdoor temperature, the indoor temperature directly and immediately affects the thermal comfort of the occupants and is changed by the opening and closing of windows. However, the probability varies significantly depending on the indoor temperature, influencing the window state prediction.

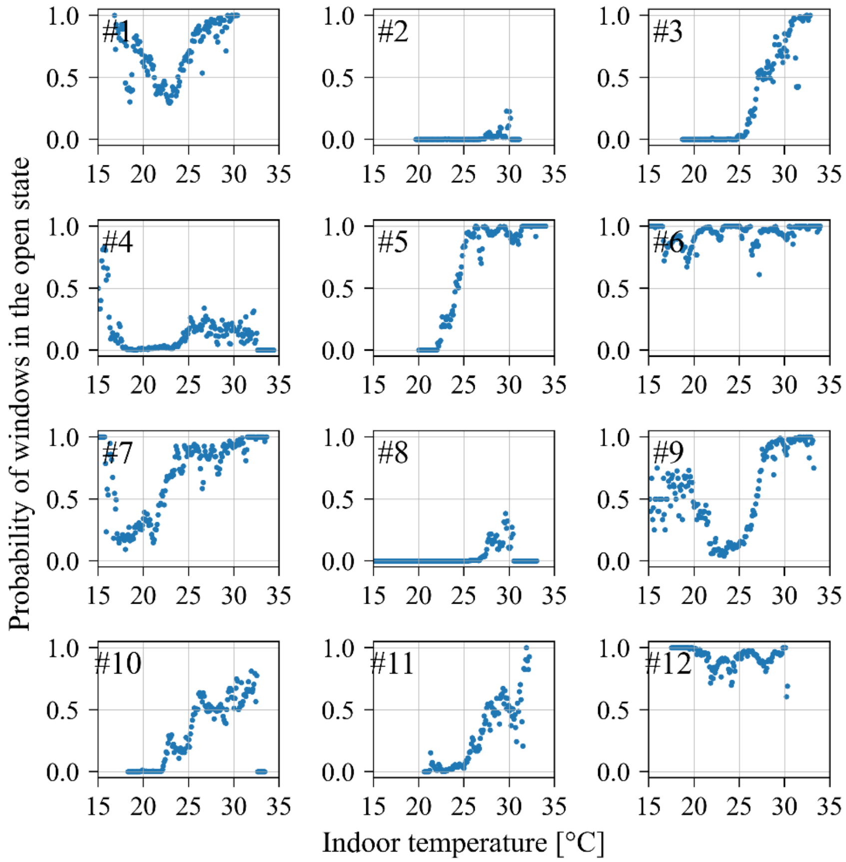


Figure 4.7 Indoor temperature versus probability of windows in the open state of the 12 households for one year

In addition to thermal comfort, the occupants open or close windows to improve indoor air quality. It has been reported that indoor CO2 and PM2.5 concentrations significantly influence the window-opening/closing behavior compared to the window's state (Anderson et al., 2013; Cali et al., 2016; Fabi et al., 2015). Figure 4.8 displays the correlation between the probability of occupant window-opening behavior and indoor CO_2 concentration. As shown, the occupants do not respond in low concentrations, but it appears that the probability of taking action to open the window increases above 1000 ppm. A threshold value of approximately 1000 ppm is indicated, similar to the results of previous studies (Yao and Zhao, 2017; Li et al., 2015; Persily, 2015). Similarly, for the indoor $PM_{2.5}$ concentration, a threshold value of approximately 100 $\mu g/m^3$ is indicated (Figure 4.9).

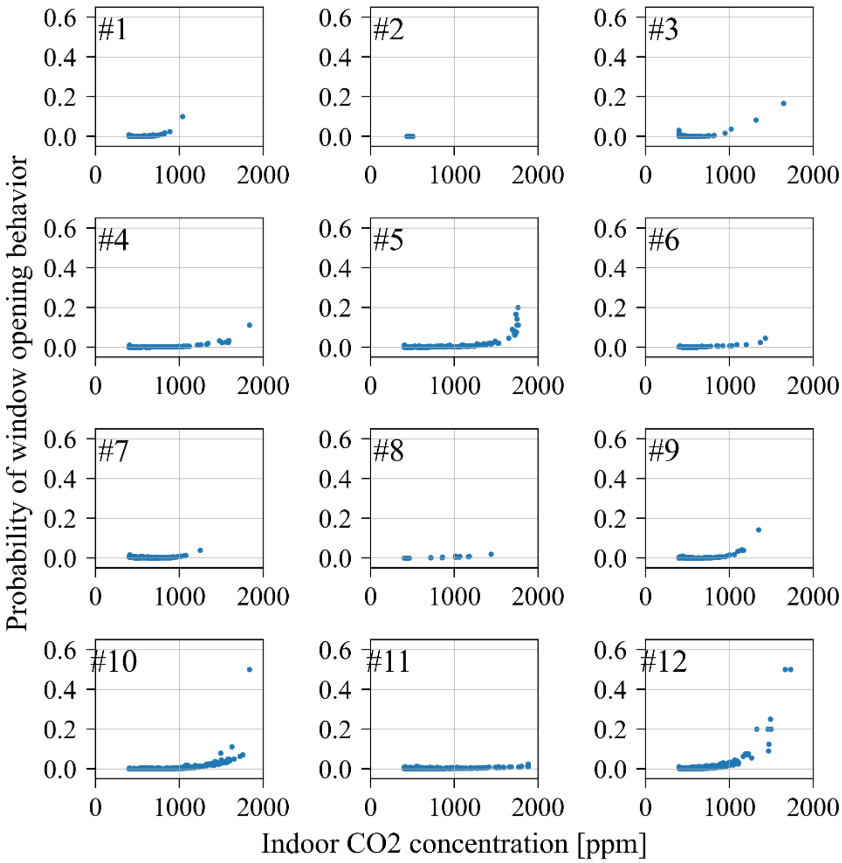


Figure 4.8 Indoor CO_2 concentration versus probability of window-opening behavior of the 12 households for one year

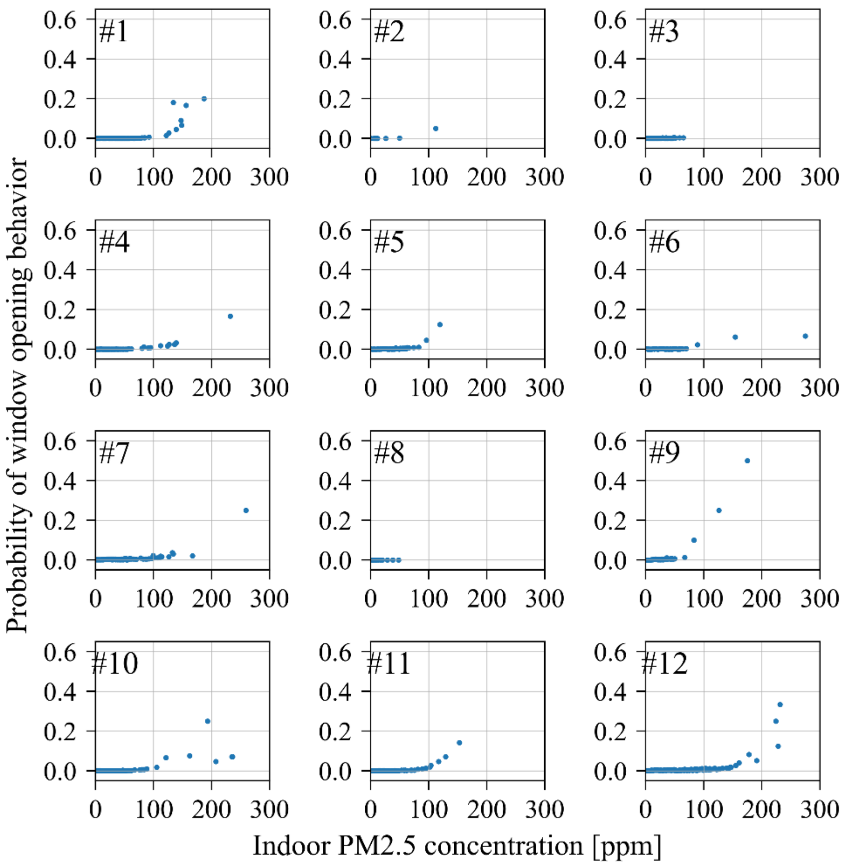


Figure 4.9 Indoor $PM_{2.5}$ concentration versus probability of window-opening behavior of the 12 households for one year

Figure 4.10 illustrates the correlation between the probability of windows in the open state and the time of day. The probability tends to increase during the daytime in several households (households #1, #3, #4, #7, #9, and #11). However, no particular tendency is observed in other households.

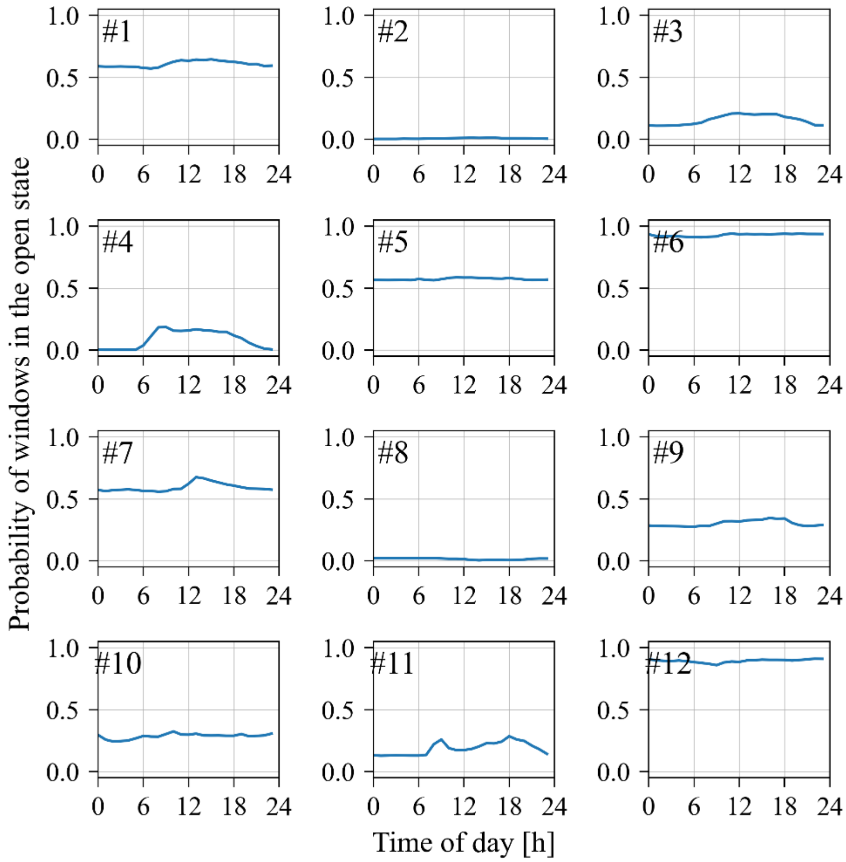


Figure 4.10 Time of day versus probability of windows in the open state of the 12 households for one year

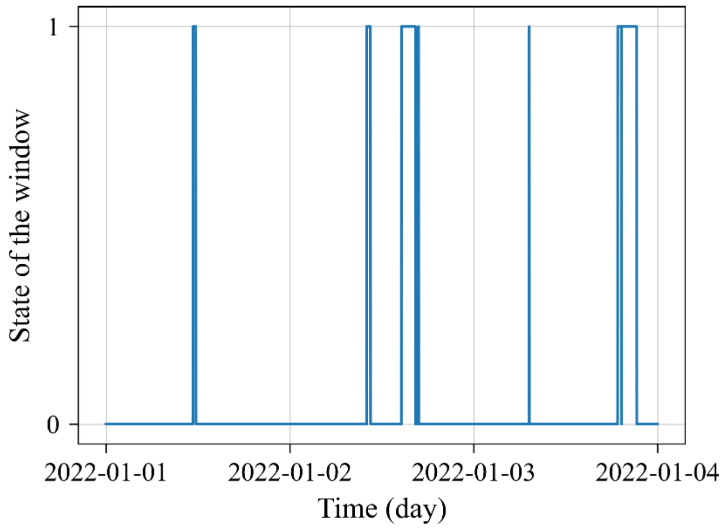
4.3.2 Implementing window state models

Table 4.3 presents the prediction accuracy of the window state models with the logistic regression and XGBoost classifier implemented for each of the 12 households. The logistic regression models show an accuracy between 79.9% and 99.3%, whereas the XGBoost models have a relatively high accuracy of ≥98.6%. The prediction accuracy is the lowest in household #7 for both models. This may be due to the influence of unknown factors that have not been collected or considered in the models. The analysis of the unknown factors will be covered in Section 4.4.

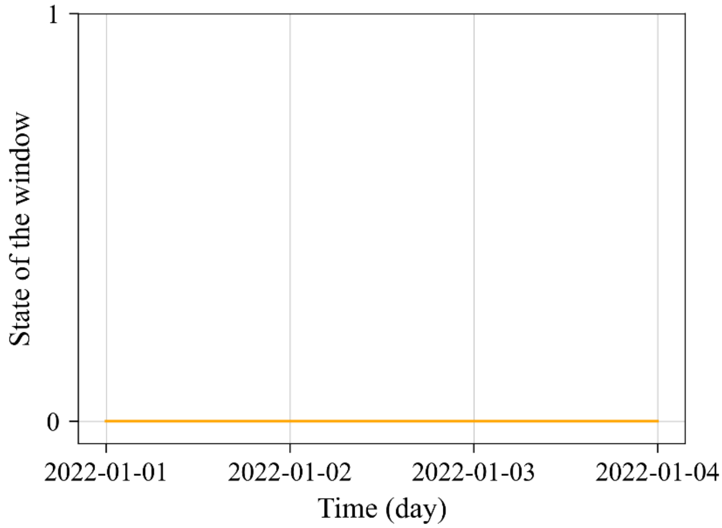
Table 4.3 Prediction accuracy of window state models using logistic regression and XGBoost (%)

Household #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Logistic regression	80.9	99.3	91.5	91.1	91.0	92.7	79.9	98.3	89.5	80.3	85.2	89.5
XGBoost	99.4	100.0	99.4	99.5	99.7	99.8	98.6	100.0	99.7	99.4	99.1	99.4

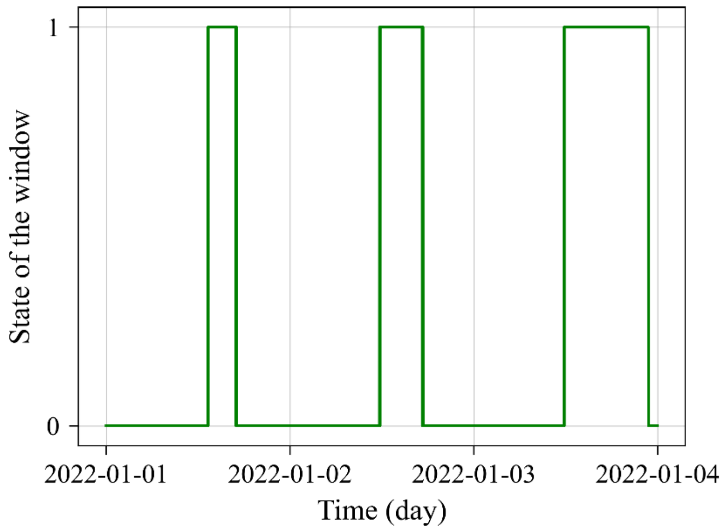
Figure 4.11 visualizes the measured and predicted window states through two models in a specific scenario. As evident from the figure, the XGBoost model, used for prediction, captures the temporary window opening behavior of occupants to some extent, whereas the logistic regression prediction model consistently predicts closed states. XGBoost is a well-known model for high prediction accuracy, and it demonstrates better performance in prediction accuracy compared to logistic regression.



(a) Measured window state



(b) Predicted window state with logistic regression



(c) Predicted window state with XGBoost classifier

Figure 4.11 Measured and predicted window state in a specific scenario

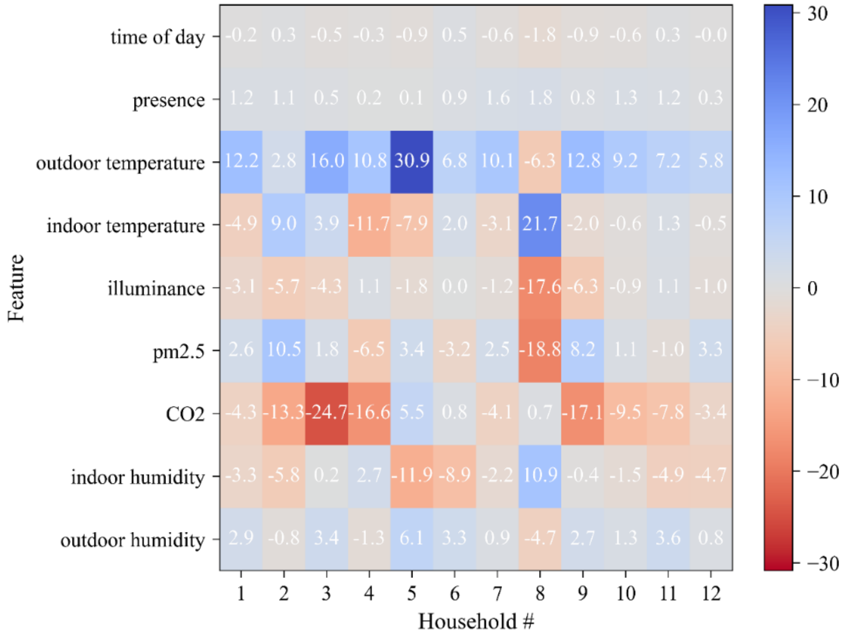
(household #1, January 1, 2022, to January 3, 2022)

4.3.3 Feature influence quantification

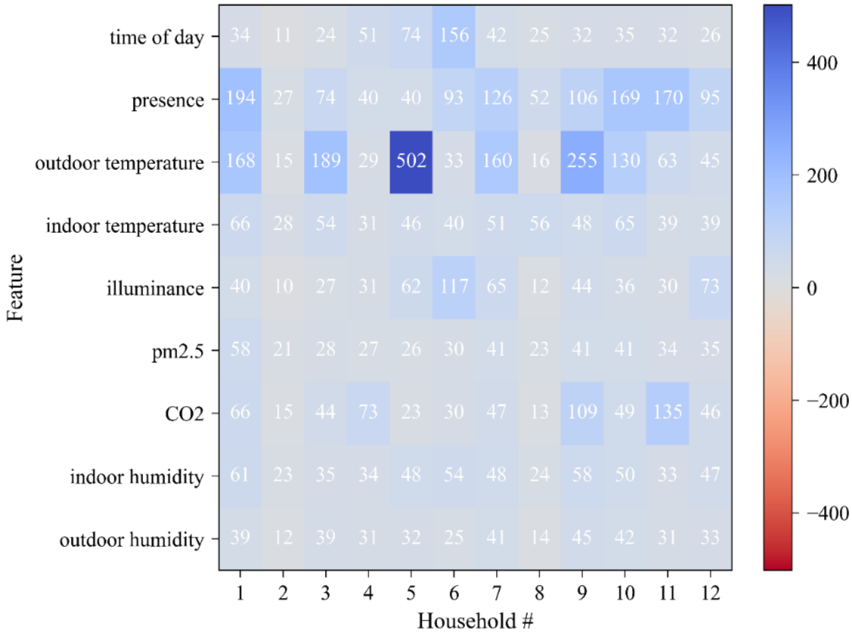
Figure 4.12 depicts the influence of each feature on the prediction of the window states of the 12 households by applying three different XAI methods to the aforementioned models. Figures 4.12(a), (b), and (c) display the results of the logistic regression, XGBoost with the feature importance method, and XGBoost with SHAP, respectively. The blue color of the cell indicates that the larger the corresponding feature value, the higher the probability of predicting the window in an open state. Conversely, the red color indicates that the larger the corresponding feature value, the higher the probability of predicting the window in a closed state. Notably, Figure 4.12(c) illustrates the mean Shapley value for the entire period. Different Shapley values can be obtained for the prediction of each time step.

Based on the results, the outdoor temperature and CO_2 concentration exhibit the largest positive and negative influences, respectively, in most households. The higher the outdoor temperature, the greater the number of households maintaining the window in the open state, and the higher the CO_2 concentration, the greater the number of households maintaining the window in the closed state. The indoor temperature also has a considerable positive or negative influence, depending on the household. The time of day has a significant influence only on specific households (Figures 4.12(b) and (c); households #4 and #6).

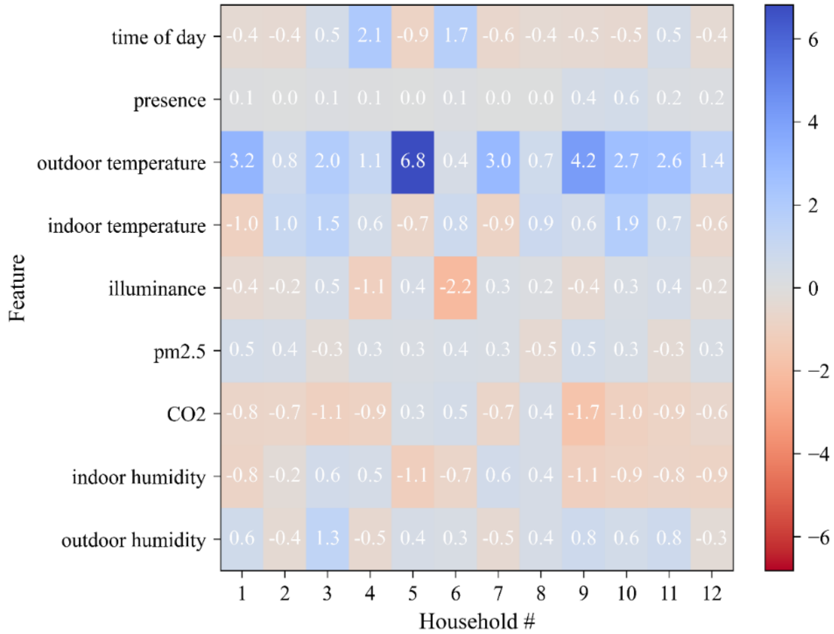
Comparing the three XAI methods, different results of the feature influence analysis are shown. Feature influence analysis using the coefficients of the logistic regression model makes it possible to quantify positive and negative influences based on a simple formula. However, the accuracy of the model is lower than that of the XGBoost model (Table 4.3), and a high intercorrelation between two or more independent features can lead to skewed or misleading results. The feature importance method supported by XGBoost has a reasonable model accuracy and low computational complexity. A significant limitation of the XGBoost feature importance method is that it cannot distinguish whether the feature has a positive or negative impact. SHAP compensates for the weaknesses of the previous two methods. SHAP can distinguish between positive and negative impacts and quantify the feature influence by considering the dependence between variables. In addition, SHAP can be applied to complex models, such as deep learning algorithms, known as unexplainable black-box models. However, the computational complexity of SHAP is high, so the time required is a hundredfold compared to that of the other two methods. Overall, SHAP is the most effective XAI technique for analyzing feature influence on predicting the window state, and it is recommended for quantifying occupant individual preferences (Kim and Park, 2023).



(a) Logistic regression



(b) XGBoost and feature importance method



(c) XGBoost and SHAP

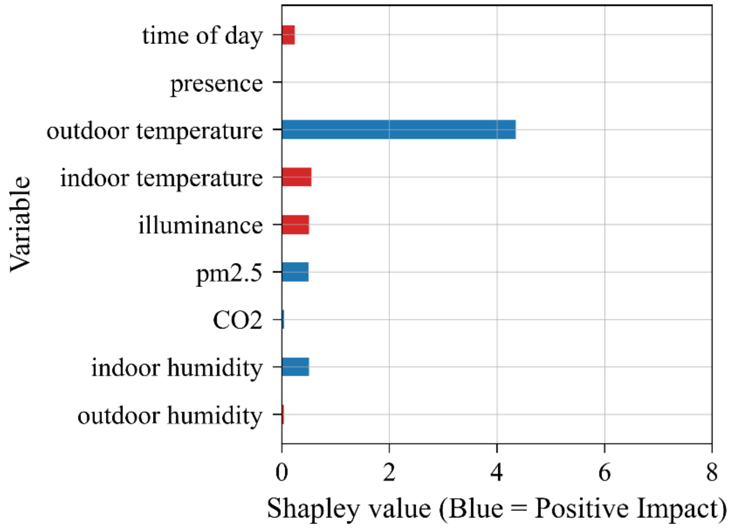
Figure 4.12 Quantified feature influence on window state prediction in the 12 households using SHAP (blue indicates a positive effect for windows in the open state)

4.4 Discussion

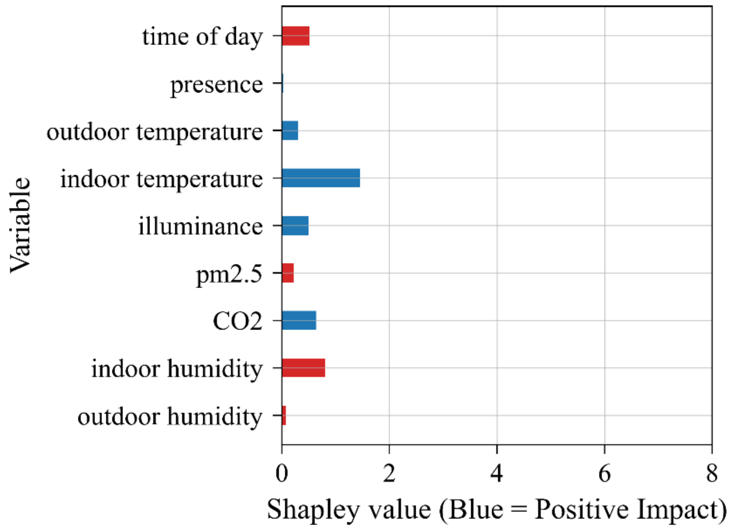
In this section, the author presents solutions or directions to address the three issues mentioned in Section 4.1, based on the results of the feature influence analysis of occupant behavior through the XAI methods.

First, there is significant spatial diversity of WAB among households. Specifically, the occupant's personal preferences vary from household to household. In Figure 4.12(c), each feature has either a positive or negative influence depending on the household, and this characteristic is prominent in the case of indoor temperature. In other words, the WAB, according to environmental/non-environmental variables, cannot be predicted deterministically, but it is important to quantify these probabilistic characteristic values, such as feature influence, with probabilistic specifications. In addition, it is difficult to describe the behavior of all occupants using one universal model. It is better to customize individual households.

Second, the occupant's preferences for WAB cannot be defined using only a single environmental parameter. WAB is a response to the interaction of environmental, non-environmental, and unknown factors. Furthermore, occupants behave differently, even in practically the same environment. Figure 4.13 shows the temporary feature influences when the window's state changes for 10 min. In this case, the occupants kept the window open for 30 h and then closed it. Therefore, the model prediction results can be interpreted as the positive effect of the reduced relatively low outdoor temperature, thus adjusting the window in the closed state. However, it is worth noting that the results of the feature influence analysis could be different from the actual intentions of the occupants. Table 4.3 indicates that the measured variables are practically the same. It is assumed that there are unknown or random factors not reflected in this study, such as psychological and social factors, which work in combination with other physical factors (Kim and Park, 2023).



(a) 2021-08-06 09:20



(b) 2021-08-06 09:30 (10 min after)

Figure 4.13 Quantified temporary feature influence on window state prediction in household #7 using SHAP

Table 4.4 Variables in the situation presented in Figure 4.13

Time	**2021-08-06 09:20**	**2021-08-06 09:30**
Window state	Open	Closed
Predicted window state	Open	Closed
Occupant presence	Present	Present
Outdoor temperature (℃)	29.7	29.9
Indoor temperature (℃)	30.4	30.4
Indoor illuminance (lx)	13	6
$PM_{2.5}$ concentration ($\mu g/m^3$)	10	8
CO_2 concentration (ppm)	513	511
Indoor humidity (% RH)	68	68
Outdoor humidity (% RH)	69	68
Air conditioner state	Off	Off

Finally, similar to logistic regression models, the current complex black-box models can also be described by applying the XAI techniques regarding feature influence. The approach presented in this chapter increases the computational complexity, but it can provide meaningful information for decision-making. Furthermore, the possibility of quantifying the diversity of occupant behavior by investigating individual preferences can be verified.

For future occupant behavior research, it should be recognized that there are unknown factors, and sufficient consideration should be given to such factors in the modeling process. Additionally, owing to the limitation of data collection, the average occupant was modeled on a household basis in this thesis. If it can be modeled based on each individual, the individual indoor environmental perception and its diversity can be analyzed.

4.5 Summary

This chapter implemented WAB models and quantified the individual preferences of households by analyzing the feature influence. Environmental data (temperature, humidity, CO_2 concentration, $PM_{2.5}$ concentration, illuminance, and time of day) and occupant data (occupant presence and window state) from 12 households were collected for one year. A logistic regression model and XGBoost classifier model were presented for window state prediction, and logistic regression, feature importance, and SHAP were used to examine the feature influence of WAB.

As a result of the preliminary analysis, the occupants adjusted the state of the window with different responses for each variable. Notably, the occupants did not react sensitively to low CO_2 concentrations, but a high CO_2 concentration of over 1000 ppm could trigger window-opening behavior. A similar result was observed for the $PM_{2.5}$ concentration, with a threshold value of 100 $\mu g/m^3$.

Window state prediction models were implemented using logistic regression and the XGBoost classifier. Logistic regression models exhibit an accuracy between 79.9% and 99.3%, whereas the XGBoost models have a relatively high accuracy of ≥98.6%. Three XAI methods, namely, logistic regression, feature importance, and SHAP, were applied to the two aforementioned prediction models to examine the effect of each variable on the prediction outcome. As a result, the outdoor temperature and CO_2 concentration were found to have the largest positive and negative influences, respectively, in most households. The indoor temperature has a significant influence as well, either positive or negative, depending on the household. Each XAI method has advantages and disadvantages, but SHAP is recommended to compensate for the disadvantages of the other two methods.

Regarding feature influence analysis, the following findings can provide insights into the issues in the current WAB research. First, different people have different personal preferences for using windows. Applying customized WAB models rather than a universal one is better. Second, the personal preferences of occupants on WAB cannot be defined using only a single environmental variable. WAB is a response to the interaction of environmental, non-environmental, and unknown factors. Finally, the current complex black-box models can also be described by applying XAI techniques regarding feature influence.

In summary, this chapter introduced a novel approach that quantifies and proves the spatial diversity of WAB.

Chapter 5. Multinomial occupant behavior model

5.1 Introduction

As mentioned in Section 1.2, occupant diversity encompasses three types: temporal, spatial, and behavioral. Previous studies have demonstrated temporal diversity (Ahn and Park, 2016; Ahn et al., 2017; Ahn and Park, 2019; Kim and Park, 2022). Similarly, spatial diversity has been explicitly acknowledged in the literature (Wang et al., 2022; O'Brian et al., 2017; Happle et al., 2020; D'Oca and Hong, 2014; Haldi et al., 2017; Liu et al., 2022; Yun et al., 2009; Markovic et al., 2018).

However, a limited number of studies modeled multiple behaviors (Dong et al., 2022). Verifying behavioral diversity is challenging due to privacy concerns and other reasons. It is crucial to validate whether reproducibility can be achieved when modeling different occupant behaviors based on the same environmental data.

Therefore, the objectives of this chapter are as follows:

- Implement a multinomial occupant behavior (OB) model using a deep-learning algorithm.
- Validate the reproducibility of the model when different behaviors are modeled using the same environmental data and algorithm.
- Analyze mutual information to quantify the degree of dependence between different behaviors with the results obtained from the implemented OB model.

By pursuing these objectives, this chapter aims to contribute to understanding behavioral diversity and investigate the potential for reproducibility in modeling diverse occupant behaviors. Furthermore, the analysis of mutual information provides insights into the intercorrelationships among behavior types and their impact on the performance of the OB model.

5.2 Modeling multinomial OB using LSTM

This section introduces the multinomial OB modeling method and analysis techniques. For sequence data-based multi-output label modeling, LSTM was utilized. Individual LSTM models were constructed for each household. Mutual information was employed to investigate the degree of dependence among the predicted behavior types.

In this chapter, six households in three residential apartment buildings in Seoul, South Korea, were selected (Figure 5.1). Occupant and environmental data measured for one year (August 3, 2021, to August 2, 2022) were used (Table 5.1).

The implemented model predicts occupants' behavior for the next one minute based on the environmental data input with a 60-minute timestep sequence. The data is sliced using the moving window technique (Figure 5.2). The sliced dataset is randomly split into training and testing data using a 7:3 ratio.

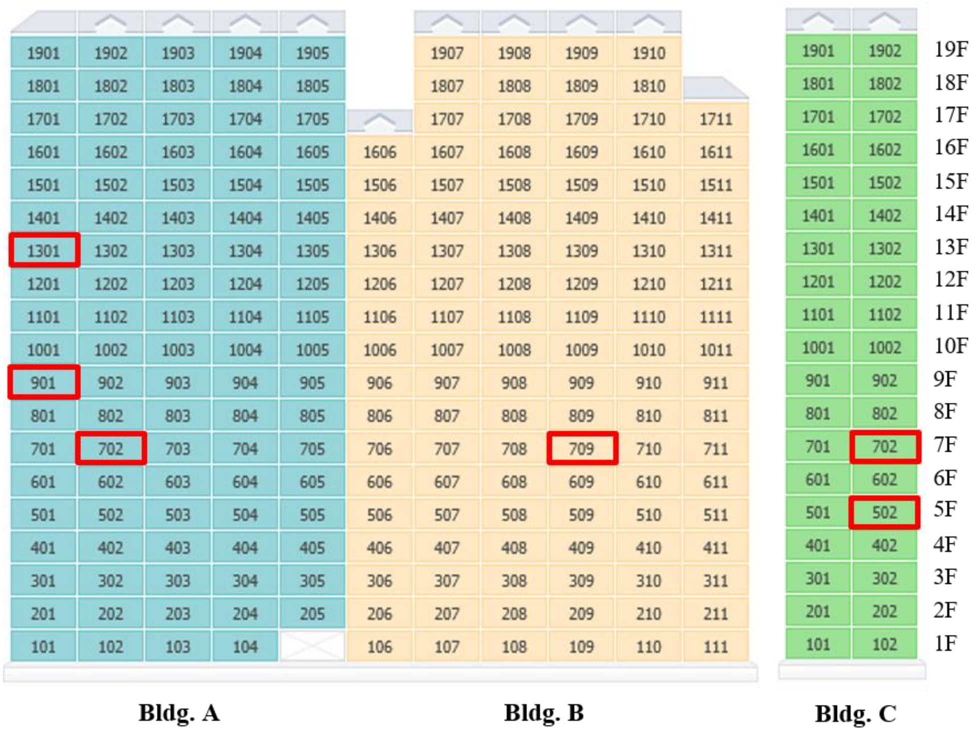


Figure 5.1 The six households selected

Table 5.1 Input and output data

Data type	Label	Sampling time	Sequence length	Number of datasets
Input	Outdoor temperature Indoor temperature Outdoor humidity Indoor humidity Indoor illuminance Indoor CO_2 concentration Indoor $PM_{2.5}$ concentration Time of day Month	1 min	60 min	
Output	Presence Window state Light switch AC switch Boiler switch	1 min	1 min	

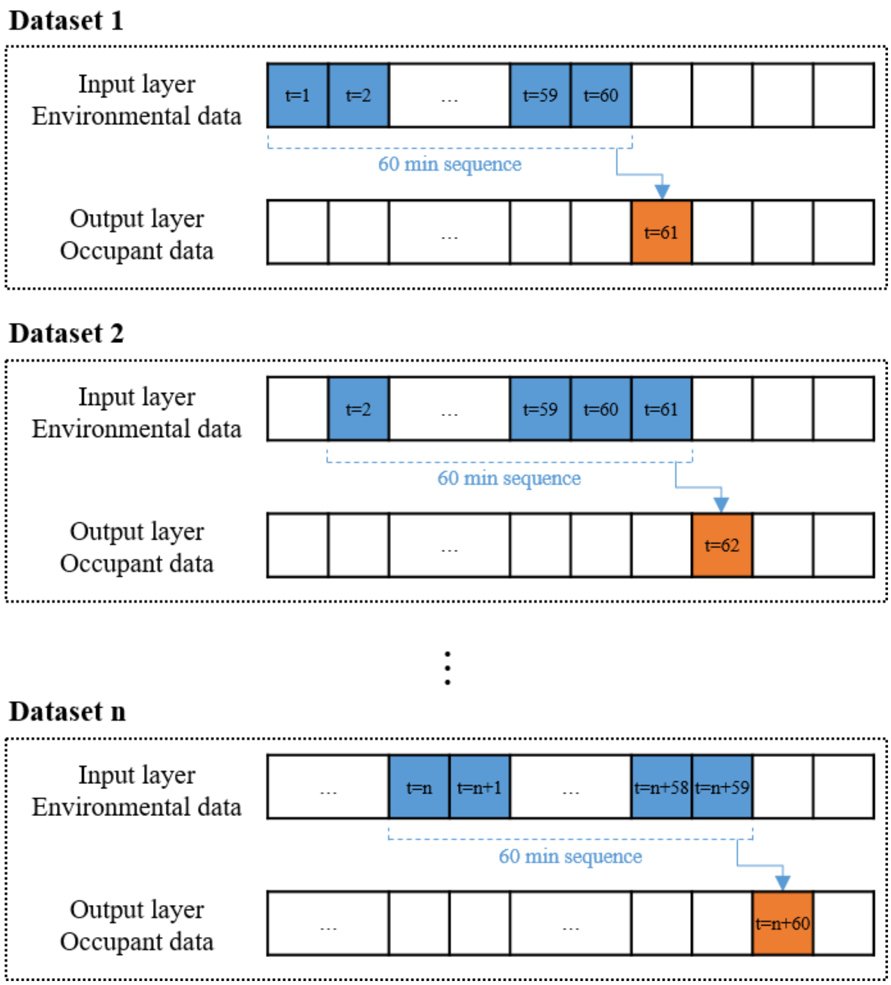


Figure 5.2 Dataset preparation for multinomial OB modeling

5.2.2 LSTM

LSTM stands for Long Short-Term Memory [Hochreiter, 1997], a recurrent neural network (RNN) type used in deep learning. LSTM is designed to handle the vanishing gradient problem that can occur in traditional RNNs when trying to learn long-term dependencies.

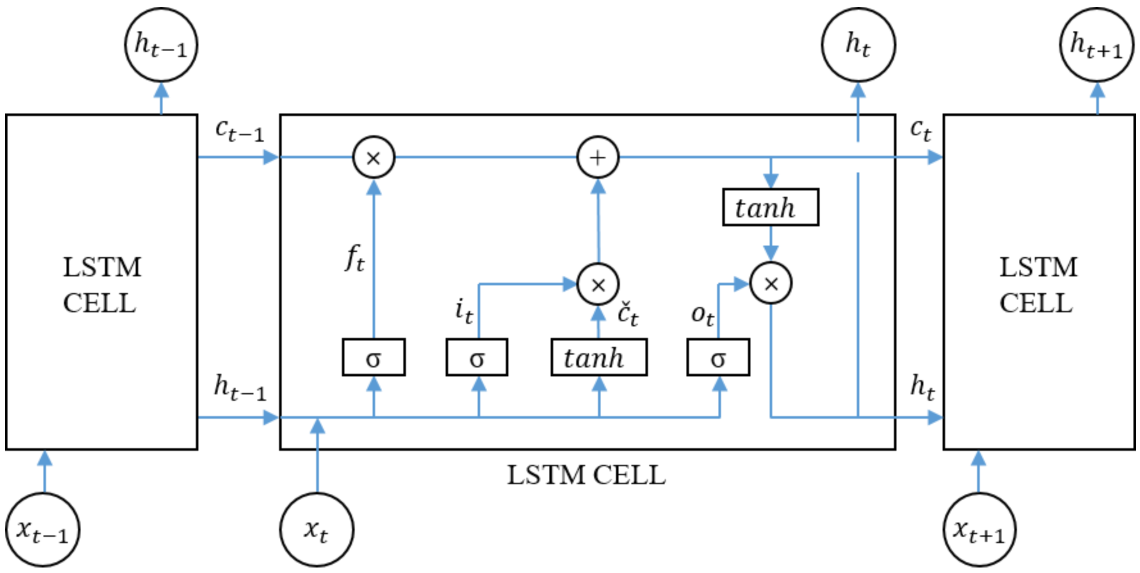
LSTM uses a memory cell and three gates (input gate, forget gate, and output gate) to control the flow of information within the network (Figure 5.3). The memory cell can store information over long periods, and the gates regulate the information flow into and out of the memory cell. This allows the network to selectively remember or forget information based on its relevance to the task and hand.

Figure 5.4 shows the whole structure of the LSTM model. The model receives nine input labels, then processed through two LSTM layers. The output values for the five occupant behaviors are then extracted through the dense layer. To avoid overfitting, a dropout layer is added after each LSTM layer. This helps to reduce the likelihood of the model becoming too closely fitted to the training data, which can lead to poor performance on new data.

During the model training process, the binary cross-entropy loss function was utilized. The loss function is commonly used in binary classification tasks in machine learning. The cross-entropy between two probability distributions p and q is defined as:

$$\mathrm{H}(P, Q) = -\mathrm{E}_p[\log q(x)] = -\sum_{x \in X} p(x) \log q(x) \qquad \text{Equation 5.1}$$

where $p(x)$ is the true probability distribution and $q(x)$ is the predicted probability distribution over a set of possible outcomes. It measures the dissimilarity between the predicted probability distribution and the true binary output. This is repeated for each example in the training dataset, and the average loss is then used to update the model parameters during training.



x_t	Input vector
h_t	Output vector
c_t	Cell state vector
f_t	Forget gate vector
i_t	Input gate vector
o_t	Output gate vector
σ	Sigmoid function
×	Point-by-point multiplication
+	Point-by-point addition

Figure 5.3 LSTM hidden Layer (Wang et al., 2018)

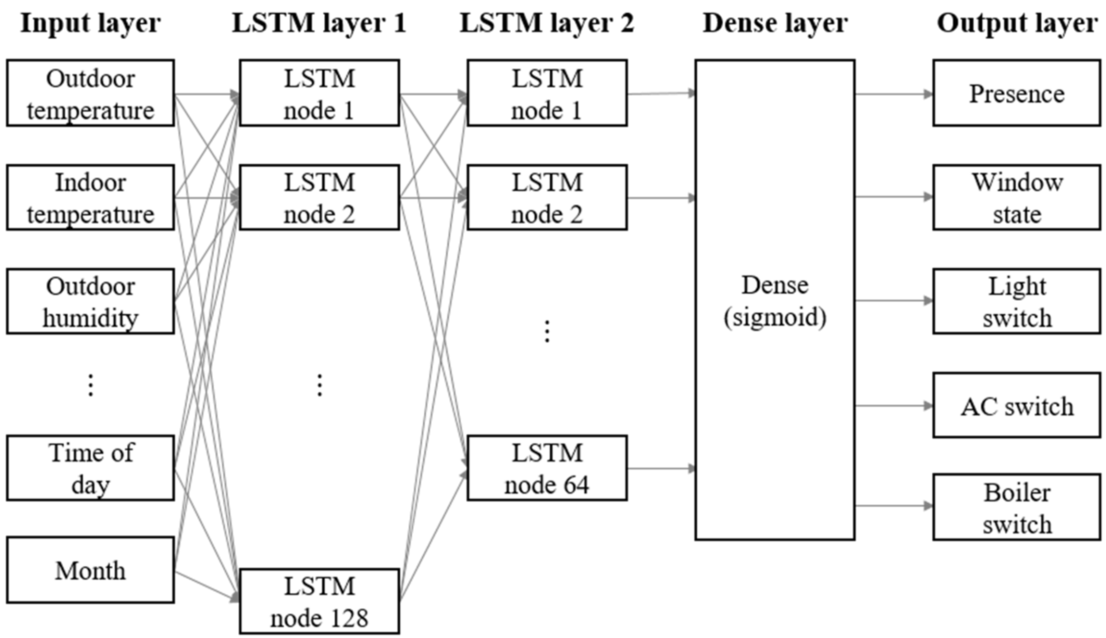


Figure 5.4 Whole Layer of the multinomial OB model using LSTM

5.2.3 Mutual information

The mutual information (MI) of two variables represents the mutual dependence between the two variables (Figure 5.5). It is used to identify the amount of information one variable provides about the other variable. Claude Shannon first introduced the quantity [C.E. Shannon, 1948], expressed as Equation 5.2.

$$\mathrm{I}(X;Y) = \sum_{y\in Y}\sum_{x\in X} P_{(X,Y)}(x,y)\log\left(\frac{P_{(X,Y)}(x,y)}{P_X(x)P_Y(y)}\right) \qquad \text{Equation 5.2}$$

Where (X,Y) is a pair of random variables, $P_{(X,Y)}$ is the joint probability mass function of X and Y, and P_X and P_Y are the marginal probability mass functions of X and Y, respectively. Notably, $\mathrm{I}(X;Y)$ is non-negative and equal to zero precisely when the joint distribution coincides with the product of the marginal, i.e., when X and Y are independent.

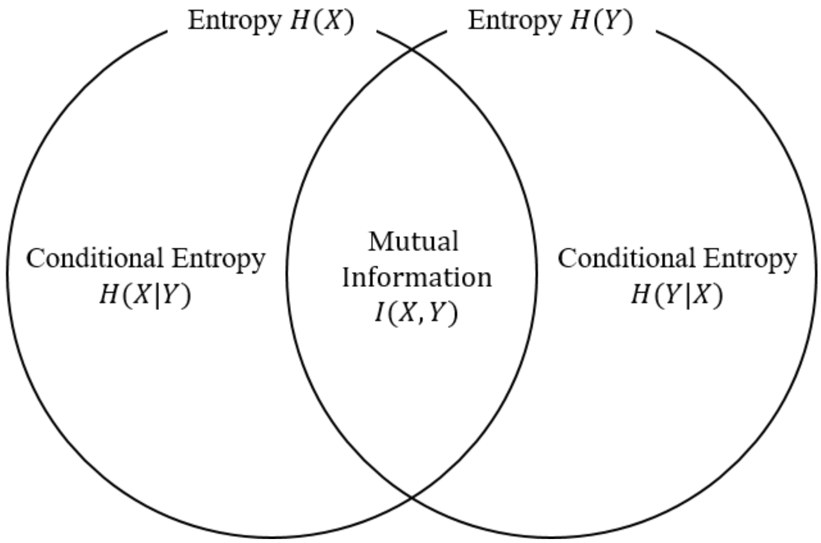


Figure 5.5 Mutual information and conditional entropy

5.3 Results

5.3.1 Analysis of measured occupant behavior by households

This section presents an overview of the measured occupant data. Figure 5.6 displays the proportion of the state of each behavior type for the six households. The household occupancy rates range from 0.75 to 0.91, showing relatively similar levels to other behavior types. However, there is a significant variation in the proportions of window and boiler states among the households. Regarding AC usage, it is observed that the proportion of the "on" state is relatively low compared to other behavior types. This is due to the temporary operation of the AC during the cooling season.

The household proportions' differences reflect their varying preferences and perceptions regarding indoor environment control and energy saving. For example, household #1 appears to prefer active indoor thermal environment control (AC and boiler) rather than relying on natural ventilation for temperature regulation. Conversely, in the case of household #4, there is a preference for utilizing natural ventilation by keeping the windows open for indoor environment control.

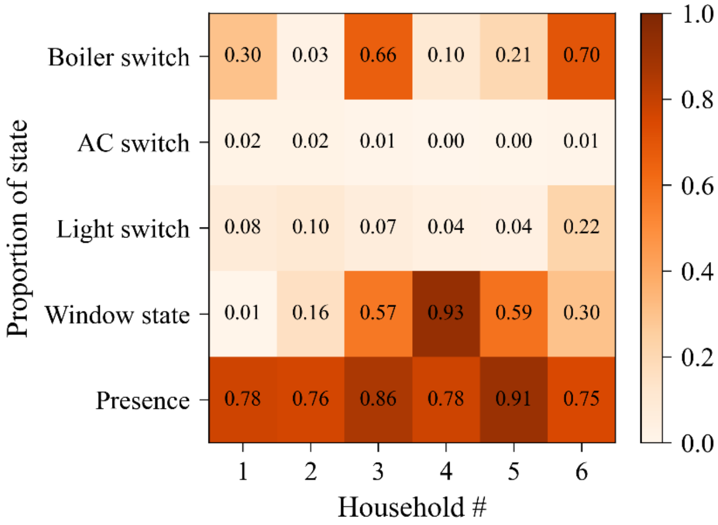


Figure 5.6 Proportion of measured state of each occupant behavior among six households (Presence: Presence, Window state: Open, Light switch: On, AC switch: On, Boiler switch: On)

5.3.2 Implementing multinomial OB model

The implemented multinomial OB model based on environmental variables demonstrates a high level of predictability, achieving label accuracies of 95% or higher (Table 5.3). However, the accuracy of predicting occupant presence and light switch state is relatively lower than other behaviors. The total accuracy, which represents the proportion of correctly predicted labels for all behaviors, ranges from 88.2% (household #2) to 97.7% (household #3) and varies among households. There is a noticeable trend of decreased total accuracy in households where specific labels exhibit significantly lower accuracy.

Notably, the model's accuracy varies depending on the specific behavior type and the households. Despite being based on the same environmental data, different behavior types and individual household characteristics contribute to the variation in accuracy.

Table 5.2 Prediction accuracies of implemented multinomial occupant behavior model using LSTM

Household #	**Label accuracy**					**Total accuracy**	**Loss**
	Presence	Window state	Light switch	AC switch	Boiler switch		
1	0.966	0.998	0.966	0.997	0.990	0.919	0.041
2	0.956	0.988	0.948	0.991	0.992	0.882	0.063
3	0.995	0.997	0.988	0.998	0.999	0.977	0.013
4	0.982	0.997	0.983	1.000	0.990	0.954	0.029
5	0.981	0.969	0.963	0.998	0.992	0.911	0.055
6	0.988	0.996	0.986	0.996	0.999	0.965	0.020

5.3.3 Mutual information analysis

Figure 5.7 displays the calculated mutual information between the multinomial OB model output labels for each household. There is a clear variation in the tendency of mutual information among the households. In general, occupant presence shows a relatively high level of dependence on other behaviors across all households. Two households (households #4 and #5) exhibit the highest dependence on occupant presence and window state. Another two households (households #1 and #3) show the highest dependence between window state and boiler switch. The remaining two households (households #2 and #6) demonstrate the highest dependence between occupant presence and light switch.

Overall, a mathematical correlation exists between the total accuracy (Table 5.3) and the mutual information. For instance, household #2 has relatively lower mutual information than other households, which is interpreted as lower total accuracy than label accuracy. Further detailed interpretation of the results will be discussed in Section 5.4.

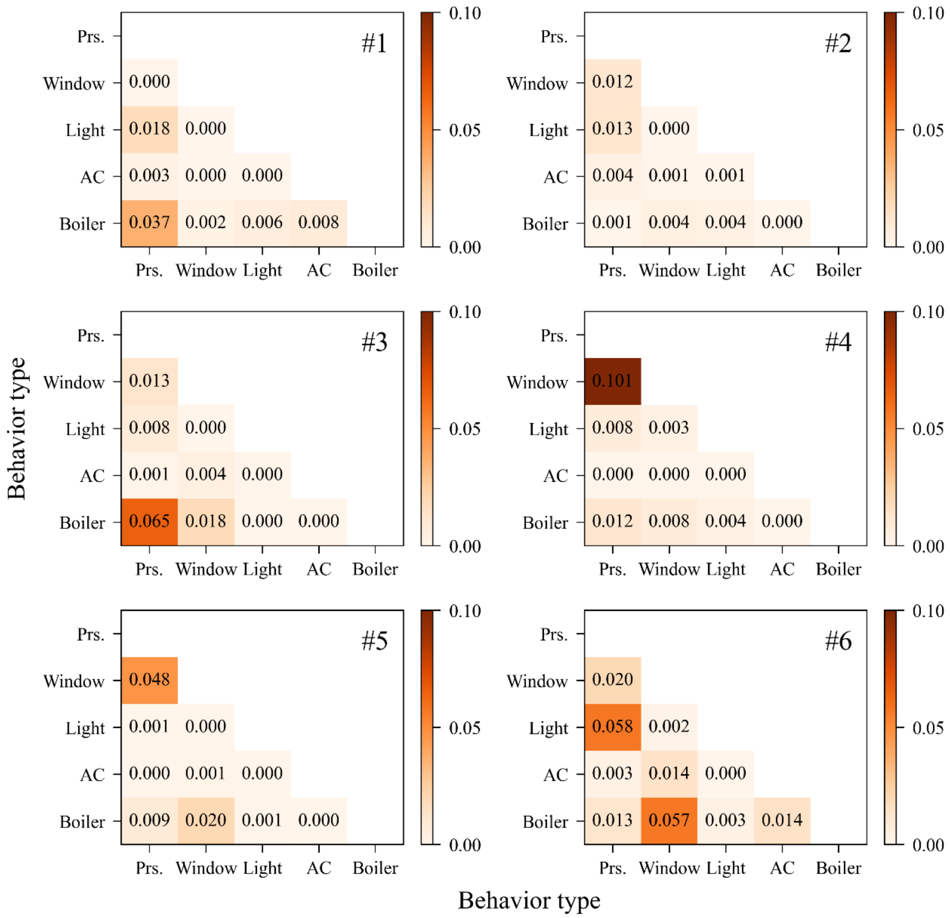


Figure 5.7 Degree of dependence between different behavior types in six households

5.4 Discussion

In this section, the author presents insights based on the results obtained from the implemented prediction model and mutual information analysis discussed in Section

5.3. These findings address the potential challenges that may arise in multinomial OB modeling. By leveraging the prediction model’s outcomes and mutual information metrics, the author suggests a guidance for enhancing the understanding and interpretation of occupant behavior in real-world settings.

- The accuracy of the prediction model varies across different behavior types, with occupant presence and light switch demonstrating relatively lower label accuracy, primarily due to the heavy reliance on thermal environmental variables for model training. In the case of the light switch, excluding the indoor illuminance variable, no other significant predictors are available for reliable prediction. To improve the model performance, it is suggested to incorporate additional factors that influence each specific behavior type as input data. Moreover, the behavior diversity is evident, as observed from the modeling conducted using the same dataset. Interestingly, even within the same model framework, noticeable performance discrepancies arise. This suggests behavioral diversity can hinder reproducibility, as it introduces variations in the model’s predictive performance.
- The proportion of each behavior type, label accuracies, and degree of dependence among behaviors vary across households, which can be attributed to the differences in energy efficacy among households. The spatial diversity within households contributes to the model's prediction performance variations. Unique characteristics of each household likely impact the modeling results, highlighting the importance of considering household-specific factors when developing predictive models for

occupant behavior. By accounting for these spatial diversities and tailoring the model to individual households, it is possible to improve the prediction performance and capture the nuances of occupant behavior more accurately.

- As mentioned in Section 5.3, the results of mutual information analysis indicate that occupant presence exhibits a high degree of dependence on other behavior types. Therefore, when occupant presence is measurable, incorporating it as an input variable in the model can potentially improve the model's performance. Additionally, considering the causal relationships among variables, models such as Bayesian Neural Networks (BNN) that can be structured with multiple layers could be a promising avenue for improvement. By leveraging such models, the accuracy and interpretability of the predictions can be enhanced, thereby addressing the limitations of the current modeling approach.

5.5 Summary

This chapter implemented a multinomial OB model and quantified the degree of dependences among behavior types of households by analyzing the mutual information. Environmental data (temperature, humidity, CO_2 concentration, $PM_{2.5}$ concentration, illuminance, time of day, and month) and occupant data (occupant presence, window state, light switch, AC switch, boiler switch) from six households were collected for one year. An LSTM model was presented for multinomial OB prediction, and mutual information was utilized to assess the degree of dependence between different behavior types.

Preliminary observation revealed significant variations in behavior types among households. While the occupancy rate showed relatively minor differences, the proportion of the window state exhibited a wide range of values from 0.01 to 0.93. These differences in proportions reflect varying preferences and perceptions regarding indoor environment control and energy conservation.

Moreover, the LSTM model demonstrated a high prediction accuracy of over 94.8% for each label. The accuracy varied depending on the label, indicating the presence of behavioral diversity. The total accuracy differed across households, ranging from 88.2% to 97.7%, showing a notable decrease in households where specific labels exhibited lower accuracy.

The mutual information analysis revealed that occupant presence had a high degree of dependence on other variables, suggesting that measuring or predicting occupant presence may be essential for modeling other behaviors. Additionally, the degree of dependence between behavior types differed among households, suggesting various perspectives and user efficacy for each behavior.

This chapter conducted multinomial OB modeling from the perspective of occupant diversity, providing evidence and quantification of behavior diversity. While there is a limitation in the number of households studied, the methodology and findings of this research can be applied to other OB modeling studies, contributing to the field.

Chapter 6. Conclusion

In conclusion, this thesis conducted three studies to analyze occupant diversity: (1) Predictability quantification in occupant presence, (2) Feature influence quantification in window adjustment behavior using XAI, and (3) Multinomial occupant behavior model. The experiments were conducted in three residential buildings in Seoul, South Korea, with 31 households selected for data collection. The collected data included various aspects of occupant behavior such as occupant presence, window state, light switch, AC switch, and Boiler switch. Environmental data deemed influential on occupant behavior, including indoor and outdoor temperature, indoor and outdoor humidity, indoor CO_2 concentration, indoor PM2.5 concentration, and indoor illuminance, were also collected.

Firstly, the study revealed that occupant presence in specific households followed a random walk pattern for short measurement periods, indicating the presence of significant temporal diversity in occupant behavior. The predictability of occupant presence varied significantly between households, and the degree of variation is shown in Figure 3.7. It is important to note that the predictability of occupant presence in residential buildings cannot be generalized as a single state and varies widely based on occupant features such as the number of family members and occupancy patterns. Therefore, to implement a reliable occupant behavior model, it is necessary to assess its applicability to each occupant and building/space in terms of spatial diversity.

Second, the analysis of feature influence provided insights into the current research on window adjustment behavior (WAB). It was observed that individuals have different personal preferences for using windows, suggesting the need for customized models rather than a universal approach. Furthermore, the personal preferences of occupants regarding WAB cannot be solely defined based on a single environmental variable. A combination of environmental, non-environmental, and unknown factors influences WAB. Applying XAI techniques to understand feature influence can help describe complex black-box models.

Lastly, significant variations in behavior types were observed among households. While the occupancy rate showed relatively minor differences, the proportion of window states exhibited a wide range of values, indicating varying preferences and perceptions regarding indoor environment control and energy conservation. The mutual information analysis highlighted a high degree of dependence between occupant presence and other variables, suggesting the importance of measuring or predicting occupant presence when modeling other behaviors. Moreover, the degree of dependence between behavior types varied among households, indicating diverse perspectives and user efficacy for each behavior.

Overall, this thesis has thoroughly analyzed the causes behind the performance gap in building simulations resulting from occupant behavior modeling. The insights gained from this research highlight the importance of considering temporal, spatial, and behavioral diversity in occupant modeling to enhance the accuracy and effectiveness of building performance simulations.

Future research in this field can build upon these findings and focus on developing occupant models that are more reliable, reproducible, and replicable by incorporating occupant diversity. The author anticipates that the outcomes of this thesis will contribute significantly to understanding occupant behavior and bridging the gap between actual and simulated building energy use.

References

Ahn, K. U., & Park, C. S. (2016). Correlation between occupants and energy consumption. *Energy and Buildings*, *116*, 420-433.

Ahn, K. U., & Park, C. S. (2019). Temporal and spatial variation in the predictability of building occupancy. *Building and Environment*, *149*, 477-489.

Ahn, K. U., Kim, D. W., Park, C. S., & de Wilde, P. (2017). Predictability of occupant presence and performance gap in building energy simulation. *Applied Energy*, *208*, 1639-1652.

Altomonte, S., & Schiavon, S. (2013). Occupant satisfaction in LEED and non-LEED certified buildings. *Building and Environment*, *68*, 66-76.

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.

Aragon, V., Gauthier, S., Warren, P., James, P. A., & Anderson, B. (2019). Developing English domestic occupancy profiles. *Building Research & Information*, *47*(4), 375-393.

Azar, E., & Menassa, C. C. (2012). A comprehensive analysis of the impact of occupancy parameters in energy simulation of office buildings. *Energy and buildings*, *55*, 841-853.

Balvedi, B. F., Ghisi, E., & Lamberts, R. (2018). A review of occupant behaviour in residential buildings. *Energy and Buildings*, *174*, 495-505.

Barthelmes, V. M., Heo, Y., Fabi, V., & Corgnati, S. P. (2017). Exploration of the Bayesian Network framework for modelling window control behaviour. Building and Environment, 126, 318-330.

Bartlett, M. S. (1967). Some remarks on the analysis of time-series. *Biometrika*, *54*(1-2), 25-38.

Brager, G., Paliaga, G., & de Dear, R. (2004). Operable windows, personal control and occupant comfort.

Calì, D., Andersen, R. K., Müller, D., & Olesen, B. W. (2016). Analysis of occupants' behavior related to the use of windows in German households. Building and Environment, 103, 54-69.

Carlucci, S., De Simone, M., Firth, S. K., Kjærgaard, M. B., Markovic, R., Rahaman, M. S., ... & Van Treeck, C. (2020). Modeling occupant behavior in buildings. *Building and Environment*, *174*, 106768.

Carlucci, S., Lobaccaro, G., Li, Y., Lucchino, E. C., & Ramaci, R. (2016). The effect of spatial and temporal randomness of stochastically generated occupancy schedules on the energy performance of a multiresidential building. *Energy and Buildings*, *127*, 279-300.

Chakraborty, D., & Elzarka, H. (2019). Early detection of faults in HVAC systems using an XGBoost model with a dynamic threshold. Energy and Buildings, 185, 326-344.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

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.

D'Oca, S., Hong, T., & Langevin, J. (2018). The human dimensions of energy use in buildings: A review. *Renewable and Sustainable Energy Reviews*, *81*, 731-742.

Diggle, P. J., & Fisher, N. I. (1991). Nonparametric comparison of cumulative periodograms. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *40*(3), 423-434.

Dong, B., Markovic, R., Carlucci, S., Liu, Y., Wagner, A., Liguori, A., ... & Kang, X. (2022). A guideline to document occupant behavior models for advanced building controls. *Building and Environment*, 109195.

Dong, B., Yan, D., Li, Z., Jin, Y., Feng, X., & Fontenot, H. (2018, October). Modeling occupancy and behavior for better building design and operation—A critical review. In *Building Simulation* (Vol. 11, pp. 899-921). Springer Berlin Heidelberg.

Fabi, V., Andersen, R. K., & Corgnati, S. (2015). Verification of stochastic behavioural models of occupants' interactions with windows in residential buildings. Building and Environment, 94, 371-383.

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.

Fan, C., Xiao, F., & Zhao, Y. (2017). A short-term building cooling load prediction method using deep learning algorithms. Applied energy, 195, 222-233.

Frontczak, M., Schiavon, S., Goins, J., Arens, E., Zhang, H., & Wargocki, P. (2012). Quantitative relationships between occupant satisfaction and satisfaction aspects of indoor environmental quality and building design. *Indoor air*, *22*(2), 119-131.

Gorissen, D., Dhaene, T., & De Turck, F. (2009). Evolutionary model type selection for global surrogate modeling. Journal of Machine Learning Research, 10, 2039-2078.

Gu, Y., Cui, T., Liu, K., Yang, F., Wang, S., Song, H., ... & Li, Y. (2021). Study on influencing factors for occupant window-opening behavior: Case study of an office building in Xi'an during the transition season. Building and Environment, 200, 107977.

Gunning, D. (2017). Explainable artificial intelligence (xai). Defense advanced research projects agency (DARPA), nd Web, 2(2), 1.

Haldi, F., & Robinson, D. (2009). Interactions with window openings by office occupants. Building and Environment, 44(12), 2378-2395.

Haldi, F., Calì, D., Andersen, R. K., Wesseling, M., & Müller, D. (2017). Modelling diversity in building occupant behaviour: a novel statistical approach. Journal of BuPerformance Simulation, 10(5-6), 527-544.

Haldi, F., Calì, D., Andersen, R. K., Wesseling, M., & Müller, D. (2017). Modelling diversity in building occupant behaviour: a novel statistical approach. *Journal of Building Performance Simulation*, *10*(5-6), 527-544.

Han, M., May, R., Zhang, X., Wang, X., Pan, S., Da, Y., & Jin, Y. (2020). A novel reinforcement learning method for improving occupant comfort via window opening and closing. Sustainable Cities and Society, 61, 102247.

Happle, G., Fonseca, J. A., & Schlueter, A. (2020). Impacts of diversity in commercial building occupancy profiles on district energy demand and supply. *Applied Energy*, *277*, 115594.

Hastie, T. J. (2017). Generalized additive models. In Statistical models in S (pp. 249-307). Routledge.

Hipel, K. W., & McLeod, A. I. (1994). *Time series modelling of water resources and environmental systems*. Elsevier.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hong, T., D'Oca, S., Turner, W. J., & Taylor-Lange, S. C. (2015). An ontology to represent energy-related occupant behavior in buildings. Part I: Introduction to the DNAs framework. Building and Environment, 92, 764-777.

Hong, T., Yan, D., D'Oca, S., & Chen, C. F. (2017). Ten questions concerning occupant behavior in buildings: The big picture. *Building and Environment*, *114*, 518-530.

Kim, S. H., & Park, C. S. (2022). Predictability quantification of occupant presence in high-rise residential apartment buildings. *Energy and Buildings*, *275*, 112494.

Kim, S. H., & Park, C. S. (2023). Quantification of occupant response to influencing factors of window adjustment behavior using explainable AI. (under review)

Langevin, J., Wen, J., & Gurian, P. L. (2015). Simulating the human-building interaction: Development and validation of an agent-based model of office occupant behaviors. *Building and Environment*, *88*, 27-45.

Lee, Y. S., & Kim, S. K. (2008). Indoor environmental quality in LEED-certified buildings in the US. *Journal of Asian Architecture and Building*

Engineering, *7*(2), 293-300.

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.

Li, Z., & Dong, B. (2017). A new modeling approach for short-term prediction of occupancy in residential buildings. Building and Environment, 121, 277-290.

Liu, Y., Chong, W. T., Cao, Y., Liu, H., Yu, H., Cui, T., ... & Pan, S. (2022). Characteristics analysis and modeling of occupants' window operation behavior in hot summer and cold winter region, China. Building and Environment, 216, 108998.

Liu, Y., Chong, W. T., Yau, Y. H., Chang, L., Cui, T., Yu, H., ... & Pan, S. (2022). Rethinking the limitations of research on occupants' window-opening behavior: A review. Energy and Buildings, 112552.

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

Luo, N., Pritoni, M., & Hong, T. (2021). An overview of data tools for representing and managing building information and performance data. *Renewable and Sustainable Energy Reviews*, *147*, 111224.

Markovic, R., Frisch, J., & van Treeck, C. (2019). Learning short-term past as predictor of window opening-related human behavior in commercial buildings. Energy and Buildings, 185, 1-11.

Markovic, R., Grintal, E., Wölki, D., Frisch, J., & van Treeck, C. (2018). Window opening model using deep learning methods. *Building and Environment*, *145*, 319-329.

Mo, H., Sun, H., Liu, J., & Wei, S. (2019). Developing window behavior models for residential buildings using XGBoost algorithm. Energy and Buildings, 205, 109564.

Nason, G. P., & Savchev, D. (2014). White noise testing using wavelets. *Stat*, *3*(1), 351-362.

Newton, H. J. (1988). *Timeslab: a time series analysis laboratory*. Wadsworth Publ. Co.

Nicol, J. F. (2001, August). Characterising occupant behaviour in buildings: towards a stochastic model of occupant use of windows, lights, blinds, heaters and fans. In Proceedings of the seventh international IBPSA conference, Rio (Vol. 2, pp. 1073-1078).

Niu, B., Li, D., Yu, H., Mahyuddin, N., Liu, Y., Wu, J., ... & Pan, S. (2022). Investigation of occupant window opening behaviour during the summer period in a Beijing maternity hospital. Journal of Building Engineering, 45, 103441.

Norouziasl, S., Jafari, A., & Zhu, Y. (2021). Modeling and simulation of energy-related human-building interaction: A systematic review. *Journal of Building Engineering*, *44*, 102928.

O'Brien, W., Wagner, A., Schweiker, M., Mahdavi, A., Day, J., Kjærgaard, M. B., ... & Berger, C. (2020). Introducing IEA EBC annex 79: Key challenges and opportunities in the field of occupant-centric building design and operation. *Building and Environment*, *178*, 106738.

O’Brien, W., Gaetani, I., Gilani, S., Carlucci, S., Hoes, P. J., & Hensen, J. (2017). International survey on current occupant modelling approaches in building performance simulation. *Journal of Building Performance Simulation*, *10*(5-6), 653-671.

O'Brien, W., Gunay, H. B., Tahmasebi, F., & Mahdavi, A. (2017). A preliminary study of representing the inter-occupant diversity in occupant modelling. *Journal of Building Performance Simulation*, *10*(5-6), 509-526.

Page, J., Robinson, D., Morel, N., & Scartezzini, J. L. (2008). A generalised stochastic model for the simulation of occupant presence. *Energy and buildings*, *40*(2), 83-98.

Pan, S., Han, Y., Wei, S., Wei, Y., Xia, L., Xie, L., ... & Yu, W. (2019). A model based on Gauss Distribution for predicting window behavior in building. Building and Environment, 149, 210-219.

Pan, S., Xiong, Y., Han, Y., Zhang, X., Xia, L., Wei, S., ... & 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.

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.

Park, J., Jeong, B., Chae, Y. T., & Jeong, J. W. (2021). Machine learning algorithms for predicting occupants' behaviour in the manual control of windows for cross-ventilation in homes. Indoor and Built Environment, 30(8), 1106-1123.

Pearson, K. (1905). The problem of the random walk. *Nature*, *72*(1865), 294-294.

Persily, A. K. (2015, September). Indoor carbon dioxide concentrations in ventilation and indoor air quality standards. In 36th AIVC Conference Effective Ventilation in High Performance Buildings (pp. 810-819).

Plieninger, T., Draux, H., Fagerholm, N., Bieling, C., Bürgi, M., Kizos, T., ... & Verburg, P. H. (2016). The driving forces of landscape change in Europe: A systematic review of the evidence. Land use policy, 57, 204-214.

Qu, S., Wang, Z., & Liu, W. (2021). Clothing adjustment in outdoor environment: A new clothing model based on temperature change. *Building and Environment*, *206*, 108395.

Richardson, I., Thomson, M., & Infield, D. (2008). A high-resolution domestic building occupancy model for energy demand simulations. *Energy and buildings*, *40*(8), 1560-1566.

Robinson, D., Wilke, U., & Haldi, F. (2011). Multi agent simulation of occupants' presence and behaviour. *Proceedings of building simulation,* 14-16.

Salimi, S., Liu, Z., & Hammad, A. (2019). Occupancy prediction model for open-plan offices using real-time location system and inhomogeneous Markov chain. *Building and Environment*, *152*, 1-16.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Shapley, L. (1953). Quota solutions op n-person games1. Edited by Emil Artin and Marston Morse, 343.

Stazi, F., Naspi, F., & D'Orazio, M. (2017). A literature review on driving factors and contextual events influencing occupants' behaviours in buildings. Building and Environment, 118, 40-66.

Sun, C., Zhang, R., Sharples, S., Han, Y., & Zhang, H. (2018). A longitudinal study of summertime occupant behaviour and thermal comfort in office buildings in northern China. Building and Environment, 143, 404-420.

Van Lent, M., Fisher, W., & Mancuso, M. (2004, July). An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the national conference on artificial intelligence (pp. 900-907). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Wagner, A., O'Brien, W., & Dong, B. (2018). Exploring occupant behavior in buildings. *Wagner, A., O'Brien, W., Dong, B., Eds*.

Wang, C., Hao, P., Wu, G., Qi, X., & Barth, M. (2018). *Predicting the number of uber pickups by deep learning* (No. 18-06738).

Wang, C., Yan, D., & Jiang, Y. (2011). A novel approach for building occupancy simulation. *Building simulation,* 4(2), 149-167.

Wang, Y., Cooper, E., Tahmasebi, F., Chalabi, Z., Stamp, S., Burman, E., & Mumovic, D. (2022). A multilevel window state model based on outdoor environmental conditions that captures behavioural variation at room and apartment levels. *Energy and Buildings*, *277*, 112562.

Wei, S., Jones, R., & De Wilde, P. (2014). Driving factors for occupant-controlled space heating in residential buildings. Energy and Buildings, 70, 36-44.

Wei, Y., Yu, H., Pan, S., Xia, L., Xie, J., Wang, X., ... & Li, Q. (2019). Comparison of different window behavior modeling approaches during transition season in Beijing, China. Building and Environment, 157, 1-15.

Yan, D., Hong, T., Dong, B., Mahdavi, A., D'Oca, S., Gaetani, I., & Feng, X. (2017). IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings. *Energy and Buildings*, *156*, 258-270.

Yang, T., Bandyopadhyay, A., O'Neill, Z., Wen, J., & Dong, B. (2022, June). From occupants to occupants: A review of the occupant information understanding for building HVAC occupant-centric control. In *Building simulation* (Vol. 15, No. 6, pp. 913-932). Beijing: Tsinghua University Press.

Yao, M., & Zhao, B. (2017). Window opening behavior of occupants in residential buildings in Beijing. Building and Environment, 124, 441-449.

Yilmaz, S., Firth, S. K., & Allinson, D. (2017). Occupant behaviour modelling in domestic buildings: the case of household electrical appliances. *Journal of Building Performance Simulation*, *10*(5-6), 582-600.

Yun, G. Y., & Steemers, K. (2008). Time-dependent occupant behaviour models of window control in summer. Building and environment, 43(9), 1471-1482.

Yun, G. Y., & Steemers, K. (2010). Night-time naturally ventilated offices: Statistical simulations of window-use patterns from field monitoring. Solar Energy, 84(7), 1216-1231.

Yun, G. Y., Tuohy, P., & Steemers, K. (2009). Thermal performance of a naturally ventilated building using a combined algorithm of probabilistic occupant behaviour and deterministic heat and mass balance models. Energy and buildings, 41(5), 489-499.

Zhou, X., Ren, J., An, J., Yan, D., Shi, X., & Jin, X. (2021). Predicting open-plan office window operating behavior using the random forest algorithm. Journal of Building Engineering, 42, 102514.

국문 초록

재실자 행동 모델링을 위한 재실자 다양성 분석

김승현
건축학과 건축공학전공
서울대학교 대학원

재실자 행동은 건물 에너지 시뮬레이션에서 주요한 불확실성 요소 중 하나로 간주되며, 건물 내 재실자의 다양한 행동 양상을 정확하게 모델링하는 것은 건물 내 환경 제어와 에너지 수요 예측을 위하여 필수적이다. 이 연구는 건물 내 재실자 행동의 다양성이 재실자 모델링에서 미치는 영향을 조사하고자 한다. 이를 위하여 이전 연구들에서 제안된 재실자 행동 모델링 접근 방식들을 검토하고, 건물 에너지 시뮬레이션에서의 재현성과 복제성을 어떻게 보완할 수 있는지에 대해 탐구한다. 이를 위해 본 연구에서는 서울시 소재 3개의 주거 건물에서 31개 세대를 선정하여 데이터를 수집하였다. 수집된 데이터는 재실자의 재실 상태, 창문 상태, 조명 스위치, 에어컨 스위치, 보일러 스위치의 다양한 재실자 행동 유형 데이터와 함께, 실내 및 실외 온도, 습도, 이산화탄소 농도, 미세먼지 농도, 조도의 재실자 행동에 영향을 미치는 환경 데이터도 수집되었다.

연구 결과, 재실자의 재실 상태는 짧은 측정 기간 동안 특정 가구에서 무작위한 이동 패턴을 보였으며, 이는 재실자 행동의 시간적 다양성을 나타내는 것으로 해석 가능하다. 재실자 재실 상태 예측의 정확도는 세대마다 큰 차이를 보이며, 신뢰할 수 있는 재실자 행동 모델을 구현하기 위해서는 공간적 다양성 측면에서 특정 재실자 및 건물/공간에 대한 적용 가능성을 평가하는 것이 필요하다.

또한, 창문 조절 행동에 대한 특성 영향 분석 결과 각 세대는 창문 사

용에 대한 개인적인 선호가 존재하며, 일반적인 접근 방식보다는 세대별 맞춤형 모델이 필요함을 시사한다. 창문 조절 행동에 영향을 미치는 요인은 단일 환경 변수에만 근거할 수 없으며, 환경, 비환경 및 알려지지 않은 요소들의 조합이 영향을 미친다.

마지막으로, 행동 유형 간 차이가 크게 나타났다. 서로 다른 행동 유형 간에는 세대에 따라 종속성의 정도가 다르며, 이는 각 세대들의 다양한 관점과 사용자의 에너지 유효성을 나타낸다.

본 논문에서는 건물 에너지 시뮬레이션에서 재실자 행동 모델링으로 인한 성능 차이의 원인을 분석하였다. 연구 결과는 재실자의 시간적, 공간적, 행동적 다양성을 고려하는 것이 건물 에너지 시뮬레이션의 정확성과 효과성을 향상시키는데 중요함을 강조한다. 앞으로의 재실자 연구는 이러한 결과를 바탕으로 재실자 다양성을 포함한 더 신뢰성 있고 재현 가능한 재실자 모델을 개발할 수 있을 것이다. 이를 통해 건물의 실제 에너지 사용과 시뮬레이션 결과의 일치를 높일 수 있는 재실자 행동 이해에 기여할 것으로 기대된다.

주요어: 재실자 행동, 재실자 다양성, 주거 건물, 재실자 모델링, 건물 에너지 시뮬레이션

학번: 2021-27940