A data-driven approach to support the automation of thermostats in residential buildings

Mozhdeh Bertina

A Thesis

In

the Department

of

Building, Civil, and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Applied Science (Civil Engineering) at

Concordia University

Montreal, Quebec, Canada

August 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Mozhdeh Bertina

Entitled: A data-driven approach to support the automation of thermostats in residential buildings

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Civil Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Fuzhan Nasiri

_____ Examiner
Dr. Fuzhan Nasiri

_____ Examiner
Dr. Ursula Eicker

_____ Co-Supervisor
Dr. Fariborz Haghighat

_____ Co-Supervisor
Dr. Karthik Panchabikesan

Approved by _____

_____ 2023 _____

Abstract

A data-driven approach to support the automation of thermostats in residential buildings

Mozhdeh Bertina

Programmable thermostats represent a significant advancement in home automation technology, offering the potential for maintaining comfort and energy efficiency. However, the frequent overriding of default schedules indicates the necessity of flexibility to accommodate the dynamic occupant behavior and requirement. This thesis delves into this challenge, leveraging data-driven insights to understand thermostat override behaviors and hence develop supportive automation strategies that minimize human interaction. The introductory focus of this research lies in examining how individual comfort preferences, outdoor conditions, and daily schedules influence thermostat override behaviors. The data set for this exploration comprises thermostat and occupancy data from two residential buildings in Quebec, Canada, equipped with ecobee smart thermostats from the heating and cooling seasons of 2017 to 2019. The research subsequently explores the frequency of override behaviors across different Heating, ventilation, and air conditioning (HVAC) modes, schedules, temperatures, and years.

A key novelty of this research lies in its extensive exploration of occupancy, temperature, and setpoint trends over specific periods, facilitating the identification of patterns in thermostat override cycles and daily adjustments. Machine learning algorithms, such as decision trees and random forests, are employed to ascertain the importance of various features influencing thermostat override behaviors. Association rule mining techniques then reveal the relationship between variables, suggesting adaptive automation strategies based on temperature, occupancy, time, and outdoor conditions.

After conducting a comparative data analysis for two households, we identified significant shifts in occupant behavior and temperature preferences. From these insights, we have derived four various automation strategies: temperature-based, occupancy-based, outdoor temperature-based, and time-of-day and weekday-based. These strategies exemplify the adaptability in occupant

behaviors. Recognizing the factors that influence thermostat overrides makes it possible to equip smart thermostats with more intuitive automation strategies. These strategies can proactively adjust settings in line with user behavior and prevailing outdoor conditions, enhancing comfort and energy efficiency. To further fine-tune and widen the applicability of these strategies, it would be beneficial to conduct additional research with more extensive and diverse datasets.

ACKNOWLEDGEMENTS

First and foremost, I express my most profound appreciation to my supervisor Dr. Fariborz Haghighat for his continuous support, patient guidance, and encouragement. I am very grateful for the opportunity to work under his supervision and for his substantial effort and guidance during this research period.

Secondly, I would like to extend my deepest gratitude to my family. My parents, whose endless love and constant support have been a bedrock throughout my life, deserve special mention. I am also profoundly grateful to my husband, whose unwavering support, patience, and love have made this journey possible and enjoyable. His companionship and understanding, especially during those late nights and stressful moments, have been my source of strength and motivation.

Furthermore, I would like to thank Dr. Karthik Panchabikesan for his guidance and providing the data for this study. His invaluable contribution has been instrumental in this research's success, and I sincerely appreciate his support.

I extend my gratitude to ecobee for generously providing the 'Donate Your Data' program dataset, which was instrumental in facilitating my research and analysis.

Last but not least, I want to express my gratitude to the dedicated scholars who have used their professional expertise to reduce building energy consumption, enhance human comfort, and combat global warming.

Contents

List of Figures

List of Tables

List of Abbreviations

ARM	Association Rule Mining
BEMS	Building Energy Management Systems
CART	Classification and Regression Trees
CTs	Connected Thermostats
DLC	Direct Load Control
DR	Demand Response
DYD	Donate Your Data
EDA	Exploratory data analysis
GPS	Global Positioning System
HVAC	Heating, ventilation, and air conditioning
MURB	High-rise Multi-unit Residential Buildings
PIR	Passive Infra-Red sensor
PMV	Predicted Mean Vote
RF	Random forest

Chapter 1:

1. Introduction

1.1. Background

Almost 20% of North America's energy consumption and more than 12% of its carbon dioxide emissions come from residential buildings. As a result, residential structures consume significant amounts of energy, which, in turn, result in emissions. This energy consumption is primarily attributed to space conditioning, which regulates room temperature and air quality via an HVAC system. Typically, a thermostat is deployed to manage the HVAC system. Although thermostats can control much energy, the current control method is still reactive. Therefore, it is up to well-informed households to take responsibility for minimizing waste, such as energy, emissions, or cost (Huchuk, Sanner, & O'Brien, 2019).

The technology behind residential thermostat control has remained unchanged since its inception in the late 1800s. Programmable thermostats have been the primary option for decades, allowing users to set temperature schedules for periods of occupancy, absence, and dormancy (such as during sleep). However, studies conducted in the field have revealed that users were unable or unwilling to use the programming features of thermostats. Additionally, it was found that programmable thermostats did not result in any energy savings (Pigg & Center ofWisconsin, 2000). The evolution of thermostat control methods has been slow, with uncertain savings. However, user expectations and functionality have steadily increased(Peffer, Pritoni, Meier, Aragon, & Perry, 2011).

The latest iteration of thermostats, connected thermostats (CTs), can be integrated with the internet and accessed through various channels such as web, mobile, voice, and smart-home systems. In addition, connected thermostats are equipped with advanced capabilities to detect factors such as room temperature, motion, and location and can interact with other home products and services. These functionalities enable data collection, transmission, and receipt from the thermostat deployed in the client's domicile.

Using CT data presents numerous opportunities to enhance the management of residential buildings and houses. By analyzing this data, valuable insights can be gained, which can be utilized to make suggestions for performance improvement. Additionally, the information can be used to

create customized controls adaptable to the specific thermostat, house, or user preferences and behaviors. By improving the controls based on occupancy patterns, achieving a 30% reduction in energy consumption in the United States may be feasible (Pang et al., 2021).

However, there needs to be more comprehensive research on effective methods for utilizing the vast amount of data available to generate dependable population-wide recommendations and tailored solutions on peripheral devices situated away from centralized computing resources.

1.1.1 Motivation:

Optimizing residential building performance has been a recurring subject of discussion within the academic community. However, a unique opportunity has arisen to substantially enhance Households' operational efficiency. This opportunity is primarily driven by three key factors:

1. Before the introduction of CTs, there needed to be more historical data on thermostats. As a result, researchers had to rely on additional methods, such as deploying sensors and conducting surveys, questionnaires, or interviews, to obtain data for their evaluations. This time-consuming process often resulted in smaller sample sizes, limited geographic regions, and shorter periods. However, with CTs, there is now an abundance of high-quality historical data available for each house at a very detailed level. This enables researchers to understand better the entire home system, including the users, building, and environment. It also provides opportunities to customize the thermostat for each house(Meier, Aragon, Peffer, Perry, & Pritoni, 2011).

2. Recent progress in data science and machine/deep learning, coupled with more readily available implementations, has led to notable expansion and exploration of new data-driven techniques. These have brought about significant transformations in various industries. Additionally, advancements in faster and less expensive technology have made it more feasible to process large swathes of data. Moreover, embedded devices have now been endowed with the capacity to handle more computations required for model training and prediction. This further broadens the scope of solutions that can be devised and tested.

3. As the automation and interconnectivity of residential buildings continue to advance, managing their systems efficiently has become increasingly challenging. While traditional dead band controls are adequate for predictable schedules and fixed costs, they must catch up regarding

complex cost structures such as energy costs, greenhouse gas emissions, and comfort. In addition, more sophisticated control decisions are required as the thermostat becomes just one component in a network of devices that must be coordinated to achieve potentially conflicting objectives.

Researchers have dedicated their efforts to comprehending how individuals operate thermostats within their households for many years. However, previous data collection methods have yet to allow in-depth analysis of personalized setpoint schedules. Fortunately, with the introduction of smart thermostats, researchers now possess access to a continuous longitudinal stream of data about thermostat usage. This presents an exciting opportunity to enhance energy efficiency within residential structures through advancements in data science, machine learning, and more sophisticated control decisions.

1.2. Objectives

This research aims to create a systematic methodology using a data-driven approach to support the automation of programmable thermostats considering occupant's override behavior. The framework will be tested in two residential apartments in Quebec, Canada, using ecobee smart thermostats for a year. The two households were selected based on the distribution of overrides from 10 Households. The study aims to achieve the following objectives:

- Developing a methodological framework to extract the behavior patterns of individual households concerning their smart thermostat override activities.
- Proposing helpful strategy to support thermostat automation informed by the frequent override events.

The investigation aims to ascertain the feasibility of automating smart thermostats by observing occupants' override behavior. Furthermore, analyzing the occupants' override behavior would allow for a tailored temperature control system, which may enhance user satisfaction. To achieve these objectives, the study will focus on analyzing the patterns and trends in override activities of individual households using ecobee intelligent thermostat data, emphasizing factors that influence the occupants' override behavior.

1.3. Organization of the thesis

This study is structured into five sections. Chapter 1 serves as the introduction and delineates the goals of this thesis. Chapter 2 surveys existing literature to pinpoint principal strategies for smart thermostats, occupant behavior, and data mining techniques. Chapter 3 outlines the dataset incorporated in this examination and thoroughly explains the methods implemented to realize the research goals. Chapter 4 presents this analysis's findings and assesses the predictive models' performance. Lastly, Chapter 5 offers a summary, conclusions, and key takeaways and suggests directions for subsequent research.

Chapter Two

2. Literature review

2.1 Introduction

The rising trajectory of energy consumption in buildings has emerged as a significant global concern. Buildings contribute to over one-third of worldwide energy consumption, and this figure is anticipated to experience a substantial increase in the future (Policy, 2013). Addressing this challenge, the utilization of high-level controllers in building automation systems has been proposed to improve building performance and conserve energy (Palensky & Dietrich, 2011). In advanced building control systems, occupancy information is essential (Ahmad, Mourshed, & Rezgui, 2017). Occupancy information is essential for applications such as automatic lighting control (Casals, Gangolells, Forcada, & Macarulla, 2016) and building conditioning (Erickson, Carreira-Perpiñán, & Cerpa, 2014). In addition to the mentioned applications, using occupancy information in thermostat control for regulating a setback temperature during unoccupied hours is considered a practical energy-saving approach (Shen, Newsham, & Gunay, 2017).

According to (Peffer et al., 2011), residential thermostats account for approximately 9% of the total energy consumption in the United States. These devices regulate cooling and heating systems to maintain a comfortable indoor temperature. Three distinct thermostat types are available on the market: programmable, non-programmable, and smart thermostats. Non-programmable thermostats necessitate manual adjustments for changing the indoor temperature, while programmable thermostats let occupants set various temperatures at different time periods. This feature enables thermostats to self-adjust the temperature based on user settings, making them more convenient and energy-efficient.

Modern smart thermostats now offer advanced features that were once exclusive to programmable thermostats, including occupancy-based control. These intelligent devices utilize advanced technology to provide more efficient and convenient heating and cooling solutions for households and are becoming increasingly popular among homeowners. In addition, smart thermostats can connect with Wi-Fi and mobile phone applications, enabling users to operate them remotely. In addition, it is widely accepted that programmable thermostats are energy-efficient as they automatically regulate setpoint temperatures (Meier et al., 2011).

2.2 Occupancy resolution levels

The study conducted by (Melfi et al., 2011) has identified four distinct levels of occupancy resolution in three dimensions, as illustrated in Figure 1. These levels range from the direct detection of the presence of occupants to the identification of their activities at every time step. The information gleaned from such observations can be precious for various building applications, including but not limited to Building Energy Management Systems (BEMS), parking management, space management, and emergency response. However, it is essential to note that different applications may require different levels of occupancy resolution. Unfortunately, the definition of "occupant information" is not standardized, leading to significant variations in data collection ranges.

To address this concern, (Liu et al., 2015) proposed two additional levels of occupancy resolution and reorganized the existing levels based on their relevance to building energy consumption. The resulting framework comprises six distinct levels:

- Level 1 - Occupancy Presence: This level detects whether occupants are present in a particular zone using conventional sensors such as PIR. This information can be utilized to operate devices like intelligent lighting systems, thereby conserving energy.
- Level 2 - Occupant Location: This level focuses on determining an individual's location within the building, which can be achieved through nonintrusive load-monitoring algorithms or GPS. The collected data can be utilized to regulate the HVAC system to provide a comfortable environment for the occupants.
- Level 3 - Occupant Number: This level entails detecting the number of individuals in a particular zone, achievable via sensors such as PIR, ultrasonic sensors, or more sophisticated devices like Wi-Fi equipment and cameras.
- • Level 4 - Occupant Activity: This level involves recognizing the activities occupants are engaged in, which can assist in determining the acceptability of the indoor thermal atmosphere.
- • Level 5 - Occupant Identity: This stage focuses on the occupants, encompassing their facial characteristics, personal computer addresses, and mobile accounts.
- • Level 6 - Occupant Track: At this final stage, the movement paths of occupants across the building's various zones are tracked. This can be utilized in the creation of anticipatory comfort systems.

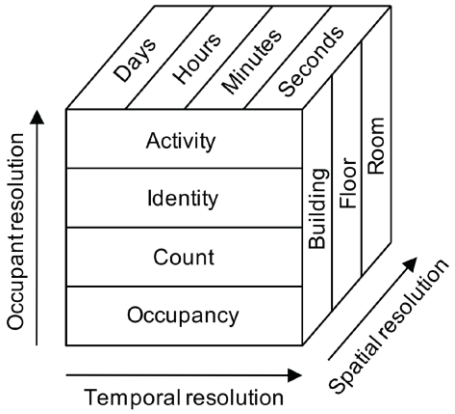

Figure 1-1:Occupancy resolution in three dimensions (identified by (Melfi et al., 2011))

2.3 Occupant behavior and thermostat usage

In cold climates, one approach to conserve energy during the heating season is to program a lower setpoint temperature when one intends to be away from home. This method could decrease the runtime of heating equipment compared to manually adjusting the temperature. However, there exist varying perspectives on the effectiveness of this energy-saving strategy, as some studies offer support while others do not. (Peffer et al., 2011) have reported that several studies have indicated that households equipped with programmable thermostats tend to consume more energy. Nevertheless, (Meier, 2010) have also suggested that the extent of energy conservation achieved through such thermostats is contingent upon how the occupants are programmed and regulated.

Research has indicated that occupant behavior regarding thermostats is critical in achieving building energy efficiency and meeting energy saving goals. For example, (Urban & Gomez, 2013) study shows that discrepancies in thermostat settings could result in uncertainty in building energy models. In addition, (Moon & Han, 2011) emphasized the importance of comprehending the range of ways occupants utilize thermostats to enhance energy efficiency in residential spaces. Therefore, understanding how occupants interact with thermostats is essential for ensuring dependable energy reduction targets and effective energy modeling.

The domain of mechanical systems, encompassing furnaces, compressors, and heat exchangers, has been subject to extensive research. However, there needs to be more investigation into the interaction between occupants and thermostats, as (Meier et al., 2011) noted.

To provide a more holistic perspective on building interfaces, (Day et al., 2020) have emphasized the need for comprehensive research on thermostats, windows and other appliances to gain insight into their control and design logic's impact on usability energy efficiency. However, additional research is required in this area due to the limited number of observational studies and inadequate sample sizes in current research.

Gaining insight into individuals' behavior when operating thermostats can be complicated. (L. Yan, Liu, Xue, & Zhang, 2020) have identified a limitation in programmable and non-programmable thermostats in terms of their inability to access and record real-time thermostat data. This limitation directly impacts data acquisition and should be considered when evaluating the effectiveness of these types of thermostats. As a result, initial studies primarily depend on data that individuals report themselves. (Pritoni, Meier, Aragon, Perry, & Peffer, 2015) suggest exercising caution when approaching low-granularity datasets, as they may be unreliable and contain flaws that could affect their accuracy. For example, (Parker, Sutherland, Chasar, & Center, 2016) encountered limitations in their study due to the need for actual setpoint data and were only able to monitor hourly indoor temperature, which was then assumed to be the setpoint temperature. This may have reduced the validity of their study. Furthermore, some studies and pilot reports have limited monitoring periods, typically lasting only one or two months, during which only representative months, such as the coldest and hottest months, are selected for analysis. However, these short-term monitoring analyses may not accurately represent occupants' long-term patterns of thermostat usage. As such, it is essential to acknowledge the limitations of the data and studies when interpreting and analyzing their findings. Understanding occupant behavior can be complex. According to (Urban, Elliott, & Sachs, 2012), it can be challenging to determine the energy-saving advantages of thermostats because the behavior of the occupants is unpredictable. This, in turn, limits the effectiveness of setback options in programmable thermostats, as noted by (Parker et al., 2016). Moreover, (Méndez et al., 2020) have found that improper usage of connected thermostats can lead to increased energy consumption. This. can be attributed to occupants' inadequate understanding of HVAC systems and their limited knowledge regarding thermostat settings and environmental impact. A study conducted by (Moon & Han, 2011) investigated the effect of thermostat settings on energy saving for residential buildings.

Using energy simulation models, they determined that the most effective approach is to utilize appropriate setback and setpoint temperatures and setback periods. These findings hold significant implications for reducing energy consumption and promoting sustainability in the built environment. This source considers a limited range of situations; in reality, the potential thermostat setting options exceed those examined in this context. The final component involves the approach. Analyzing occupant behavior requires a multidisciplinary effort, encompassing data mining, human-machine interaction, statistics, machine learning, and more. As a result, it is crucial to incorporate methodologies from various fields to explore these complex behaviors. For example, data mining techniques, which involve extracting valuable insights and patterns from extensive data sets, have recently gained increasing prominence.

2.4 Data mining and statistical modeling of occupant behavior

Data mining is a proposed occupant behavior analysis and simulation method that can help uncover valuable information from large datasets. This information can then be employed to create statistical models that depict various occupant actions, including window manipulation, shade adjustment, and light switch usage (D. Yan & Hong, 2018). Additionally, (Hong, D'Oca, Turner, & Taylor-Lange, 2015) concur with the notion that data mining holds promise as a beneficial tool in occupant behavior modeling endeavors.

Furthermore, specific, measurable interactions between occupants and their buildings have been the subject of research using data mining methodologies and statistical models. As an example, window usage has been investigated through methods such as logistic regression and association rule mining (D'Oca & Hong, 2014), and similar techniques have been applied to the study of ventilation system operation (C. Zhang, Xue, Zhao, Zhang, & Li, 2019) in order to uncover patterns of behavior and the factors that drive these actions.

There are challenges in collecting detailed information about thermostat settings and other indoor environmental conditions. Fortunately, solutions to address these challenges are in progress. One solution that is gaining popularity is using smart thermostats, particularly internet-connected ones. For example, Canadian manufacturer ecobee has developed the 'Donate Your Data' program (DYD), allowing tens of thousands worldwide to anonymously and voluntarily share data from their thermostats. This data includes information on settings, indoor environmental runtime, and

equipment conditions. This program presents a unique opportunity to gather data on human-thermostat interaction on a large scale, which was previously a significant obstacle. Numerous research studies have used the ecobee database to analyze the dynamics involved in the interaction between individuals and thermostats, owing to the heightened accessibility of data.

The study conducted by (Huchuk, O'Brien, & Sanner, 2018) analyzed dataset submitted by 2500 users from two distinct viewpoints. Firstly, the study examined the differences in thermostat behavior based on seasonal changes, varying climates, and utility rates. Secondly, it explored the possibility of categorizing users based on their thermostat usage patterns. The results demonstrated that seasonal changes and varying climates influenced users' thermal preferences but not utility rates. Furthermore, the data did not reveal any identifiable user categories.

A recent investigation by (Stopps & Touchie, 2020) examined residents' thermal satisfaction and responses in a pair of newly constructed high-rise multi-unit residential buildings (MURBs) in Toronto, Canada. The researchers assessed various elements, including deviations in indoor temperature from the thermostat's designated level, the use of windows, and supplementary heating and cooling systems, to gauge the comfort levels experienced by occupants and those estimated by the Predicted Mean Vote (PMV) model. Upon evaluation, the researchers determined that there was an inconsistency between the PMV model's predictions and the comfort levels reported by the residents. Furthermore, the study revealed that over-conditioning was a common issue in these structures, occurring 35% of the time during heating and 23% during cooling periods. The air conditioning systems within the buildings also displayed an exceptionally brief operational duration, with the most extended runtime not surpassing 16 minutes per hour. The findings of this study have significant implications for the management of building operations.

In an investigation conducted by (Huchuk, O’brien, & Sanner, 2020), the focus was placed on smart thermostats' effectiveness in bypassing predetermined schedules. An analysis of data from more than 20,000 households was performed to delve deeper into the frequency and variations of overrides used and their consequent influence on energy consumption. The widely held belief that user-generated holds pose a considerable problem and are a primary factor in unsatisfactory programmable thermostat performance might not be wholly justified.

The study suggests that user-holding actions may not be as damaging to aggregate energy savings as conventionally believed, considering that many thermostats experienced rare overrides, and the

limited duration of each override contributed to only a minimal increase in energy consumption. Furthermore, a mere fraction of thermostats, under 5%, displayed suboptimal energy performance due to persistent overrides.

The study's findings imply that there is potential for devising effective strategies to refine thermostat user interfaces to discourage the implementation of prolonged energy-inefficient hold settings. The occurrence of overrides was approximately 30-35%, a lower percentage than what previous research on programmable thermostats has indicated. The study also differentiated between distinct user categories, such as "frequent holders" and "infrequent holders," among others. In the case of "infrequent holders," examining override initiations about contextual factors revealed that the frequency of holds varied depending on specific situations.

In 2021, research was conducted by (H. Stopps & M. F. Touchie, 2021) on two high-rise multi-unit residential buildings (MURBs) to examine the habits of occupants concerning thermostat set point and override settings while at home. Results from the study revealed that the setpoint temperatures were infrequently adjusted by more than 60% of residents, despite the variability in individual preferences. Furthermore, it was suggested by the investigation that holds behavior, which involves users overriding their thermostats, might not pose a significant concern due to the low frequency of overrides and the brief intervals between adjustments.

Nonetheless, it is imperative to recognize that the dataset only partially consists of infrequent users and the energy consumption implications for those who regularly exhibit hold behavior must be considered.

2.5 Smart thermostats and occupant behavior

Utilizing a methodology suggested by the Environmental Protection Agency (2016), (Huchuk et al., 2020)conducted a quantitative analysis of hold behaviors' energy consequences for a sample of 1500 ecobee users in Ontario, Canada. The study outcomes revealed that minimal energy consumption was linked to users who frequently enacted hold behavior and made thermostat adjustments, while users who rarely participated in hold behavior followed. In contrast, the highest energy consumption correlated with users who often held but scarcely adjusted thermostats. Importantly, no studies have been found that specifically address the energy implications of hold behaviors about high-rise residential buildings.

The ecobee dataset has undergone rigorous analysis utilizing advanced machine learning techniques. Machine learning involves training models to recognize data relationships, enabling them to make informed predictions or decisions based on the training they receive. For example, in a study conducted by (Huchuk et al., 2019), different machine-learning models were utilized to predict and analyze residents' occupancy of a space. Similarly, in another study by (H. Stopps & M. Touchie, 2021), predictive models based on decision trees were employed to anticipate the duration of equipment operation in high-rise residential buildings. In contrast, (Huchuk, Sanner, & O'Brien, 2022) used different models, including a time-series model, a linear model and a grey box model, to accurately predict indoor temperatures over a short term period. However, it is essential to note that these studies primarily focused on prediction rather than identifying patterns in how occupants interact with thermostats. A more comprehensive investigation is required to investigate residents' methods of interacting with thermostats through data mining techniques. For example, a study by (Xiaoxin Ren, Yan, & Hong, 2015)examined affordable housing during the heating season using clustering analysis to identify indoor temperature patterns. However, the absence of thermostat data made it challenging to determine the settings precisely. Another study by (Xinyuyang Ren et al., 2019)scrutinized data of room air conditioning units using data mining techniques. This investigation uncovered distinct occupant behaviors related to temperature-setting and established their connections with factors such as energy consumption per hour, operation duration, and total energy usage of the air conditioning systems.

(Tomat et al., 2022) studied user interactions with smart thermostats in the context of Demand Response (DR) events. Users are categorized into distinct groups through clustering techniques, and specific user behaviors that could undermine the effectiveness of the Direct Load Control (DLC) strategy are identified. Consequently, the study proposes the necessity for personalized DR events catering to diverse user types. The research is grounded in real-world data from the Donate Your Data dataset, which investigates user engagement with smart thermostats during DR events.

The authors employed clustering techniques to discern user categories based on their actions before and during the DR event. Furthermore, a building energy simulation tool was utilized to model various scenarios and assess the power reduction and energy impact that was not achieved. In conclusion, the authors advocate for customized DR events to improve the effectiveness of the DLC strategy and assert that specific user behavior could hinder its success. Additionally, the

distinct patterns revealed through clustering could be employed to develop tailored DR events for various user archetypes.

The study by (Deng, 2021) delves into occupants' behavior when utilizing smart thermostats and using various techniques, such as association rule mining, logistic regression, clustering analysis, and predictive models. To identify patterns in behavior, potential factors that drive these behaviors, and their impact on energy consumption. This paper's limitations include using a single dataset from a specific geographic location, which may limit the generalizability of the findings to other regions. Additionally, the study only focuses on hold behaviors and does not consider other factors affecting energy consumption, such as occupancy and building characteristics. Finally, the study does not investigate the reasons behind the hold behaviors, which may provide further insights into the occupants' decision-making process.

2.6 User comprehension

According to a recent study, the success of achieving savings in high-performance building designs is influenced by the occupants' understanding of the system and their comfort expectations (Day & Gunderson, 2015). However, as smart thermostats become more advanced, research suggests that elderly users may struggle to comprehend the features (Combe, Harrison, Craig, & Young, 2012). Furthermore, multiple research studies have demonstrated that existing programmable thermostats could be simpler for users. Consequently, it is of utmost importance to prioritize user-friendliness and accessibility when designing smart thermostat interfaces, with particular consideration given to elderly individuals and those with disabilities or impairments.

2.7 Gaps in the literature

After conducting a thorough literature review, it is clear that overriding smart thermostats remains a significant challenge to achieving energy efficiencies, despite technological advancements. This behavior is influenced by various factors, including comfort needs, outdoor environment, energy awareness, and the occupants' interaction with technology. In addition, it is important to consider human behavior when designing and utilizing these systems, as individuals, in general, prioritize personal comfort over energy efficiency, even with advanced automation. The literature shows opportunities for more personalized intervention strategies, especially considering the varying impact of interventions on different demographics and settings. Relating the findings to

automation, the smart thermostat technology has to evolve from fixed, pre-programmed schedules to more adaptive and dynamic models that can learn from previous override behaviors. In summary, prior studies using smart thermostat data have focused on prediction tasks like forecasting equipment runtimes or occupancy. There needs to be more emphasis on discovering granular behavioral patterns and modeling user interactions with thermostats. This thesis helps fill this gap through detailed data mining to uncover occupancy override patterns. Most analyses have been limited to one household or a single year of data. The comparative analysis across different households and years for having a more accurate result is a novel contribution of this thesis, providing insights into evolving behaviors. While some works have proposed general automation strategies, this thesis uniquely tailors the recommendations to each home based on discovered override patterns. This personalized approach is novel and contributes to advancing automation. This study utilized the Python programming language for all data analysis steps. Specifically, packages such as pandas for data manipulation, numpy for numerical operations, matplotlib and seaborn for data visualization, scikit-learn for machine learning, and statsmodels for statistical models were employed.

Chapter Three

3. Methodology

3.1. Data Description and understanding

The data utilized in this research was gathered from smart thermostat users who voluntarily agreed to share their thermostat usage through ecobee Inc.'s 'Donate Your Data (DYD)' program. The dataset comprises five-minute interval data collected from the thermostat and sensors throughout the Households, along with user metadata that details the house's characteristics. Moreover, for each home, outdoor weather data from the nearest weather station is also included at the same level of detail as real-time thermostat data, with 5-minute intervals. Table 2-1 contains information regarding the DYD dataset.

Table 2-1: The interval dataset by ecobee DYD

Data Point	Description	Units
Indoor air temperature from remote sensor(s)	Temperature readings from remote sensors	°F
Outdoor air temperature (T_out)	from local weather station	°F
Outdoor relative humidity	from local weather station	%
Temperature setpoints (T_stp_heat), (T_stp_cool)	Heating and cooling setpoint temperatures	°F
Control Temperature (T_Ctrl)	Indoor temperature used by the thermostat to compare against setpoints. It represents a combination of temperatures across the home to reduce energy or increase comfort.	°F
HVAC mode	Heat/Cool	NA
Motions	State of occupancy detected by PIR sensor	boolean
Schedule Options	User-defined comfort period (e.g., home, away, sleep, etc.)	NA
Event Options	Items that override the set schedule (e.g., holds, vacations, demand response events, etc.)	NA

In addition to providing five-minute interval data, each thermostat is equipped with corresponding metadata. This metadata contains pertinent details regarding the device's location, including the

city and state/province, as well as information regarding the characteristics of the building, such as its floor area, age, and type. Additionally, information is available regarding the HVAC

equipment installed on the property, including whether or not a heat pump is present and how auxiliary heating should be utilized. Table 2-2 contains information regarding the available buildings for this research.

Table 2-2:General information regarding the available buildings

Identifier	Country	Provic e State	City	Floor Area [ft2]	Style	Number of Floors	Number of Occupan ts	Auxiliar y Heat Fuel Type	Number of Remote Sensors
1	CA	QC	St-Lambert	1500	detached	1	6	Electric	9
2	CA	QC	L'Ile-Bizard	2500	detached	2	4	Gas	4
3	CA	QC	Anjou	3000	detached	3	4	Gas	5
4	CA	QC	Gatineau	1500	detached	2	6	Gas	4
5	CA	QC	La Prairie	3000	detached	3	4	Gas	3
6	CA	QC	St-Luc	1500	detached	1	2	Electric	3
7	CA	QC	Vaudreuil-Dorion	2500	detached	2	4	Gas	3
8	CA	QC	Cap-rouge	4000	detached	2	4	Electric	3
9	CA	QC	Brossard	2000	detached	2	5	Electric	3
10	CA	QC	st-bruno	2500	detached	2	3	Gas	5
11	CA	QC	Laval	1500	detached	2	2	Electric	1
12	CA	QC	Mascouch e	3000	detached	3	4	Electric	1
13	CA	QC	Laval	2500	detached	2	5	Electric	1
14	CA	QC	Laval	2000	detached	3	2	Gas	3
15	CA	QC	brossard	2500	detached	3	5	Electric	1

3.2 Data availability

The data availability for each dwelling was based on the user's enrolment in the DYD program. The data collection's start and end dates for the ecobee dataset used in this study were from the heating and cooling seasons of 2017-2018 to 2018-2019. Therefore, data for each dwelling represented a subset of the abovementioned period. For example, some dwellings have the data measured from September 2015 till September 2019, whereas some are only available for 2019. The datasets used in this study belong to two residential buildings in Quebec, Canada. Out of the

fifteen houses available in the dataset, only two were selected for analysis based on the distribution of thermostat overrides. This selection was driven by the study's primary objective, which was the provision of individualized rules for distinct households. Given the depth and detail required for such individualized analysis, it was deemed that focusing on just two households would be sufficient to fulfill the research aims. Table 2-3 describes the general information about selected household. The dwellings selected for this study possessed sufficient data for analysis collected during the heating and cooling seasons of 2017-2018and 2018-2019. Employing this methodology made the inclusion of homes that offered the most valuable information for our research feasible. The two households were selected based on the number of overrides and data availability.

Table 2-3:Selected residential units with the number of overrides

Identifier	City	Floor Area [ft2]	Style	Number of Floors	Age of Home [years]	Number of Occupants	Number of Remote Sensors	Number of Override in Heating mode	Number of Override in Cooling mode
3	Anjou	3000	detached	3	10	4	5	1137	29854
5	La Prairie	3000	detached	3	5	4	3	5312	6043

3.3. Methodology

Figure 2-1 represents a schema developed with a data-centric approach, aiming to meet the goals set for the present research. Two residential buildings provide the requisite data sets for this exploration. The data is available for the years 2015 to 2019. For this study, the heating and cooling seasons of 2017 to 2019 were considered. The suggested structure is independently implemented for each building's data to see whether it can be generalized to households with different characteristics. The thermostat will select the appropriate HVAC mode based on the detected season - heat for colder and cool for warmer months. This decision is not just about on/off but also about maintaining an optimal balance of indoor temperature, humidity, and energy use. In the initial stage, the HVAC Mode (heat or cool) is selected based on the season under scrutiny, whether heating or cooling. Following the selection, the data preprocessing was performed. This involves cleaning data and extracting pertinent features such as occupancy state, temperature settings, and event schedules. After preprocessing step, the frequency of occupancy overrides was calculated. This calculation provides insight into the behavior patterns of the occupants and their impacts on the indoor temperature settings and comfort levels. An analysis is conducted to determine average temperature differences for 'Event' in different 'Schedule' categories to understand the effect of overrides on indoor temperature settings and occupant comfort. The dataset is then aggregated to hourly intervals to provide a more concise representation of the data. During this process,

categorical columns are consolidated using the mode method, while numerical columns are aggregated using the mean value for that hour. An occupancy state (presence/absence) analysis is performed by calculating the median value for each hour for each building, which is then considered a threshold. The data is then transformed based on this threshold, and occupancy during nighttime hours is modified to account for typically low occupant movement. A temperature analysis is conducted to identify trends, peaks, and troughs in indoor and outdoor temperatures during the heating season. Additionally, 'Hold Cycles' are analyzed to identify periods when the schedule is overridden, and these 'Hold Cycles' are calculated for each day. Subsequently, feature importance is evaluated by applying machine learning techniques such as Decision Trees and Random Forests (explained in detail later in section 3.9) to identify the key factors affecting override actions. Furthermore, association rule mining is applied to discover relationships between occupancy states, temperature settings, and override events. To understand the evolution and potential changes in occupant behavior and the effects on indoor temperature settings, a comparative analysis of data from two heating seasons, 2017-2018 and 2018-2019, is conducted. This comparative analysis aims to identify any significant changes in the various factors influencing energy consumption over the two years. In addition, this comparison aids in refining predictive models and provides recommendations for optimal thermostat settings to balance energy efficiency and occupant comfort.

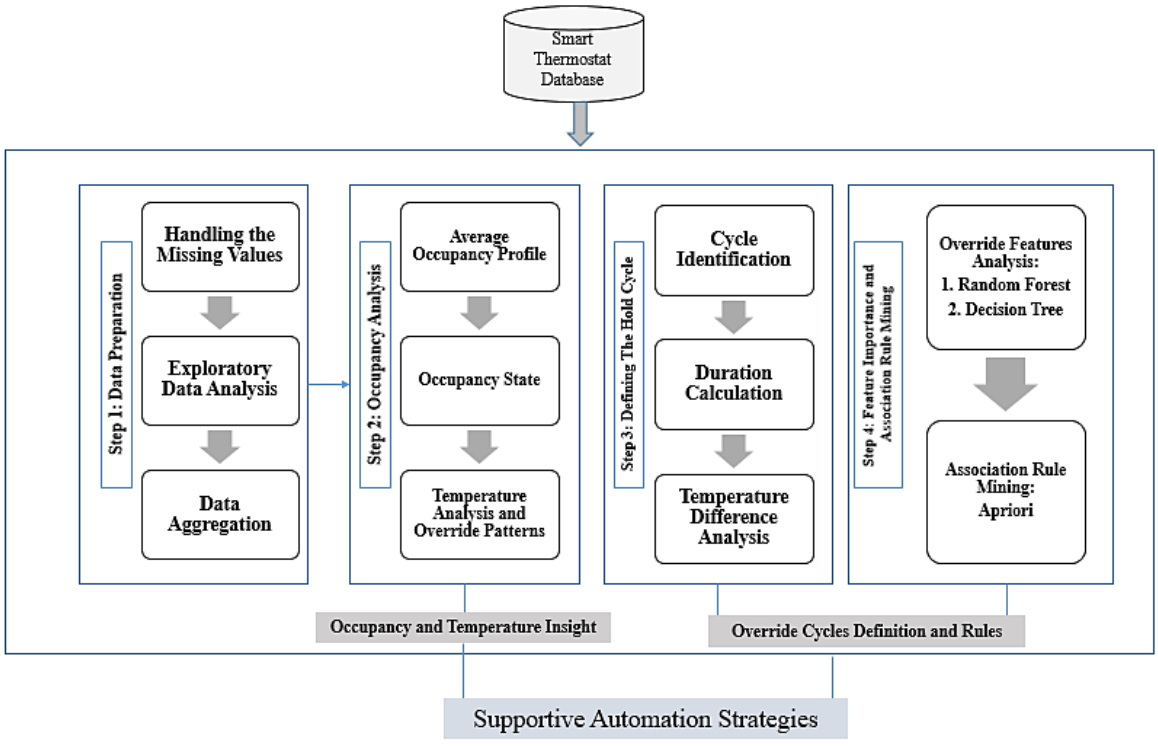

Figure 2-1: Methodology Framework

3.4 Data preparation

Data imperfections can occur due to various factors, such as human or mechanical errors, resulting in noisy, missing, or inconsistent data (Panchabikesan, Haghighat, & El Mankibi, 2021). Before the application of data mining technology, data preprocessing significantly eliminates noise and inaccurate data. The source data may initially contain missing values and outliers, which, if addressed, could positively impact the accuracy of predictions. Furthermore, given that the variables in the raw data may vary in scale, utilizing features with varying scales may not contribute equally to the analysis. Therefore, data cleaning constitutes the primary phase in data preparation.

3.4.1. Handling the Missing values

The dataset initially presented temperature values in Fahrenheit degrees. However, as part of our data preprocessing steps, all temperature values were first converted into Celsius degrees to maintain a consistent unit of measurement across the study.

Moreover, the data selection was stringent to ensure the most complete and accurate picture of daily temperature and occupancy variation. Specifically, only those days with a full 24-hour record of thermostat data were included in our study for each dwelling. This approach allowed to maintain a consistent daily timeline for our analysis, thereby enhancing the reliability of our findings.

The next step was handling the missing values. Addressing missing values is crucial in analyzing the thermostat dataset within the research for several reasons:

1. It ensures the findings are based on comprehensive data, enhancing the study's reliability. Missing values can distort the data's trends, patterns, and relationships, leading to inaccurate conclusions.
2. Treating missing values improves the study's statistical power by retaining as many cases as possible.
3. It allows for accurate modeling and prediction in the dataset, which can be particularly critical when dealing with time series data or ecological models, as is often the case with the smart thermostat dataset.

In the first step, the 'Event' column might contain missing values, indicating instances when no specific event was recorded or overridden. To deal with these missing values and maintain data consistency, the 'fillna' function was used. This function fills the missing values or 'NaN' entries with a specific value. In this case, every missing value in the 'Event' column was replaced with 'No

Override'. This implies that where there were no recorded events, it is assumed that no override event occurred.

3.4.2 Exploratory data analysis (EDA)

Exploratory Data Analysis (EDA) plays a pivotal role in data analysis as it involves employing data visualization techniques to examine hypotheses and comprehensively understand the dataset at hand. Through EDA, researchers can uncover patterns, relationships, and critical insights within the data, enabling them to make informed decisions and draw meaningful conclusions. By visually exploring the data, EDA aids in identifying trends, outliers, and potential issues or biases that may impact subsequent analytical processes. It is an essential preliminary step before applying more advanced statistical or machine-learning techniques to the data. (X. M. Zhang, Grolinger, Capretz, & Seewald, 2018). Typically, EDA is carried out after data acquisition and preprocessing. The primary steps of EDA can be explained as below:

(1) detection of outliers; (2) comprehension of the database structure; (3) preliminary selection of suitable models; (4) extraction of essential parameters by uncovering the relationship between variables; and (5) visualizing potential relationships between variables and outcomes.

EDA encompasses graphical and non-graphical methods. Graphical techniques include histograms, boxplots, scatterplots, line plots, and heat maps. Non-graphical methods involve statistical tests, summary statistics, and tabulation. (DuToit, Steyn, & Stumpf, 2012).

Boxplots summarize data distribution features and display essential statistics, while scatter plots visualize relationships between two variables, revealing correlations, linearity, and trends. (Sandels, Widén, Nordström, & Andersson, 2015). Boxplots provide concise information about the central tendency, skewness, symmetry, and outliers of a variable. They are particularly useful for comparing the characteristics of multiple groups of data using side-by-side boxplots (Tukey & Tukey, 1985).

The frequency of occupancy overrides was systematically investigated, conducted individually for each household, and distinguished between the heating and cooling seasons in our research. The distribution pattern of these overrides within our dataset was grasped through this approach.

Furthermore, the average setpoint temperatures during different events and schedules were also explored in-depth. The average heating setpoint temperature during the heating season and the average cooling setpoint temperature differences for the cooling season were calculated. These were organized according to the 'Event' and 'Schedule' categories and sorted into bins based on three-degree outdoor temperature increments.

Box plots were used to visualize the discrepancies in average setpoint temperatures during override and non-override events. This represented the data's distribution, highlighting how average temperatures differed when occupants chose to override their pre-set schedules compared to when they did not.

The same procedure was also applied to analyze the thermostat temperature. The overarching goal of these analyses was to provide a deeper understanding of the differential impacts of override and non-override events on the average temperatures.

3.4.3 Data aggregation

Data aggregation was an essential step in preparing the data for analysis. In this study, data recorded at five-minute intervals were transformed into one-hour values to enhance the interpretability and understandability of the variables. During this process, the aggregation of categorical columns was performed using the mode method, which identified the most common value within each hourly period. This allowed for an uncomplicated interpretation of the general category within each hour. Conversely, for numerical columns, aggregation was executed using the mean value calculated over each hour, yielding an average representation of these figures throughout the specified time frame. This method's benefits were dual-pronged. Firstly, it ensured that the critical trends and features of the original data were effectively encapsulated in the resampled dataset. Secondly, it streamlined the data, making it more accessible for subsequent analysis and interpretation, thereby augmenting the efficiency of future stages of the study.

3.5 Average Occupancy Profile

Enhancements were made to the dataset by incorporating a new column termed 'total_sensor'. This column is specifically constructed to represent the aggregate sum of motion captured by all available remote sensors and the thermostat motion sensor. Calculated for each dwelling on an hourly basis, this column aims to offer a holistic perspective of the total motion activity within the monitored spaces.

An additional column, named 'total_sensor_avg' (the average of total sensors), has been introduced. This column, computed for each entry in the dataset, embodies the average motion detection value accrued from all remote and thermostat motion sensors. The purpose of calculating this average is to provide a representative insight into the standard level of motion activity across the monitored spaces.

These modifications to the dataset contribute to a more robust understanding of the patterns of motion activity within the dwellings under study. They enable more informed decision-making processes underpinned by a more comprehensive and nuanced dataset.

In order to have a thorough understanding of occupancy and setpoint temperatures, it is essential to consider monthly variations, differences between weekdays and weekends, and daily variations on weekdays, all based on the hour of the day. In addition, this approach provides valuable insights into how to override behaviors.

By examining each month separately, distinct patterns in occupancy and setpoint temperatures may emerge, reflecting changes in environmental conditions and energy usage behaviors. In addition, these patterns are exciting during seasonal changes, as they reveal how occupants adapt their behavior and thermostat settings in response to external temperatures.

Comparing weekdays with weekends adds another layer of complexity to behavior analysis. Generally, occupancy rates are higher during weekends when occupants are likelier to be home, influencing setpoint temperatures. Understanding these differences is crucial in predicting when overrides may occur and how they will impact overall energy consumption. Analyzing day-to-day variations in occupancy and setpoint temperatures provides further insights. For instance, occupants may have different schedules on Mondays than Fridays, resulting in varying occupancy patterns and influencing setpoint temperatures.

It is important to note that these trends are not static throughout the day. Hourly changes in occupancy and setpoint temperatures offer the most detailed insight into occupants' behavior and how it affects smart thermostat usage. Capturing these hourly variations is crucial to understanding when overrides occur and under what conditions. Figure 3-1 provides a sample of heat map depicting occupancy averages over time, showcasing the dataset's inherent patterns. The figure is for Home 3 during the heating season of 2018. Each cell in the heat map pertains to a specific hour and date, with color intensity representing average occupancy. The visual representation allows an immediate understanding of when and how frequently spaces were occupied. The figure reveals occupancy patterns across different times and days, shedding light on human presence within monitored spaces. For instance, in the provided heat map, differences in the absence and presence of occupancy during various hours of the day over a year are discernible. As anticipated, occupancy is observed to be significantly reduced during midnight hours. It can be attributed to the typical sleep patterns and decreased activity during these hours. By taking into account these finer-grained analyses, along with the Average Occupancy Profile, a complete picture of occupancy and setpoint temperatures can be achieved. This level of detail is essential in developing sophisticated and adaptive automation algorithms that can better predict and respond to override behaviors, leading to improved occupant comfort.

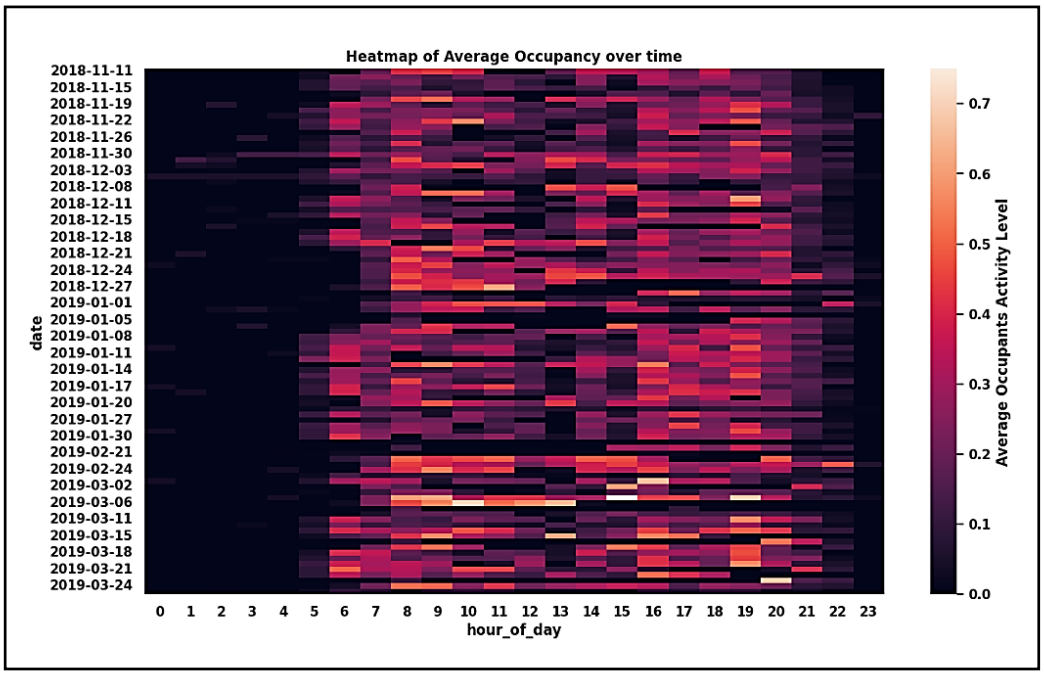

Figure 3-1: A Sample of Heat map for The Average Occupancy Over the Time-Home 3-Heating Season-2018

3.6 Occupancy State Analysis

Thermostat operation can be classified into various user tendencies based on their actions. These actions may include the treatment of occupied and unoccupied periods, the allowance of schedules to run, and the selection of setpoint temperatures. The initial focus of our exploration was to examine how users operate a thermostat differently during their occupied and unoccupied periods.

The dataset was subjected to careful transformation for occupancy level analysis. Each house's hourly median value was determined and utilized as a reference for categorizing the data into 'Occupied' and 'Unoccupied.' During a specific hour, if the average occupancy was higher than the median value, it was classified as 'Occupied.' Conversely, if it was lower than the median value, it was categorized as 'Unoccupied.'

However, the quieter hours between 10 pm and 7 am presented a challenge due to reduced movement. In response, a unique strategy was devised whereby any hour within this period labeled as 'Occupied' resulted in the entire night being classified as 'Occupied.'

The assumption was that occupants were likely present for the entire night if they were home for part of the night. Conversely, if all hours were labeled as 'Unoccupied,' it was assumed that the occupants were absent and the house was unoccupied.

3.7 Temperature analysis and override patterns

A temperature analysis was carried out to identify trends, peaks, and troughs in indoor and outdoor temperatures, control temperature, and setpoint temperatures during heating and cooling seasons. Simultaneously, an average occupancy value was computed for analysis purposes.

The experiment involved varying the conditions and characteristics of the holds, including different thermostat setpoints and durations. The holds were set at various outdoor and indoor temperatures, with varying motion patterns observed during the holds. A detailed analysis was conducted on each temperature's maximum, minimum, and mean values for override and non-override events.

3.8 Analyzing the Hold Cycle

3.8.1 Schedule Override

The ecobee thermostat allows users to customize their temperature schedule by setting unique times and temperature points for each day. Without a pre-established schedule, the device defaults to a cycle of sleep and home periods each day, governed by predetermined start and finish times.

The ecobee thermostat's scheduled periods, including sleep, home, away, or a personalized option, are linked to specific heating and cooling setpoints. The thermostat adheres to the temperature points associated with the current time slot in the schedule without requiring manual intervention, a common feature of programmable thermostats. However, the device adjusts setpoints anticipating schedule changes, a feature not common in standard programmable thermostats. This way, seamless temperature transitions can be ensured, corresponding to predicted schedule alterations. Consequently, the desired temperature is maintained in the home precisely during a scheduled shift, enhancing comfort and energy efficiency.

Furthermore, specific device versions allow minor adjustments (within a few degrees) to the planned temperature points based on detected occupancy or vacancy. User-initiated changes to the schedule, or "holds" in popular parlance, can be implemented provided the system is not in a deactivated mode (e.g., heating, cooling, or auto). These changes can be made directly on the device or remotely via a mobile device, web interface, voice assistants, or other third-party applications registered with the device.

Unfortunately, the DYD dataset did not provide insights into the method used to instigate a thermostat hold. The data captured at regular intervals demonstrated a hold being implemented, and the temperature points conveyed the new temperatures being regulated. The interval data confirmed the hold status every five minutes until the hold was lifted. During the hold, the device's other smart features are disabled, and the thermostat does not modify its temperature points following the schedule.

3.8.2 Duration of Override in Ecobee Smart Thermostat

The length of a thermostat hold depends on the user's actions and the preset preferences within the settings. Although a user can discontinue a hold at any moment, the thermostat's default

programming ensures that the override stays effective for a specific period. Duration options include

1. two-hour,
2. four-hour,
3. Until the next scheduled period begins,
4. an unlimited duration (the default choice), or
5. A prompt to the user to select their preferred duration (from the previously mentioned options) each time a hold is started or adjusted.

This user-specific preference is not included in the DYD project's data and is subject to changes made by the user within the thermostat settings. Owing to the lack of a dependable reference for user preferences on duration, an analysis was not undertaken concerning the reasons for terminating a hold or the duration of an individual hold.

3.8.3 Defining the Hold Cycle

Comprehending the duration of the override cycle, commonly referred to as a "hold" in the context of Ecobee's smart thermostats, is paramount for two primary reasons: personalized comfort and energy efficiency. Furthermore, it is imperative to grasp the occupants' hold behavior, which is crucial for optimizing user comfort and energy consumption.

Personalized Comfort: The override feature of the Ecobee smart thermostat empowers users to adapt their environment's temperature to suit immediate needs. For instance, an individual may desire a more relaxed environment during physical exercise or a warmer ambiance for relaxing activities like reading. The hold feature facilitates these modifications, allowing for temporary and immediate deviation from the thermostat's pre-set schedule. By adjusting the hold duration, users can tailor their interior temperature to correspond with their activities and comfort preferences in real time. Additionally, it is crucial to understand occupants' hold behavior, as it provides insights into their comfort preferences and how they interact with the thermostat's settings in various situations.

Energy Efficiency: The Ecobee smart thermostat is designed with features aimed at enhancing energy efficiency, including occupancy detection, a 'Follow Me' feature that modulates temperature based on occupied rooms, and an eco+ mode that adjusts heating and cooling

according to current humidity and electricity cost. These features operate based on a pre-set schedule intended for optimal energy usage. However, using an override or hold could disrupt this optimization. The ability to adjust the hold's duration so it only lasts until the next scheduled activity helps to minimize potential energy waste. This balancing act between user comfort and energy conservation underscores the intelligent design of the Ecobee smart thermostat. Furthermore, comprehending the duration and frequency of holds provides valuable insights into occupants' behavior and can be used to optimize energy efficiency strategies further.

The impact of override behaviors on energy efficiency remains a topic of ongoing debate within academic and industry circles. While some scholars argue that reducing overrides could lead to more significant energy savings, others suggest that the effects of overrides might be more nuanced, and their energy implications may not be as substantial as initially expected (Huchuk et al., 2020).

In order to further our research objectives, we have undertaken a meticulous analysis of the manual overrides within the realm of smart thermostats. To this end, we have established a "Hold" cycle as a discrete occurrence in which the user supersedes the previously programmed temperature settings, thus commencing a new period of temperature regulation.

1. **Cycle Start Identification:** Within each date-specific group, we iterate through the data to identify the start of Hold cycles. This process tracks the 'Hold' event in the dataset, marking the beginning of a new cycle.

2. **Duration Calculation:** Any non-Hold event in the dataset marks the end of a Hold cycle. The cycle duration is calculated in hours and stored upon encountering such an event.

3. **Handling Unfinished Cycles:** Some Hold cycles may extend until the end of the day. We add a specific check to ensure that these cycles are accounted for accurately. If a cycle's start is marked but not concluded within the day, the end of the day (timestamped at 23:59:59) is assumed as the end of that Hold cycle. The duration of such cycles is calculated accordingly and added to the list of durations.

4. **Cycle Count and Duration Recording:** At this stage, the cycle count is incremented, signifying the end of one Hold cycle and the potential start of another.

5. **Data Collation:** Finally, we collate each date's count and durations of Hold cycles into a new data frame. This information offers a comprehensive view of the Hold cycles, providing a thorough temporal analysis.

3.8.4 Temperature difference during the Hold Cycle

This section delves into two essential temperature differential measurements during the Hold Cycle: the average temperature difference between the control temperature (T_ctrl) and the setpoint temperature (T_stp_heat/T_stp_cool) and the difference between the indoor (Thermostat_Temperature) and outdoor (T_out) temperatures.

These measures offer significant insights into user comfort levels, preferences, and interactions with their thermostat.

1. Calculating the average temperature difference between control temperature and setpoint temperatures:

Observing the temperature difference between the control temperature (T_ctrl) and the setpoint temperature (T_stp_heat/T_stp_cool) can quantitatively measure user comfort levels and preferences. By statistically analyzing this temperature difference, it is possible to identify patterns in occupancy override behavior, serving as a proxy for user comfort. For instance, when overrides occur with considerable differences between the control and heating setpoint temperatures, it could indicate that the existing settings are not adequately satisfying the user's comfort needs.

The average temperature difference can be calculated using statistical measures such as the mean or median of the differences at each corresponding timestamp during the event. A more significant average temperature difference could suggest a greater discrepancy between the user's desired temperature and the control settings, necessitating more frequent overrides.

2. Calculating the difference between inside (Thermostat_Temperature) and outside (T_out) temperatures:

Another critical aspect to consider when analyzing user behavior regarding thermostat settings and overrides is the difference between indoor and outdoor temperatures. This calculation can reveal the relationship between these variables and their potential influence on occupancy override behavior.

The difference between these two temperatures can be calculated for each corresponding timestamp as (Thermostat_Temperature - T_out). Evaluating this difference can uncover trends in user preferences and comfort levels. Overrides are more likely to occur when the indoor temperature is significantly higher or lower than the outdoor temperature, indicating that users are trying to counteract the temperature difference to maintain their comfort levels. A more significant average temperature difference might indicate a greater need for the HVAC system to counteract external temperature influences, leading to more frequent overrides.

3.8.5 Categorizing the hold cycle based on the number of override

Categorizing the hold cycle based on the number of daily overrides can provide valuable insights into user behavior and the thermostat's operation. These overrides can occur from none to multiple times daily, reflecting the dynamic nature of user needs and environmental conditions. The number of these overrides during a day can be categorized as follows:

1. **Hold Cycle 0:** Cycles with no adjustments (no override during the whole day).
2. **Hold Cycle 1:** Cycles with one-time adjustment.
3. **Hold Cycle 2:** Cycles with two times adjustments.
4. **Hold Cycle 3 and beyond:** Cycles with three or more adjustments during the day, with the pattern continuing for further cycles.

3.9 Analyzing Feature Importance Related to Hold Events

In the context of smart thermostats and user behavior analysis, understanding the key factors or "features" that influence the occurrence of override is crucial.

Feature importance measures the relative contribution of different factors in predicting or explaining a specific outcome. Our focus lies on the factors that exert the most significant influence on the probability of a hold event.

3.9.1 Data mining techniques

This section gives a brief overview of two machine learning algorithms: DT (decision tree) and RF (random forest). The algorithms above were selected by two fundamental factors, namely their prevalence and heterogeneity level (Fan, Xiao, & Wang, 2014). The diversity in the application of

various studies and the resolution of different problems is derived from the distinct mathematical principles that underlie each algorithm. Each algorithm that is chosen possesses its own set of unique advantages and limitations.

1. Decision tree

Decision tree algorithms, such as the classification and regression trees (CART), provide an essential feature of importance scores. These scores are based on reducing the criterion used for selecting split points, such as Gini or entropy. This method is acknowledged in academic circles as a valuable tool for analyzing data and making informed decisions. In the context of a regression problem, decision trees are a powerful tool for feature importance analysis, which involves comprehending the most influential variables in predicting the target variable.

Decision trees split the data into distinct branches based on certain conditions or rules established using the feature variables. The feature that offers the most optimal split, as determined by a particular criterion, such as information gain or Gini impurity in classification tasks and variance reduction in regression tasks, is chosen as the root node, and this process is iterated on the resulting subsets. The recursive binary splitting mechanism employed by decision trees facilitates the establishment of a hierarchical structure for feature importance analysis.(Linero, 2018).

2. Random forest

The Random Forest (RF) algorithm is a robust machine learning methodology under ensemble learning strategies, leveraging the bagging technique more explicitly. This methodology harnesses the power of many decision trees, where each contributes to the final output (Scornet, Biau, & Vert, 2015). Depending on the problem, RF employs a majority voting system for classification tasks or an averaging mechanism for regression problems. These methods enhance the precision and generalization capabilities of the model.

Notably, each decision tree is constructed independently in the RF algorithm, introducing an additional level of randomness. This design aspect is instrumental in reducing the overall variance of the prediction model. Consequently, this RF feature also eliminates the need for extra pruning steps, a common practice aimed at improving model generalization and curbing overfitting.

When confronted with binary classification labels, the RF algorithm addresses the inherent instability issues associated with decision trees. Rather than relying on a single tree, RF generates a forest of trees, thereby enhancing the stability of predictions (Wang, Wang, Zeng, Srinivasan, & Ahrentzen, 2018).

An essential capability of RF is its inherent ability to evaluate and rank feature importance. This allows the model to identify which variables significantly influence label predictions, providing valuable insights into the data.

In this study, the RF algorithm was implemented using the Scikit-learn library, a popular Python library for machine-learning applications (Pedregosa et al., 2011). The choice of Python and Scikit-learn for this task underscores the accessibility and flexibility of these tools in conducting sophisticated machine learning analysis.

3.9.2 Performance evaluation

Both Random Forest and Decision Tree are widely employed machine learning algorithms that can be utilized for regression tasks. It is imperative to comprehensively evaluate the performance of these models upon training them on the designated dataset to understand their predictive capabilities concerning the target variable. The R-squared (R^2) score, an extensively utilized metric in regression problems, denotes the coefficient of determination. This metric indicates the proportion of the variance in the dependent variable that can be predicted from the independent variables. Notably, an R^2 of 100% indicates a complete explanation of changes in the dependent variable through changes in the independent variable(s). However, in most real-world problems, achieving an R^2 of 100% remains unattainable. The R^2 is determined using Equation 1 and ranges from 0 to 1 inclusive.

$$R^2 = 1 - \left(\frac{RSS}{TSS}\right) = 1 - [\,\Sigma(yi - \hat{y}i)^2 \,/\, \Sigma(yi - y\_mean)^2\,] \qquad \text{(Equation 1)}$$

RSS is the sum of squares of residuals, which are the differences between predicted and observed outcomes. TSS measures the total variance in the data. In a perfect model, TSS equals RSS

resulting in an R^2 value of 1. $\hat{y}i$ is predicted outcomes, yi is actual outcomes, and y_mean is the average of observed outcomes.

3.10 Association Rule Mining

Uncovering patterns in how occupants behave is a process that may differ based on the specific behavior being studied. Nevertheless, the overall approach for these types of investigations follows a standard format. One common approach is the accumulation of time-series data into discernable patterns, often in the form of motifs or clusters. For instance, research by (Capozzoli, Piscitelli, Gorrino, Ballarini, & Corrado, 2017) and (L. Yan et al., 2020) successfully used clustering methods to assemble time series data into diverse distinct groupings.

In another approach, (D'Oca & Hong, 2014) and (Xinyuyang Ren et al., 2019) Logistic regression models were employed to develop behavioral models utilizing data collected from occupant behaviors. Furthermore, (Funde, Dhabu, Paramasivam, & Deshpande, 2019) employed a technique known as motif discovery to pinpoint commonly recurring patterns in time series data.

Once these raw data have been processed and converted into patterns, they can be a foundation for more profound insights. Typically, the selection of methodologies is guided by the study's specific objectives, like devising a classification model that can sort new data according to patterns previously discerned.

In most studies referenced above, the data was further structured into a set of 'if-then' rules using Association Rule Mining (ARM). These rules denote connections between two or more variables, like 'if the decision to open a window is driven by temperature, then the same factor drives the decision to close it' (based on (D'Oca & Hong, 2014)). This rule illustrates a repeated pattern by an occupant throughout the data collection period, and such rules represent the occupant's behavioral tendencies.

The utility of Association Rule Mining (ARM) in uncovering occupant behavior patterns is limited. However, other research in building energy efficiency has successfully leveraged ARM to detect faults, identify opportunities for energy conservation, and characterize households. For instance, (Fan, Xiao, Madsen, & Wang, 2015)employed ARM to identify relationships among diverse parameters recorded from a chiller plant, thereby pinpointing abnormal operational states. Similarly, (Yu, Haghighat, Fung, & Zhou, 2012) utilized ARM on sub-metering data collected

over two years from an office building's air conditioning system, enabling the identification of patterns of energy wastage and potential faults by comparing the resulting rule sets from each year.

(Cabrera & Zareipour, 2013) focused on detecting wastage patterns in classroom lighting systems. They initiated their work by identifying waste patterns. They harnessed the power of Association Rule Mining (ARM) to establish relationships between these patterns and various factors, including season, time, day of the week, and occupancy, among others—the resulting rules furnished facility managers with valuable and practical knowledge to decrease the consumption of lighting energy.

3.10.1 ARM Parameters

The procedure involved in creating association rules involves a twofold approach. The initial step necessitates the creation of frequent item sets, while the subsequent step involves deriving rules based on these frequent item sets. In Association Rule Mining (ARM) context, an itemset refers to a collection of features that regularly appear together. Each feature is considered an 'item,' and a collection with 'k' features is known as a k-itemset.

Various algorithms are available for generating frequent item sets, including the Apriori algorithm (introduced by Rakesh Agrawal in 1994) and the FP-growth algorithm (Han, Pei, & Yin, 2000). The fundamental concept behind these frequent itemset generation methods is to utilize strategies to decrease the computational complexity that a brute-force search would entail. Each algorithm employs different techniques to achieve this.

However, when the objective is to generate a relatively small set of rules, both the Apriori and FP-growth algorithms demonstrate similar performance, as indicated by (Zheng, Kohavi, & Mason, 2001).In light of the present study, the Apriori algorithm is the preferred method for creating frequent itemsets.

The analysis of association rules necessitates the utilization of two essential input variables: support and confidence, both determined by particular formulas commonly referred to as Equations (2) and (3), respectively. When examining a rule expressed as $A \rightarrow B$, A is designated as the antecedent, and B is recognized as the consequent.

The "support" of a rule indicates the joint probability of A and B (expressed as $P\ (A \cap B)$) or the frequency at which A and B are found together in the dataset. Conversely, the "confidence" of the rule stands for the conditional probability of B, given the occurrence of A. In simpler terms, it measures the probability of encountering B when A has been observed.

Elevated support implies that A and B are common co-occurrences in the dataset. At the same time, increased confidence indicates a higher likelihood of B's presence when A is observed, suggesting a stronger correlation between A and B. Therefore, it is crucial to thoroughly understand these fundamental concepts while analyzing association rules to ensure the accuracy and validity of our findings.

$$Supp\ =\ P(A\ \cup\ B) \qquad \text{Equation (2)}$$

$$Conf = P(B|A) = P\ (A \cup B)/\ P(A) \qquad \text{Equation (3)}$$

Configuring support and confidence parameters is a crucial aspect that demands careful consideration, particularly in specific applications. The ideal setup of these parameters predominantly depends on the desired rules, their frequency of occurrence, and the strength of the correlation between them.

When both high frequency and strong correlation are desired, setting both parameters to higher values is recommended. However, certain studies may prioritize rules that exhibit a robust relationship, regardless of their frequency of occurrence. In such scenarios, a lower value may be allocated to support while confidence is maintained at a higher value. For instance, (Fan et al., 2015) set support at 0.1 and confidence at 0.9 to achieve their study objectives. Hence, it is crucial to tailor the configuration of these parameters to align with the specific research objectives.

3.10.2 Organization of Rules

The Association Rule Mining (ARM) algorithm is frequently employed to uncover relationships within data variables. Nevertheless, it primarily targets categorical data. Even though there are algorithms designed to handle numerical data, they are usually applied to find associations across numerous numerical datasets. For example, these could be used to investigate the correlation

between the energy usage of cooling systems and primary air-handling units. (Fan et al., 2015). In this study, both categorical and numerical attributes are present. Thus, converting the numerical data into categories is advantageous before applying conventional ARM algorithms.

Various methodologies for data categorization can be employed, such as partitioning data into equal-width ranges or equal-depth ranges (Aggarwal, 2015). However, the choice of categorization technique mainly depends on the specific application.

3.10.3 Rules Generation

A significant challenge with the ARM algorithm is its tendency to yield an overabundance of trivial and redundant rules. As such, it is crucial to select the rules post-generation for their significance meticulously.

All numerical data were transformed into categorical form to facilitate the execution of the ARM algorithm. Post categorization, the original data was replaced with the respective names of the intervals: The categories for Heating Setpoint Temperature were as follows: 16-18℃, 18-20℃, 20-22℃, 22-24℃, 24-26℃, and temperatures above 26℃. Similarly, the Cooling Setpoint Temperature categories were defined as 16-18℃, 18-20℃, 20-22℃, 22>℃, and temperatures above 24℃. Further, the Hour of day categories were defined as nighttime (22:00h to 07:00h) and daytime (08:00h to 21:00h).

It is important to note that the Ecobee thermostat and its operational range played a vital role in defining the ranges for 'Thermostat_Temperature' and 'T_out' (outdoor temperature).

The recommended indoor temperature ranges for thermal comfort, according to ASHRAE Standard 55, is between 68°F (20°C) and 78°F (25.5°C) for the heating and cooling seasons, respectively. However, it should be noted that personal preferences may vary for each household. Additionally, the outdoor temperature range depends on location and time of year. A simple categorization can be done based on the broad comfort range of outdoor temperatures, such as below freezing (≤32°F or 0°C), cold (33°F or 0.5°C to 50°F or 10°C), cool (51°F or 10.5°C to 65°F or 18°C), and mild (66°F or 19°C to 80°F or 26.5°C).

The study used the 'Hold' rule to understand the features and reasons that cause override. By analyzing when the 'Hold' status is triggered, we can gain insights into what conditions or factors—

like outdoor temperature, time of day, or setpoint temperature most often lead to the user adjusting the thermostat setting outside the usual schedule. In addition to categorizing Setpoint Temperatures, Hour of Day, Thermostat Temperature, and Outdoor Temperatures, another significant aspect to examine in this context is the 'Hold' feature in the Ecobee thermostat. As mentioned in this chapter, 'Hold' refers to when a user overrides the existing thermostat schedule to maintain a specific temperature for an extended period.

By categorizing 'Hold' events, and including them in the association rule mining process, we can seek to understand the circumstances that lead to these overrides. For instance, a 'Hold' event may occur during certain hours or when the outdoor temperature falls into specific categories. This step can provide insights into how users interact with their thermostats, which could be used to optimize default settings and user interfaces or provide more personalized comfort recommendations.

Thus, the 'Hold' feature can be seen as a target variable in our rule mining process, helping us analyze and understand the features and reasons that lead to an override in the thermostat settings.

3.11 Comparison and Automation

Analyzing data from different years (2018-2019) can provide insights into how thermostat overrides change over time, reflecting changes in user behavior or climate patterns. By applying the ARM algorithm separately to each year's data, we can identify which rules remain consistent and which vary. For example, we may discover that 'Hold' events occur more frequently during certain times of the day in one year than the other. Alternatively, changes in outdoor temperature categories frequently leading to 'Hold' events suggest that users' temperature preferences or weather conditions have changed. Identifying these patterns and changes can help inform automation strategies. The system can automatically adjust thermostat settings based on the time of day, outdoor temperature, and other factors. For instance, if rule mining reveals that users frequently set a 'Hold' at 22℃ during the daytime in cooler months, the automation system can automatically adjust the thermostat to this setting under these conditions. If rules change, the automation algorithms can be updated to reflect the users' evolving preferences. By combining association rule mining and comparing data from different years, we can develop more effective and responsive automation strategies, ultimately improving the heating and cooling system's energy efficiency and comfort.

Chapter 4:

4. Results and Analysis

4.1 Data preparation

When preparing the data, different houses had different numbers of complete days (days with all 24 hours) in each season (heating and cooling). The 24-hour timeframe was chosen as the standard measure for a complete day to ensure consistency and accuracy in the analysis. These differences could be due to various reasons, such as data collection issues, the house being unoccupied, the HVAC system being off, etc. In the context of this analysis, fewer complete days might mean that any overridden behavior makes up a more significant percentage of the available data, hence increasing the override percentage. Conversely, more complete days provide a more extensive data set, which might dilute the impact of the override behavior on the overall percentage. In this sense, a house with more complete days might exhibit a lower override percentage, even if the frequency of override events is the same.

Lastly, regional climate variations, insulation characteristics, or resident preferences could also affect each house's heating and cooling seasons. These factors, coupled with different numbers of complete days, could substantially influence the distribution of overrides and should be considered during data interpretation. Table 4-1 presents a comparative overview of the two households, elucidating their number of complete days' disparity.

Table 4-1: The number of complete days for each House

No. of Household	Complete days Heating Mode 2018	Complete days Cooling Mode 2018	Complete days Heating Mode 2017	Complete days Cooling Modes 2017
3	95	85	88	100
5	114	103	152	116

4.1.1 Distribution of Override

The next step was to consider the year 2018 and then compare the results with the year 2017. After selecting the HVAC mode, the overall override frequency for each schedule during the dataset was conducted. Figure 4-1 depicts the percentage of override for two Households for cooling and heating HVAC mode for the years 2018 and 2017.The distribution is based on the total number of override among the whole dataset.

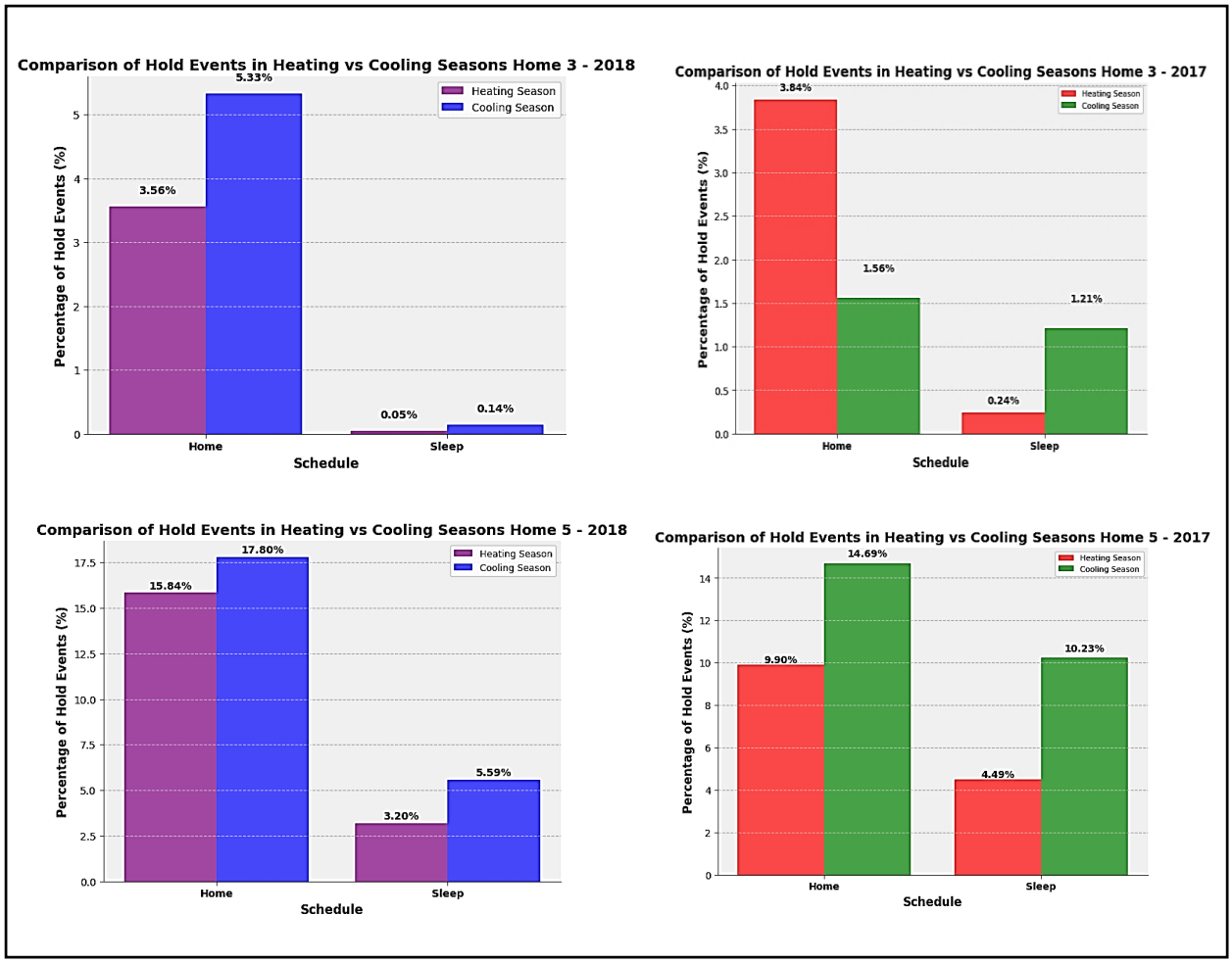

Figure 4-1: Distribution of Households' Override during the Schedules

Home 3:

For 2018, during the heating period, Home 3 had an override frequency of 3.56% during the 'Home' time and a substantially smaller 0.05% during the 'Sleep' time. The cooling period saw a slightly higher override frequency, with 5.33% during 'Home' time and 0.14% during 'Sleep' time.

When comparing these values to the 2017 data, we can observe a minor decrease in override frequency for the heating period in both 'Home' (from 3.84% to 3.56%) and 'Sleep' (from 0.24% to 0.05%) categories. This could suggest that the residents of Home 3 were becoming more accustomed to or satisfied with their pre-set heating schedule in 2018 compared to the previous year.

On the other hand, the cooling period override frequency significantly increased during 'Home' time (from 1.56% to 5.33%) and slightly decreased during 'Sleep' time (from 1.21% to 0.14%). This suggests that while the residents were more comfortable with the cooling schedule at night, they frequently adjusted it during the day, indicating a potential dissatisfaction with the pre-set cooling schedule in 2018 compared to 2017.

Home 5:

In 2018, Home 5 had a relatively higher override frequency than Home 3. During the heating period, the 'Home' time saw an override frequency of 15.84%, and the 'Sleep' time had a significantly less 3.20%. The cooling period override frequencies were even higher, with 17.80% during 'Home' time and 5.59% during 'Sleep' time.

In 2017, during the heating period, the 'Home' time saw an override frequency of 9.90%, and the 'Sleep' time had a significantly less 4.49%. The cooling period override frequencies were even higher, with 14.69% during 'Home' time and 10.23% during 'Sleep' time.

The data from 2017 and 2018 reveals discernible differences in override frequencies, showcasing the inherent flexibility and variability in occupant behavior and patterns. It could be due to various factors, such as changes in their daily routines, shifting comfort preferences, or potential changes in the external climate.

4.1.2 Average Setpoint and Thermostat Temperature Variations During Override and Non-Override Event

The results of the analysis of the average setpoint and thermostat temperatures during various events and schedules for Home 3 during the year 2018 are illustrated in Table 4-2. The data were classified and arranged based on outdoor temperature increments of three degrees for both the heating and cooling seasons. This approach lets us gain insights into building occupants' temperature control patterns and preferences.

Home 3: During the home schedule, users manually adjusted the setpoint temperature for the heating season, opting for a slightly cooler level in 2018 but higher level in 2017. During the year 2018 for the Home schedule in heating season the difference was (-1.11°C) and for the sleep schedule it was (-0.01°C). For the cooling season, the difference for home schedule was (0.18°C) and for the sleep schedule was (0.53°C) When examining the temperature measured by the thermostat during the home schedule, it was found to be slightly higher when a "Hold" was activated in both years and for both seasons. However, the distinction was more prominent in 2018, with a (-0.48°C) difference for heating and a (0.42°C) difference for cooling, compared to 2017's (0.11°C) difference for heating and (-0.49°C) difference for cooling. Surprisingly, the outside temperature did not significantly impact the thermostat's performance in maintaining the setpoint temperature during the home schedule in both years and seasons. However, an exception was observed during the heating season in 2017, where the outside temperature exhibited a significant decrease of (-1.11°C) when a "Hold" was activated, whereas, in 2018, the difference was only (0.74°C). In the sleep schedule, users consistently set a warmer setpoint temperature during the heating season and a cooler temperature during the cooling season in both years. Nonetheless, the disparity between the "Hold" and "No_Hold" modes was significantly higher in 2017 for both heating (2.22°C) and cooling (1.17°C) compared to 2018's heating difference of (-2.81°C) and cooling difference of (-1.42°C). Like the home schedule, the temperature measured by the thermostat during the sleep schedule was higher when a "Hold" was activated in both years and seasons. However, the discrepancy was notably more substantial in 2017, with a (1.32°C) difference for heating and a (0.12°C) difference for cooling, while in 2018, the differences were (0.14°C) for heating and (0.44°C) for cooling.

Home 5: During the home schedule of the heating season, users manually adjusted the setpoint temperature to a warmer level in both 2017 and 2018. However, the discrepancy between the "Hold" and "No_Hold" modes was less pronounced in 2017, with a difference of (0.89°C), compared to 2018's difference of (0.97°C). Furthermore, the temperature measured by the thermostat was also higher when a "Hold" was activated in both years, with 2017 exhibiting a more significant difference of (0.65°C) compared to 2018's difference of (0.48°C). Interestingly, the outside temperature was significantly lower when a "Hold" was activated in both years, but the difference was more notable in 2018 (-7.97°C) than in 2017 (-1.02°C). In the sleep schedule of the heating season, users consistently set a warmer setpoint temperature in both 2017 and 2018, and the difference between the "Hold" and "No_Hold" modes was more significant in 2017 (2.45°C) than in 2018 (2.22°C). Similarly, the temperature measured by the thermostat was higher when a "Hold" was activated in both years, with 2017 displaying a more significant difference of (1.43°C) compared to 2018's difference of (1.32°C). Moreover, the outside temperature was significantly lower when a "Hold" was activated in both years, but the difference was more prominent in 2018 (-3.91°C) than in 2017 (-2.96°C). During the cooling season of both years, cooling setpoint temperature was higher during the “No_Hold" event. In 2017, setpoint temperature difference was (-2.21°C) for home and (-4.60°C) for sleep schedule. The result for 2018 was different with (-1.55°C) for home and (-3.67°C) for sleep schedule. The thermostat temperature had the same pattern. In 2018, with the difference for (-0.28°C) for home and (-0.36°C) for sleep schedule. Lastly for 2017, the result was (-0.72°C) for home and (-1.10°C) for the sleep schedule.

Home 3: 2018

Table 4-2: Average Temperature difference - Home 3-2018

		Hold			No Hold			Difference		
Season	**Schedule**	Setpoint Temperature (°C)	Thermostat Temperature (°C)	Outdoor Temperature (°C)	Setpoint Temperature (°C)	Thermostat Temperature(°C	Outdoor Temperature(°C	Setpoint Temperature (°C)	Thermostat Temperature(°C	Outdoor Temperature(°C
Heat	**Home**	21.07	20.64	-2.30	22.19	21.12	1.53	-1.11	-0.48	-3.82
	Sleep	20.01	20.10	-4.42	20.02	19.96	-1.62	-0.01	0.14	-2.81
Cool	**Home**	22.48	22.37	24.01	22.30	21.95	22.32	0.18	0.42	1.69
	Sleep	21.12	21.13	16.94	20.59	20.69	18.37	0.53	0.44	-1.42

Figure 4-2 and 4-3 represent an overview of setpoint temperature distribution with outside temperature bins for heating and cooling seasons, override, and non-override events. In boxplots, the numbers inside the box represent the data points in each outdoor temperature range. The minimum and maximum lines in the box represent the 5th and 95th percentiles, respectively. The middle horizontal line in the box denotes the median value. The box's lower and upper horizontal lines denote the 25th and 75th percentiles. The rest of figures and tables were presented in section 8, appendix A.

Heating and Cooling Mode – 2018

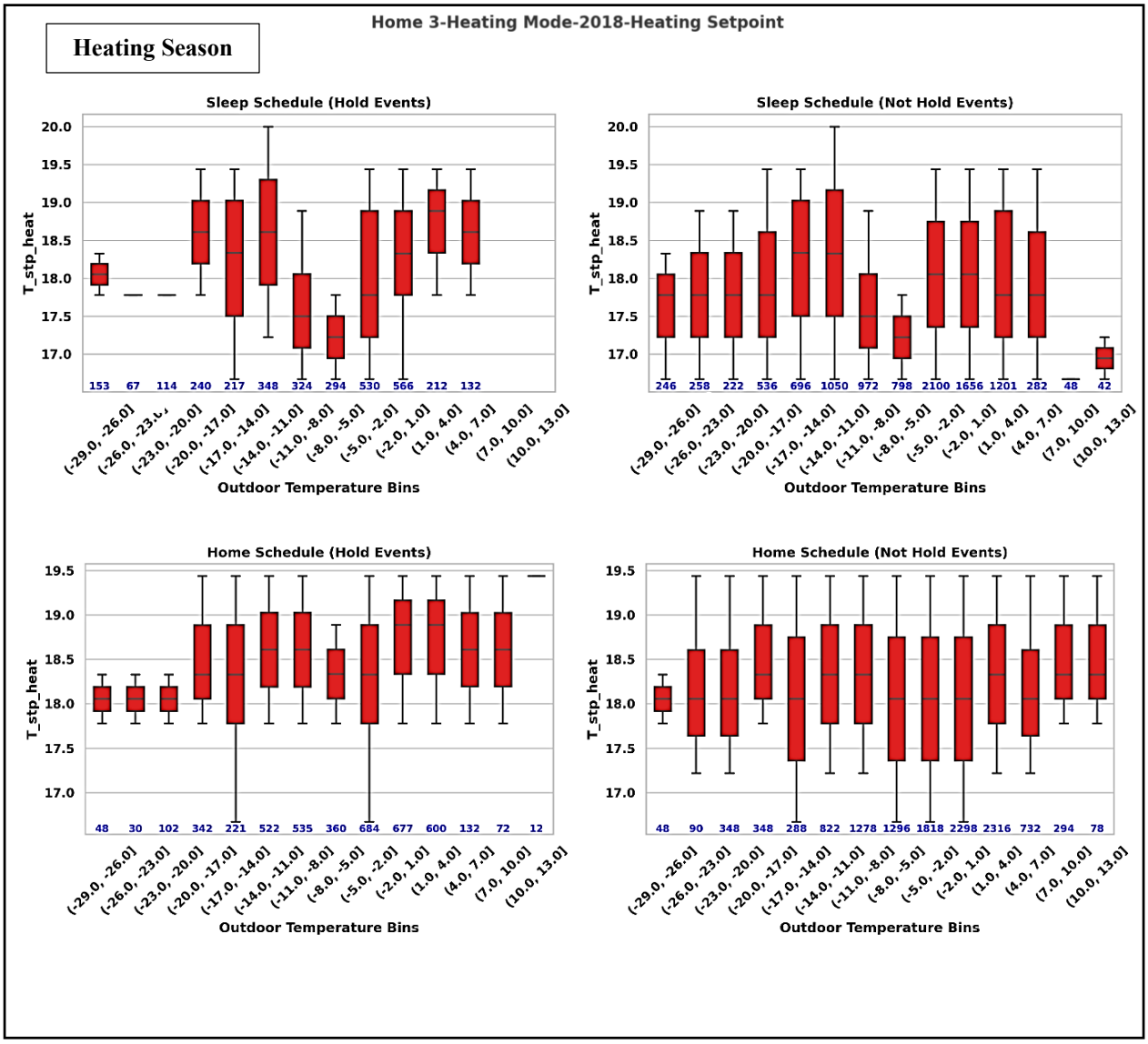

Figure 4-2: Setpoint temperature distribution -Hold and Non-Hold-Home 3-2018-Heating Season

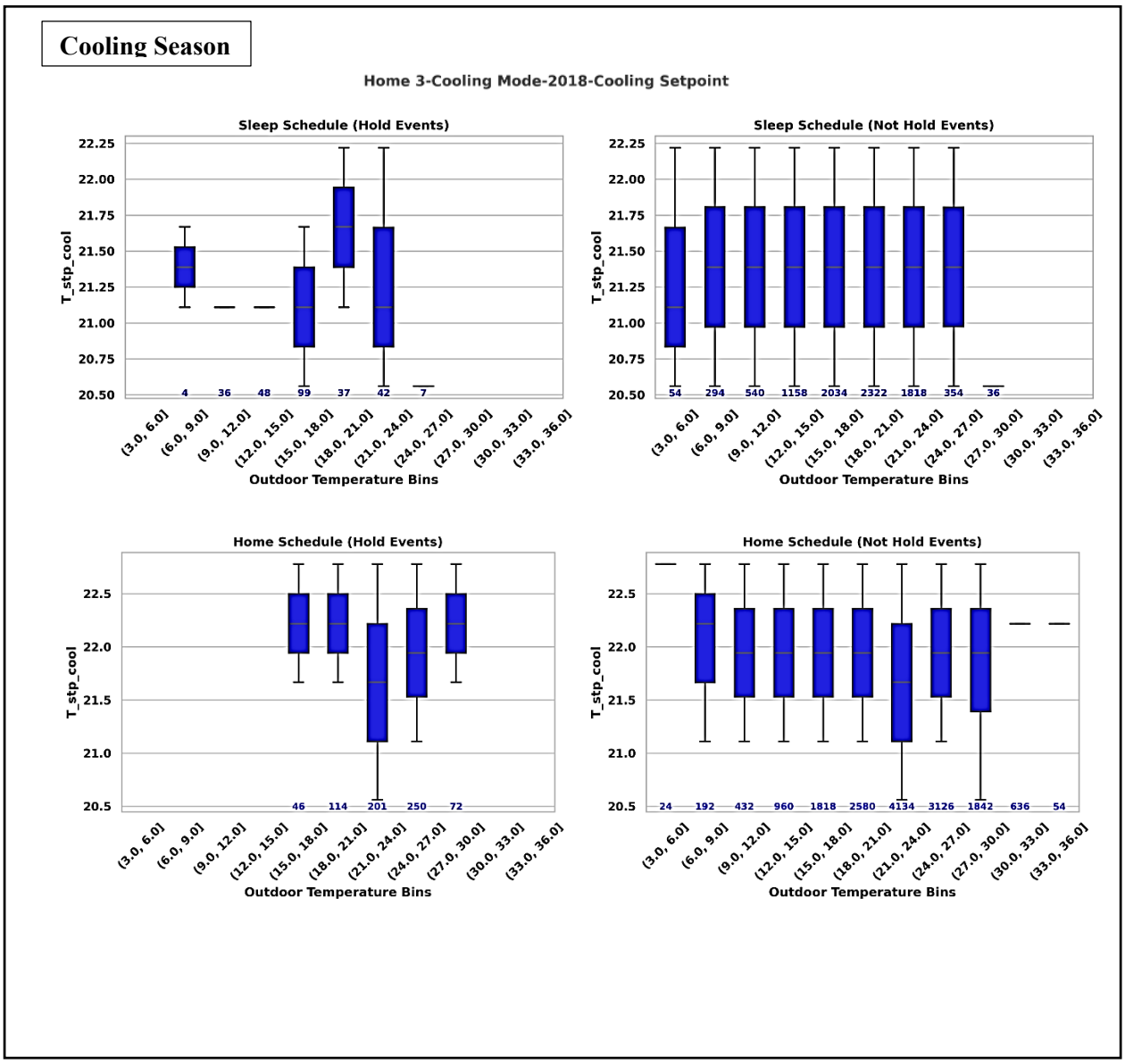

Figure 4-3: Setpoint temperature distribution -Hold and Non-Hold-Home 3-2018-Cooling Season

4.2 Average Hourly Occupant Activity Profile

To fully grasp occupancy and setpoint temperature trends, analyzing monthly changes, differences between weekdays and weekends, and daily fluctuations during weekdays, considering the specific time of day is essential. This method helps identify common patterns and better understand user preferences and habits. Following the method applied by (Hosseinihaghighi et al., 2022), occupant movement detection data were recorded every 5 minutes using ecobee thermostats. This data was

then converted to hourly values for each property, which led to occupancy values that ranged from 0 to 1. A median value for each hour was calculated and used as a threshold to transform the data. Consequently, all average occupancy values were assigned as either 0 or 1. Low movement detection between 22:00 and 07:00 presented a challenge that was handled by assigning a '1' value for the entire night whenever any movement was detected. Figure 4-4 and 4-5 depict the average occupancy and setpoint temperatures for heating and cooling season for Home 3 during the heating and cooling season of 2018.

Home 3 during 2017 and 2018

The average heating setpoint temperature in the heating season in 2018 was marginally reduced to 19.97°C, compared to 20.03°C in 2017. The aggregate average occupancy, as denoted by the total sensor reading, was somewhat elevated in 2017 (0.12) compared to 2018 (0.10), possibly attributable to resident schedules or dwelling usage alterations.

There was a subtle shift in occupancy trends across these two years, with Thursday recording the highest average occupancy in 2018, while Wednesday recorded the same in 2017 during weekdays. The hour witnessing the highest occupancy was 16:00 in 2018, recording a total sensor average of 0.15, a shift from 8:00 in 2017, with a total sensor average of 0.14. The peak heating setpoint was documented at 13:00 in 2018 and 19:00 in 2017. However, the month registering the highest average occupancy and heating setpoint varied, with January for occupancy and December for heating setpoint in 2017 and December for both metrics in 2018.During the cooling season, there was a higher average cooling setpoint temperature in 2017 (22.01°C) compared to 2018 (20.61°C), indicating a warmer internal environment in 2017. The total average occupancy was also higher in 2017 (0.14) compared to 2018 (0.11), accompanied by a shift in occupancy patterns, with Thursday witnessing the highest occupancy in 2018, while Wednesday did in 2017. Peak occupancy was recorded at 8:00 in 2018 and 19:00 in 2017, with distinct total sensor averages (0.14 in 2018 and 0.16 in 2017). The highest cooling setpoint remained consistent at 19:00 in both years. Interestingly, the house was most occupied in June and the least occupied in July 2018, while in 2017, July witnessed the highest occupancy, and August recorded the lowest. Average occupancy and setpoint temperature were consistently higher during weekends than weekdays in heating and cooling seasons.

Home 3 -2018-Heat

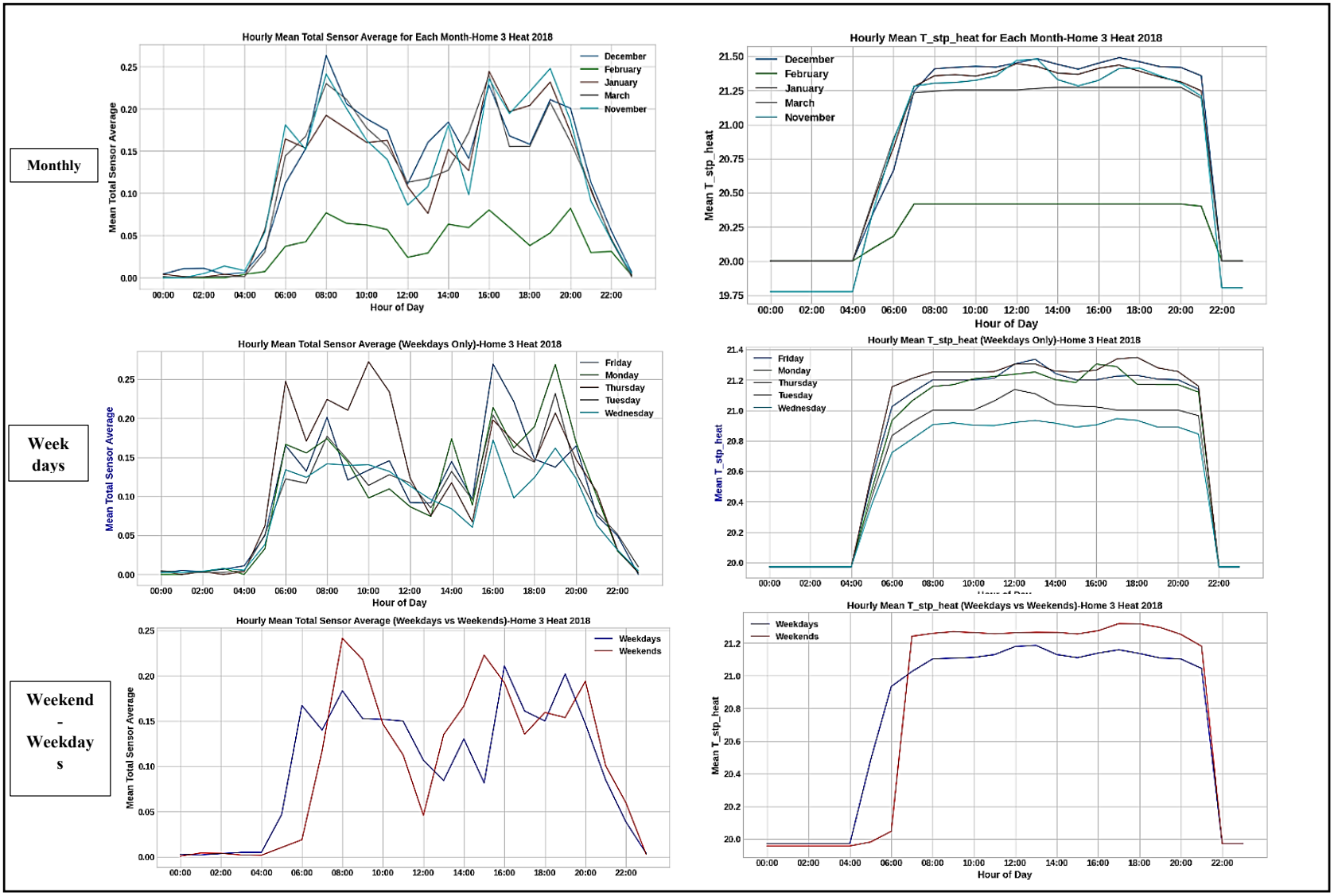

Figure 4-4: Average Occupancy and Heating Setpoint Temperature-Home 3-2018

Home 3 -2018-Cool

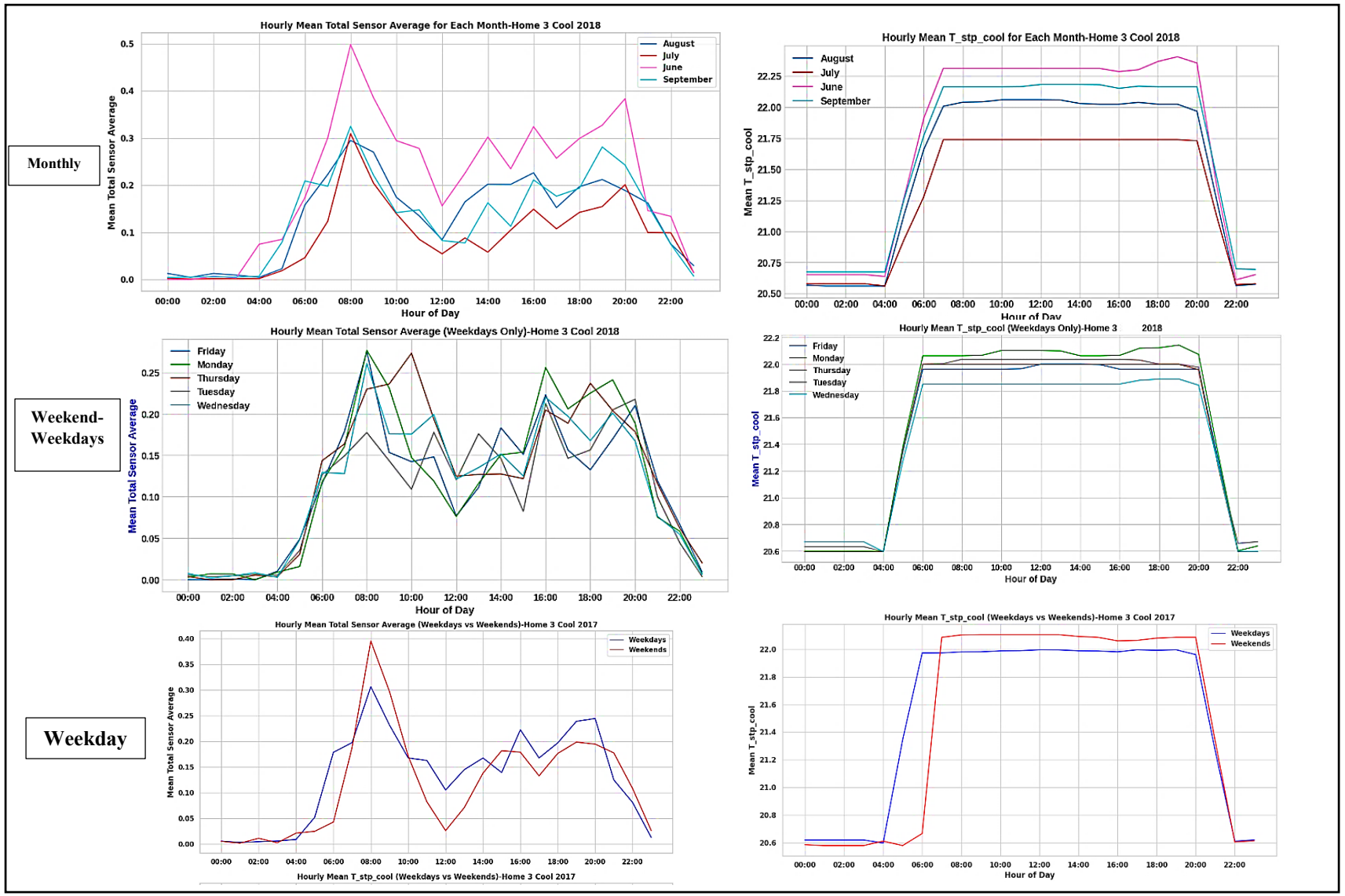

Figure 4-5: Average Occupancy and Cooling Setpoint Temperature-Home 3-2018

Home 5 during 2017 and 2018

During the heating season, the average heating setpoint temperature in 2018 was marginally lower at 19.43°C compared to 19.89°C in 2017, possibly reflecting energy conservation measures, changes in occupancy, or different weather conditions. The overall average occupancy sensor reading was slightly higher in 2017 at 0.182 compared to 0.148 in 2018, a change that could be attributed to variations in the residents' schedules or alterations in home usage patterns.

Weekday occupancy patterns shifted between the two years, with Monday (2018) and Tuesday (2017) registering the highest mean total sensor averages. Peak occupancy and heating setpoint hours were also varied, with 16:00 and 10:00 being the peak hours for occupancy and heat set in 2018, while 18:00 and 19:00 were the peak hours in 2017. January recorded the lowest mean occupancy and heat setting in both years. The months with the highest mean occupancy and heating setpoint were varied, with November in 2017 and March in 2018.

Regarding the cooling season, the average cooling setpoint temperature (T_stp_cool) in 2018 was slightly higher at 24.24°C compared to 23.98°C in 2017, suggesting a cooler indoor environment 2018. The total occupancy sensor reading was lower in 2018 at 0.166 compared to 0.371 in 2017, and the occupancy pattern changed between the two years. Wednesday was the day with the highest occupancy in 2018, and Monday in 2017 during weekdays.

The peak occupancy hours were 17:00 in 2018 and 22:00 in 2017, while the highest cooling setting hour was consistent at 10:00 in both years. The months with the highest mean occupancy and cooling setting were May for both years, while it was least occupied in September 2018 and August 2017 and had the lowest cooling setting in August. Like the heating season, average occupancy and setpoint temperature were higher during weekends than on weekdays during the cooling season.

Both years saw the highest mean occupancy and cooling setpoint in May, while the lowest occupancy was documented in September 2018 and August 2017 and the lowest cooling setpoint in August. Like the heating season, the cooling season also exhibited higher average occupancy and setpoint temperatures during weekends than on weekdays.

4.3 Temperature Statistics During the Override and Non-Override Event

Using hourly data, this part identifies the trends, peaks, and troughs in indoor and outdoor temperatures, control temperature, and setpoint temperatures during heating and cooling seasons during override and non-override events. Occupancy was also examined thoroughly to understand temperature dynamics within both events.

When we delve into the Temperature Preference and Mode Selection, inhabitants of both homes lean towards warmer settings when manually controlling the thermostat during heating seasons. To illustrate, in 2017, Home 3 residents set an average temperature of 22.4°C in 'Hold' mode, while 'Non-Hold' mode saw a lower average of 21.27°C. Such inclinations towards warmer conditions when the occupants actively influence the settings may result from an array of factors such as individual comfort levels, age, health conditions, and potentially even cultural variables.

Investigating Variability in Temperatures, we find that the temperature range (min to max) is broader during periods of manual control ('Hold' mode) compared to automated settings. This observation suggests that manual settings cause the system to exert more effort to maintain the specified temperature due to more significant fluctuations, possibly influenced by external weather changes, home insulation efficiency, or differing occupancy levels.

When examining Seasonal Differences, we observe variations in temperature preferences and mode usage between cooling and heating seasons, potentially signifying adaptations in user behavior according to the season. A notable instance in Home 5 during the cooling season of 2018 revealed a higher set-point temperature in the 'Non-Hold' mode than the 'Hold' mode, indicating higher energy consumption when the house is less occupied. Upon exploring Inter-Home Differences, we note intriguing patterns between the two homes. Home 5 consistently exhibits higher occupancy levels than Home 3, suggesting differing living patterns or schedules between the two households. Such distinctions are vital for user behavior modeling and improving the performance of smart thermostats.

Table 4-3 and 4-4 contains each home's average heating and cooling set points in both hold and non-hold modes for 2017 and 2018.

Table 4-3: Home 3- Temperature Statistics During the Override and Non-Override Event

Year	Season	Mode	Average Setpoint Temperature (°C)	Min Setpoint Temperature (°C)	Max Setpoint Temperature (°C)
2017	Heating	Hold	22.4	21.11	22.78
2017	Heating	Non-Hold	21.27	20	22.45
2018	Heating	Hold	22.04	21.74	23.41
2018	Heating	Non-Hold	20.71	19.44	21.98
2017	Cooling	Hold	22.46	20.56	23.89
2017	Cooling	Non-Hold	22.19	21.11	23.33
2018	Cooling	Hold	22.05	20.15	23.95
2018	Cooling	Non-Hold	21.5	20.42	22.58

Table 4-4: Home 5- Temperature Statistics During the Override and Non-Override Event

Year	Season	Mode	Average Setpoint Temperature (°C)	Min Setpoint Temperature (°C)	Max Setpoint Temperature (°C)
2017	Heating	Hold	21.22	19.68	22.6
2017	Heating	Non-Hold	19.73	17.78	23.6
2018	Heating	Hold	21.14	19.60	22.68
2018	Heating	Non-Hold	19.5	17.55	21.75
2017	Cooling	Hold	23.9	20.93	26.11
2017	Cooling	Non-Hold	22.13	19.58	24.17
2018	Cooling	Hold	21.64	18.67	24.61
2018	Cooling	Non-Hold	23.88	20.91	26.87

4.4 Analyzing the Hold Cycle

Table 4-4 describe the information related to each Households' Hold Cycles. As mentioned in section (3.8.4), the "Average Temp Difference (°C)" signifies the mean difference between the control and setpoint temperatures, and the "Inside-Outside Temp Difference (°C)" denotes the difference between indoor and outdoor temperatures. The Cycle is based on hour during a whole day.

1. **Home 3 (2018 - Heat):** In Hold Cycle 1, a two-hour override was observed at 72.7%, suggesting the possibility of relatively short bursts of additional heating being preferred. The thermostat's well-regulated performance in maintaining temperature consistency is indicated by the range of the average temperature difference (-0.16 to 1.33°C).
2. **Home 3 (2018 - Cool):** Hold Cycle 1 recorded a one-hour hold 43.8% of the time and a two-hour hold 31.3%. During hold cycle 1, the temperature differences were from -0.19°C to 1.13°C. This means users were adjusting the cooling temperature by roughly 1.13°C from the control settings.
3. **Home 3 (2017):** During Hold Cycle 1 for both heating and cooling modes, a one and two-hour override was recorded 33.3% of the time, indicating a possible preference for short-term adjustments. This suggests the need for only occasional manual overrides.
4. **Home 5 (2018 - Heat):** A preference for two-hour overrides was observed in Hold Cycle 1, with a frequency of 73.3%. Frequent minor adjustments to heating might be indicated by a substantial number of one-hour holds, recorded at 8.9%.
5. **Home 5 (2018 - Cool):** A two-hour override was used 60% of the time during Hold Cycle 1, with one-hour overrides also found to be expected, making up 15%. The data suggests that adjustments to cooling might be made frequently to accommodate varying comfort levels or changes in the outside temperature.
6. **Home 5 (2017 - Heat):** During both the first and second Hold Cycles, a strong preference for two-hour overrides was observed, with 79% frequency. This suggests a preference for adjustments of moderate duration.

In conclusion, usage patterns of hold cycles across different homes and years vary significantly, as indicated by these percentages. Some users primarily used short holds, which might indicate a preference for occasional temperature adjustment. Consistent reliance on long holds was seen in

others, suggesting a preference for substantial departures from the thermostat's regular programming. These patterns, when understood, can provide valuable insights for improving thermostat technology, underlining the need to offer a range of hold durations to accommodate various user preferences.

Table 4-5: Hold Cycles' details for each Households

Home No.	Year	Condition	No. of Hold Cycle	Number of Days	Override Durations (Hours) and Percentage	Average Temp Difference (Between the control and setpoint temperatures) (°C)	Inside-Outside Temp Difference (°C)
Home 3	2018	Heat	0	54	N/A	N/A	N/A
Home 3	2018	Heat	1	33	1 (6.1%), 2 (72.7%), 4(3%), 5 (3%), 9 (3%)	-0.16 to 1.33	15.44 to 41.38
Home 3	2018	Heat	2	8	2 (56.3%), 3 (6.3%), 4 (18.8%), 5 (18.8%)	-0.09 to 0.21	19.63 to 39.24
Home 3	2018	Cool	0	65	N/A	N/A	N/A
Home 3	2018	Cool	1	16	1 (43.8%), 2 (31.3%), 4 (18.8%), 9 (6.3%)	-0.19 to 1.13	4.75 to 5.99
Home 3	2018	Cool	2	4	3 (25%), 4 (75%)	-0.14 to 0.23	3.72 to 9.61
Home 3	2017	Heat	0	80	N/A	N/A	N/A
Home 3	2017	Heat	1	6	1 (33.3%), 2 (33.3%), 3 (16.7%), 5 (16.7%)	-0.13 to 0.67	18.98 to 39.66
Home 3	2017	Heat	2	2	2 (50%), 5 (50%)	-0.14 to -0.02	19.63 to 44.53
Home 3	2017	Cool	0	87	N/A	N/A	N/A
Home 3	2017	Cool	1	10	1 (40%), 2 (40%), 3 (10%), 4 (10%)	-1.62 to 0.24	2.57 to 6.13
Home 3	2017	Cool	2	3	2 (66.7%), 4 (33.3%)	-0.25 to 0.12	0.05 to 3.85
Home 5	2018	Heat	0	40	N/A	N/A	N/A

Home 5	2018	Heat	1	45	1 (8.9%), 2 (73.3%), 3 (8.9%), 4 (8.9%)	-1.62 to 0.83	14.08 to 41.49
Home 5	2018	Heat	2	20	1 (20%), 2 (65%), 3 (10%), 4 (5%)	-0.14 to 1.64	18.92 to 40.26
Home 5	2018	Heat	3	8	1 (25%), 2 (50%), 3 (12.5%), 4 (12.5%)	-0.23 to 1.29	25.94 to 41.26
Home 5	2018	Heat	4	1	2 (100%)	0.49 to 0.69	38.11 to 40.74
Home 5	2018	Cool	0	38	N/A	N/A	N/A
Home 5	2018	Cool	1	40	2 (60%), 4 (22.5%), 1 (15%), 5 (2.5%)	-1.03 to 1.24	-6.34 to 20.56
Home 5	2018	Cool	2	21	2 (81%), 4 (9.5%), 1 (4.8%), 3 (4.8%)	-3.66 to 1.81	5.99 to 9.56
Home 5	2018	Cool	3	4	2 (75%), 0 (25%)	-3.66 to 0.94	5.23 to 6.99
Home 5	2017	Heat	0	54	N/A	N/A	N/A
Home 5	2017	Heat	1	48	2 (79%), 4 (12.5%), 1 (6.25%), 5 (2%)	-0.56 to 1.65	9.44 to 45.55
Home 5	2017	Heat	2	29	2 (79%), 4 (6.9%), 1 (6.9%), 3 (3.4%), 0 (3.4%)	-0.59 to 1.71	15.58 to 49.04
Home 5	2017	Heat	3	5	2 (80%), 0 (20%)	-1.52 to 1.17	21.63 to 47.18
Home 5	2017	Heat	4	1	2 (100%)	0.03 to 0.71	28.67 to 32.55
Home 5	2017	Cool	0	55	N/A	N/A	N/A
Home 5	2017	Cool	1	40	2 (85%), 4 (10%), 5 (5%)	-1.71 to 2.46	4.02 to 9.01
Home 5	2017	Cool	2	13	2 (85%), 4 (7.7%), 1 (7.7%)	-3.37 to 1.71	2.28 to 10.28
Home 5	2017	Cool	3	7	2 (71%), 4 (29%)	-3.56 to 1.04	3.04 to 4.63
Home 5	2017	Cool	4	1	2 (100%)	-2.85 to -0.69	2.73 to 4.16

4.5 Feature Importance

The importance of various parameters was analyzed using embedded feature analysis methods. The findings in Figure 4-7 to 4-10 indicates that different prediction algorithms ranked the same variable (Hold Cycle) differently for Home 3. It could be assisted in determining the most critical and minor essential features among the selected optimal ones**.** SHAP (Shapley Additive explanations) values are a method derived from game theory used to explain individual predictions of machine learning models by assigning each feature an important value for a specific prediction. In the context of the plots in this section, the feature importance visualizations are based on SHAP values, allowing for an interpretable breakdown of which features most influence each prediction.

Every household has a unique number of Hold Cycles, each with significant features. By observing these features, we can comprehend the crucial role that outdoor temperature plays in both heating and cooling seasons. When analyzing the features related to the override of an Ecobee thermostat, several factors come into play. These features provide valuable insights into the dynamics of temperature adjustments and the overall behavior of households. Among these features, outdoor temperature emerges as a significant factor. It plays a crucial role in heating and cooling seasons, as households adjust their thermostat settings in response to changing outdoor conditions. The relationship between outdoor temperature and thermostat adjustments allows households to maintain desired comfort levels while optimizing energy usage. Another essential feature is the thermostat temperature, which reflects the current temperature reading displayed on the Ecobee thermostat. By monitoring this temperature, households can ensure that the indoor temperature aligns with their desired comfort level. Deviations from the desired temperature may prompt occupants to override the thermostat settings and adjust accordingly. Setpoint temperatures, which represent the target temperatures set by occupants, also play a significant role. These temperatures serve as a reference point for the thermostat's operation. If occupants perceive a deviation from their desired setpoint temperature, they may override the thermostat to align the indoor temperature with their preferences. Average occupancy and hours of the day are additional features that influence thermostat overrides. Occupancy patterns can vary throughout the day, and different household members may have varying temperature preferences. As a result, occupants may override the thermostat settings to accommodate their comfort needs or adapt to changing occupancy levels. Furthermore, the schedule emerges as the less influential feature in thermostat overrides. While the schedule may have some influence on temperature adjustments, it is of lesser

significance compared to other factors. Nonetheless, considering the schedule in conjunction with other features can provide a more comprehensive understanding of temperature management and household behavior. The occupancy state emerges as the least influential feature in thermostat overrides. Although occupancy patterns can impact temperature adjustments, it has a lesser impact than other factors. Occupancy state reflects the presence or absence of occupants in the household. It has been observed that each Hold Cycle is encapsulated by its distinct preferences and patterns. It has been implied that, whereas outdoor temperature may be the driving force in one cycle in another, setpoint temperatures or average occupancy may be given precedence. The granularity of this discovery has been underscored, emphasizing the necessity for an approach to understanding thermostat behavior that is both tailored and discerning, one where the individuality of each Hold Cycle is respected and prioritized.

Home 3 – Heat – 2018- Decision Tree

Decision Tree

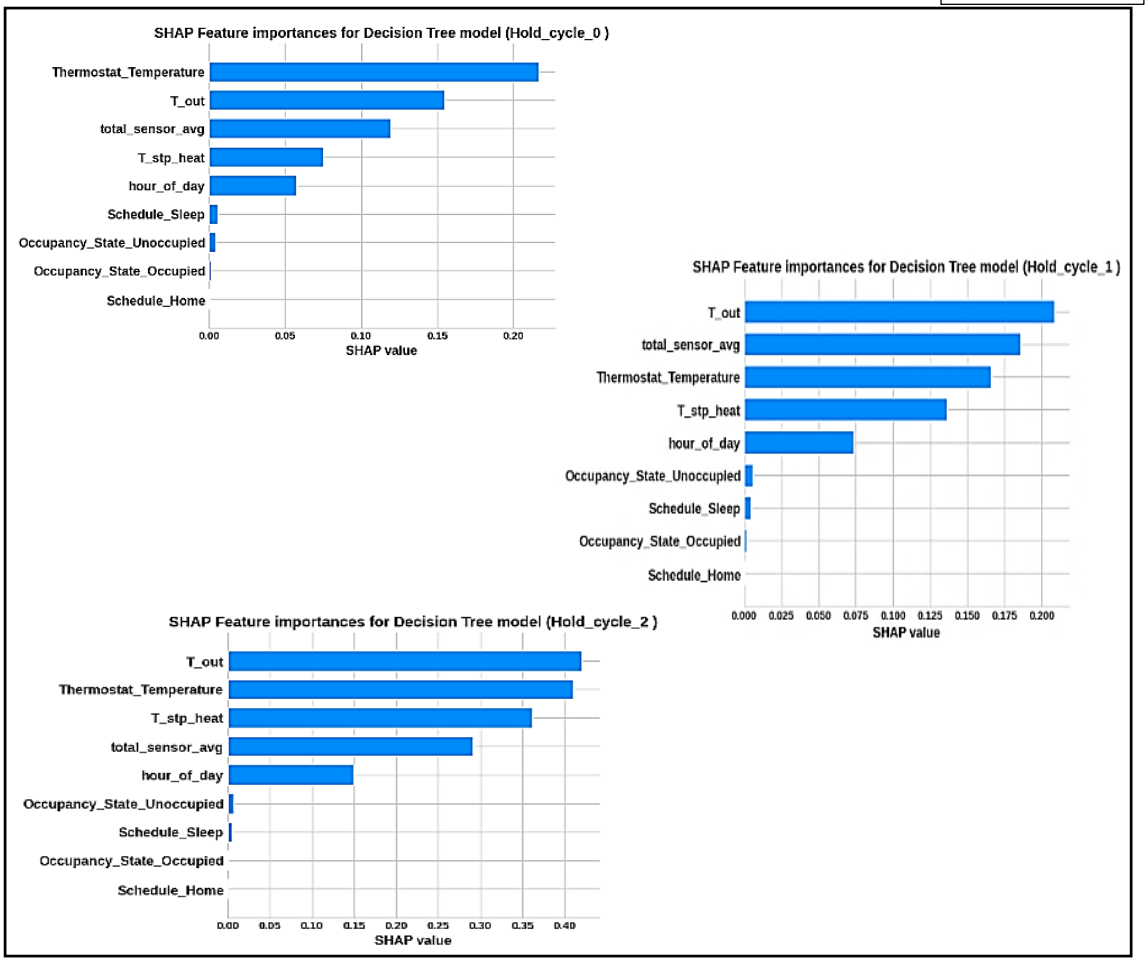

Figure 4-6: Feature Importance-Home 3-Heating Season-2018- Decision Tree

Home 3 – Heat – 2018- Random Forest

Random Forest

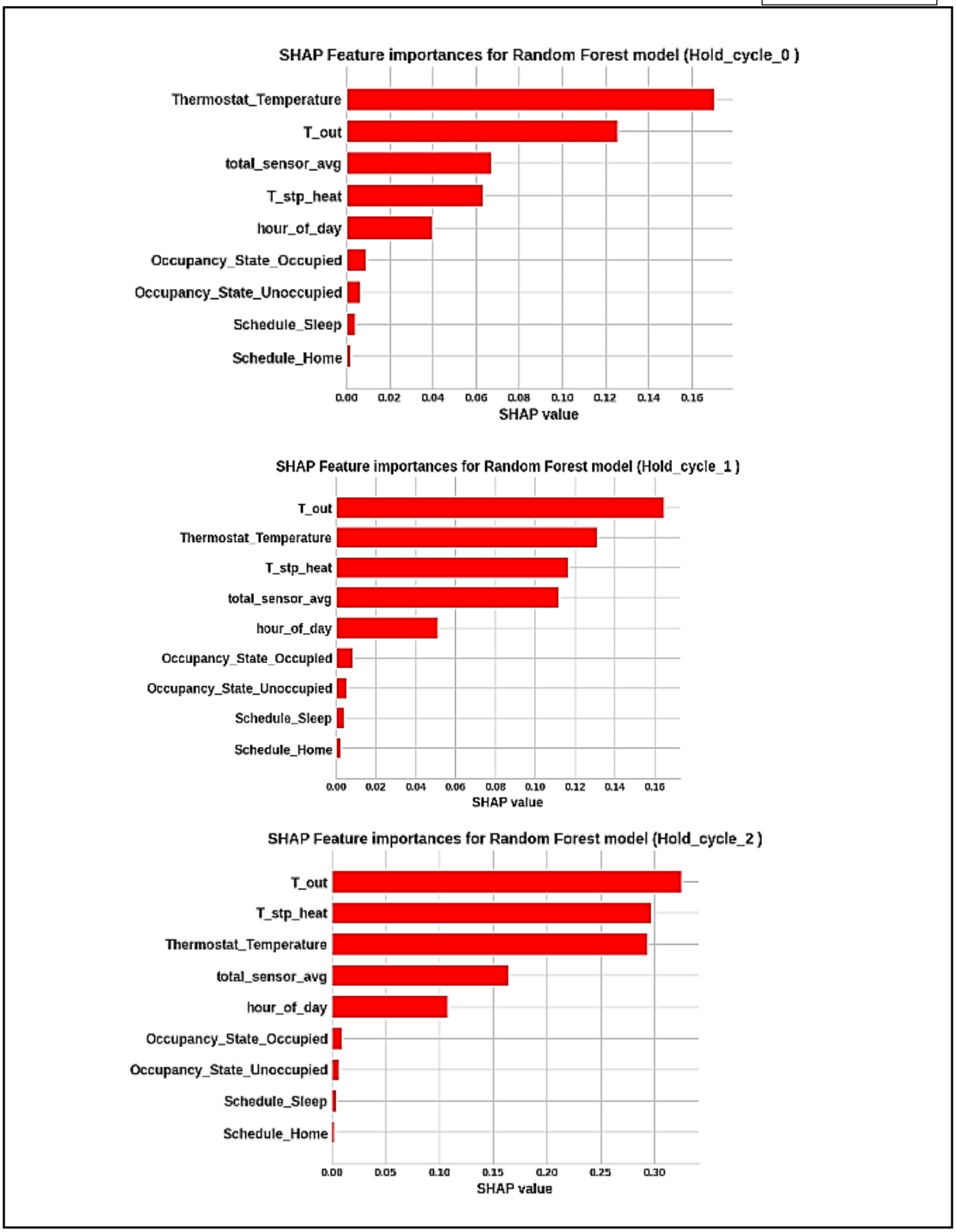

Figure 4-7: Feature Importance-Home 3-Heating Season-2018- Random Forest

Home 3 – Cool – 2018- Decision Tree

Decision Tree

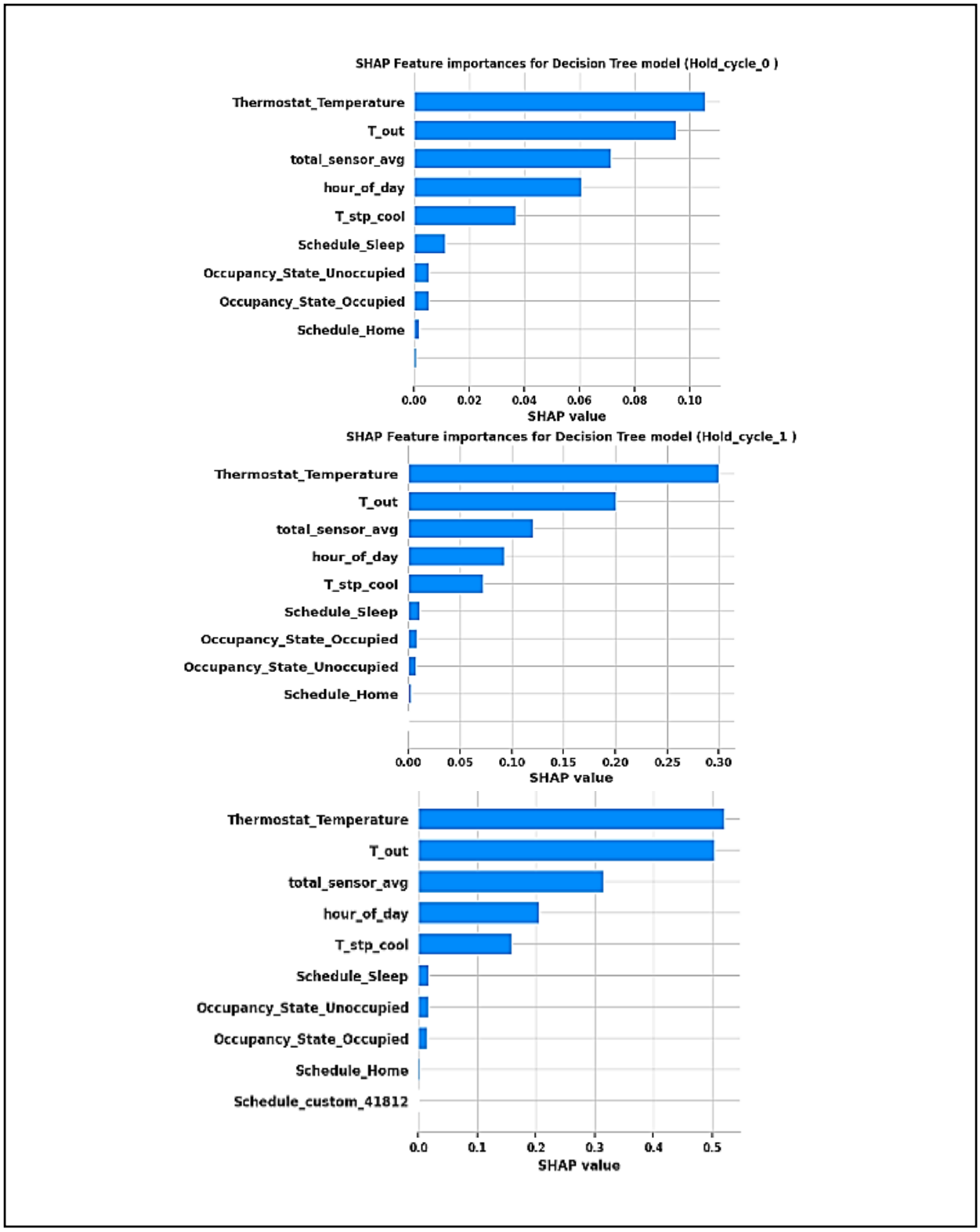

Figure 4-8: Feature Importance-Home 3-Cooling Season-2018- Decision Tree

Home 3 – Cool – 2018- Random Forest

Random Forest

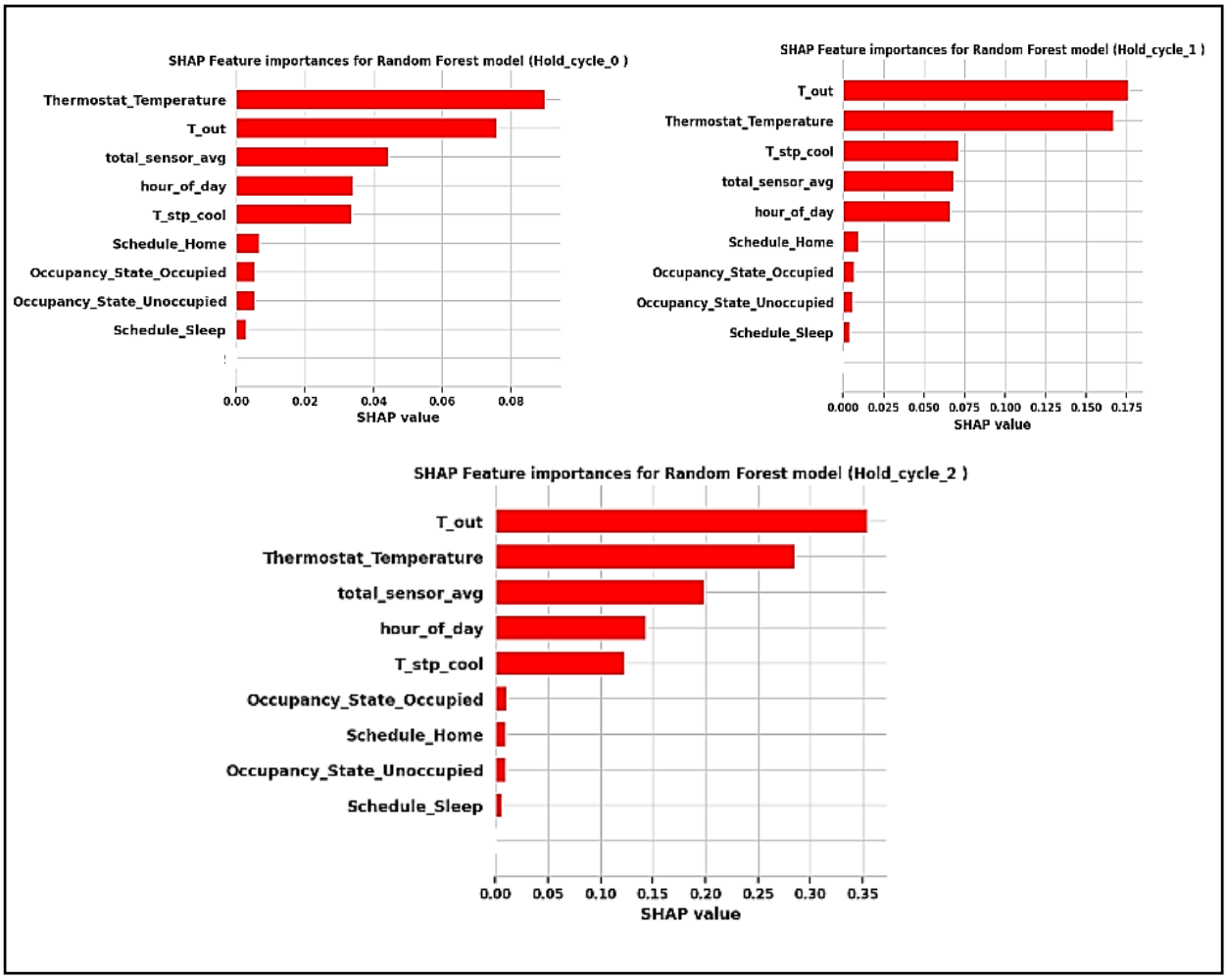

Figure 4-9: Feature Importance-Home 3-Cooling Season-2018- Random Forest

4.5.1 Performance evaluation

The performance evaluation of the models used to predict thermostat overrides was carried out by examining the R^2 score, which measures the goodness-of-fit of regression models. The R^2 score reveals how well the model captures the variation in the target variable, with a perfect fit indicated by a score of 1. The decision tree model yielded a substantial R^2 score of 0.96, demonstrating a strong correlation between the predicted and actual thermostat overrides. This result suggests that the model effectively utilized the relationships among the features, leading to accurate predictions. Moreover, an R^2 score of 0.96 means that the model could explain approximately 96% of the

variability in the target variable, which is the occurrence of thermostat overrides. This level of accuracy shows the model's capability to discern patterns and dependencies within the dataset, reinforcing its usefulness in this application. However, the random forest model performed even better, achieving an average R^2 score of 0.98. This result indicates that the random forest model could more effectively capture and use the relationships among the features, resulting in even better predictions. The high R^2 score of 0.98 indicates that the random forest model explained around 98% of the variability in the occurrence of thermostat overrides. This superior accuracy level gives us more confidence in the results produced by the random forest model. Regarding feature importance, the analysis revealed that outdoor temperature, thermostat temperature, setpoint temperatures, day hours, and schedule were the most influential features in determining thermostat overrides. On the other hand, the occupancy state had the most negligible impact on these overrides.

In summary, while both the decision tree and random forest models demonstrated high levels of accuracy, the random forest model outperformed with an average R^2 score of 0.98. This indicates its superior ability to predict thermostat overrides based on the provided features, thereby allowing us to confidently assess the importance of these features and gain valuable insights into the factors that influence household thermostat adjustments.

4.6 Association Rule Mining

In this segment, we delved into the patterns of thermostat override in two distinct households: Home 3 and Home 5. After considering 2017 and 2018 as separate entities, we took a comparative approach that transcends individual years. Based on this comparative analysis, we aim to propose more precise and universally applicable rules for future thermostat use. The conditions upon which these suggestions are based include indoor and outdoor temperatures, the set-point temperature for heating or cooling, the occupancy of the house, and the time of day. It is vital to remember that these recommendations are developed from specific patterns observed within these particular households and might not necessarily apply universally. We have employed the Apriori algorithm to discern these patterns. We aim to derive more accurate and general rules considering two distinct years.

4.6.1 Support and Confidence

During the rule generation process, support and confidence thresholds were carefully selected to identify frequent rules in the dataset and have a strong correlation. Only significant rules were extracted by setting appropriate minimum support and confidence levels. The support of a rule represents the frequency with which that particular rule occurs in the dataset as a percentage of the total records. Thus, a support level of 18% indicates that a specific rule was observed in 18% of the overall data. Additionally, confidence indicates how frequently the consequent occurs, given that the antecedent is true. The rules were further evaluated for relevance after mining the dataset using Apriori with suitable support and confidence thresholds. The percentage frequencies provided were based on those meaningful rules' support or occurrence levels. As an illustration, the "IF Setpoint is 22-24°C AND Outdoor Temp <= 0°C AND Time is Daytime AND Year is 2018 THEN Hold" rule for Home 3 heating in 2018 had support of 18%. This means the pattern was observed in 18% of the total Home 3 heating season data in 2018.

4.6.2 Automation Rules

The following sections will present detailed automation suggestions for each household based on this comparative analysis.

Home 3 - Automation for Heating

- When the heating setpoint temperature is set between 22-24°C, the outside temperature is at or below freezing, and it is daytime, the occupant frequently opts for the thermostat to override its current setting. Similarly, if the home is occupied, the heating setpoint is between 22-24°C, the thermostat's current temperature exceeds 20°C, and it is daytime with freezing external temperatures, the occupant commonly chooses the thermostat to adjust from its current state.

Home 3 - Automation for Cooling

• When the thermostat is programmed for cooling at a setpoint above 24°C (T_stp_cool_>24°C), the thermostat temperature reads above 22°C, and it is a typical weekday during the day; the occupant had preferred to override the default schedule.

• In instances where the home is occupied, the cooling setpoint of the thermostat is between 22-24°C (T_stp_cool_22-24°C), and during the daytime when the outdoor temperature is above 24°C, it is shown that the occupant preferred to override the default schedule.

Home 5 - Automation for Heating

- If the outside temperature is less than or equal to 0°C, it is the weekend, the house is occupied, the heating thermostat is set between 20-22°C, the thermostat temperature is over 20°C, and it is nighttime, then the occupant preferred to override the current setting.
- If the thermostat temperature is over 20°C, it is nighttime, and the heating thermostat is set between 22-24°C, the override observed in the current setting.

Home 5 - Automation for Cooling

- If the house is occupied, the cooling setpoint temperature is set between 22-24°C, and the outside temperature is above 24°C, then the occupant preferred to override the current setting.
- If the house is occupied, the cooling thermostat is set between 22-22°C, and the outside temperature is above 26°C, then the override observed in the current setting.

These suggestions are based on the patterns found in the dataset using the Apriori algorithm.

4.6.3 Automation Suggestions

Based on the data and rules provided for the three households, a possible automation strategy would be to make the thermostat more adaptive to the conditions. The general rules could be summarized as follows:

1. **Temperature-Based Automation**: For the heating season, the most recurring pattern for heating is when the thermostat setpoint is between 20-22°C, as observed in Home 3 and Home 5. This indicates that for both heating years across two homes, occupants frequently prefer this range. For cooling, the thermostat setpoint of 22-24°C is a consistent pattern in both households and suggesting that a temperature within this range is often preferred by occupants during the cooling season. Therefore, adjusting the default temperature ranges to align with these observed preferences could reduce the frequency of overrides
2. **Occupancy-Based Automation**: The 'Hold' state is strongly associated with the house being occupied, as observed in Home 3 and Home 5 for both heating and cooling mode. This represents that occupancy state being a determinant factor in the 'Hold' state across the given rules. Implementing an automation strategy based on occupancy would adjust temperature settings when the system detects occupancy, enhancing comfort for inhabitants.
3. **Outdoor Temperature-Based Automation**: The threshold of outdoor temperatures being less than or equal to 0°C appears in all heating rules for Home 3 and Home 5, accounting for 75% of the provided heating rules. During the cooling season, outdoor temperature greater than 20°C is a factor in both Households, representing 50% of the cooling rules presented. This suggests automating the system to adjust temperature settings when outdoor temperatures exceed these thresholds.

4. **Time of Day and Weekday-Based Automation**:

Daytime: This factor is present in all rules for Home 3 for heating and cooling mode, as well as in Home 5's cooling rule in 2017. It appears in 62.5% of the total rules, indicating that daytime plays a significant role in occupants' thermostat preferences.

Weekdays/Weekends: Weekdays are highlighted in the cooling mode rules for Home 3, while weekends are noted in the heating rules for Home 5, showing that the specific days of the week play a role in 25% of the rules.

The automation would thus be a dynamic "Hold" function that activates based on the conditions above. This would require the thermostat to be "smart" and connected, capable of collecting data on the factors listed and programmed to respond accordingly. It is also important to remember that these rules are derived from specific patterns in these households and might only apply sometimes.

Tables 4-5 to 4-8 describe the provided the most five common rules for each household, and figures 4-11 to 4-14 represent the schematic rules by Home 3 for both years. 2017 and 2018.

Table 4-6: 2018-Heat-Home 3

Home 3- 2018- Heat					
Index	Antecedents	Consequents	Support	Confidence	Lift
1	{'Home', 'T_stp_heat_22-24°C','Thermostat_Temperature_>20°C','hour_of_day_daytime}	{'Hold'}	0.50	0.75	5.21
2	{'T_stp_heat_20-22°C', 'Thermostat_Temperature_>20°C', 'T_out_≤0°C', 'hour_of_day_daytime'}	{'Hold'}	0.45	0.769	5.75
3	{'hour_of_day_daytime', 'T_stp_heat_22-24°C', 'Thermostat_Temperature_>20°C', 'Occupied'}	{'Hold'}	0.42	0.727	5.93
4	{'Home', 'T_stp_heat_22-24°C', 'Thermostat_Temperature_>20°C','T_out_≤0°C'}	{'Hold'}	0.40	0.714	4.41
5	{'T_stp_heat_22-24°C', 'Thermostat_Temperature_>20°C'}	{'Hold'}	0.38	0.9	3.89

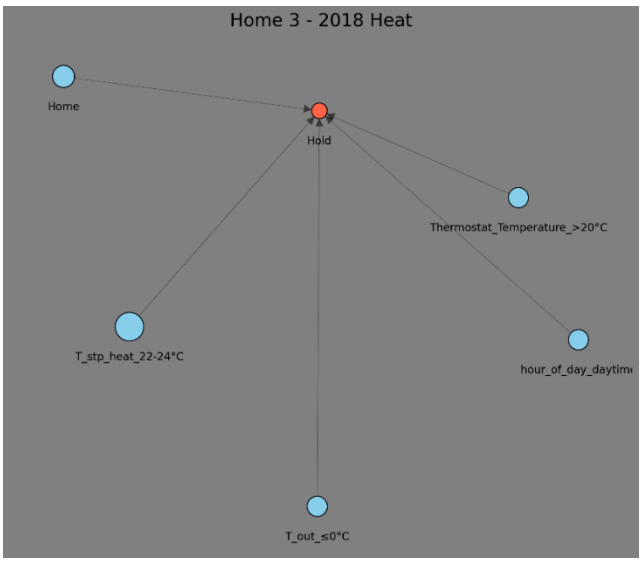

Figure 4-10: Schematic of Rules-Home 3 -2018-Heat

Table 4-7:2017-Heat-Home 3

Home 3- 2017- Heat					
Index	Antecedents	Consequents	Support	Confidence	Lift
1	{'Occupied', 'Weekend', 'T_stp_heat_22-24°C', 'Thermostat_Temperature_>20°C', 'T_out_≤0°C'}	{'Hold'}	0.65	0.601	2.693
2	{'Weekend', 'Occupied', 'T_stp_heat_20-22°C'}	{'Hold'}	0.60	0.596	2.68
3	{'Weekend', 'Thermostat_Temperature_>22°C', 'T_stp_heat_22-24°C'}	{'Hold'}	0.53	0.595	2.676
4	{'Weekday', 'Occupied', 'T_out_≤0°C', 'T_stp_heat_22-24°C'}	{'Hold'}	0.45	0.591	2.66
5	{"hour_of_day_daytime ', 'T_out_≤0°C', 'T_stp_heat_22-24°C'}	{'Hold'}	0.40	0.567	2.551

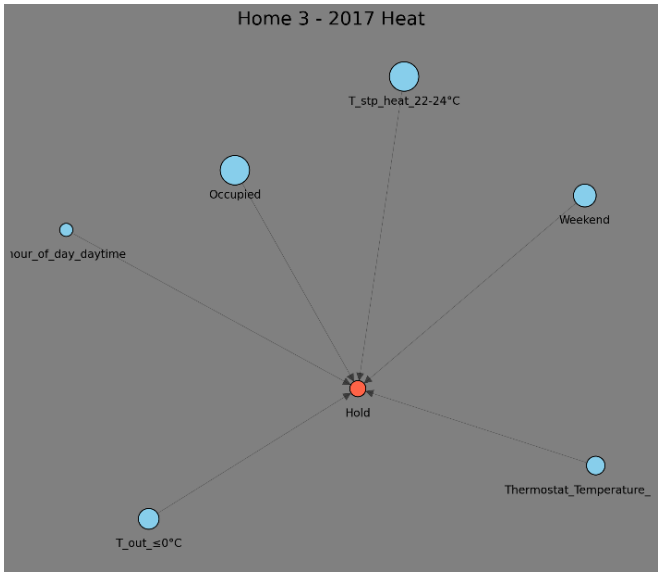

Figure 4-11: Schematic of Rules-Home 3 -2017-Heat

Table 4-8: 2018-Cool-Home 3

Home 3 2018 Cool					
index	antecedents	consequents	support	confidence	lift
1	{'Thermostat_Temperature_>22°C', 'hour_of_day_daytime', 'T_stp_cool_>24°C', 'Weekday'}	{'Hold'}	0.55	0.95	3.7
2	{'Weekday', 'hour_of_day_daytime', 'Thermostat_Temperature_>22°C', 26°C'>'T_out_>24°C'}	{'Hold'}	0.50	0.96	3.6
3	{'Weekday', 'hour_of_day_daytime', 'T_stp_cool_>22°C', 'T_out_>24°C'}	{'Hold'}	0.45	0.97	3.5
4	{'Occupied', 'hour_of_day_daytime', 'T_stp_cool_>24°C', ''}	{'Hold'}	0.43	0.98	3.4
5	{'Thermostat_Temperature_>22°C', 'Occupied', 'Weekday', 'T_out__>26°C''}	{'Hold'}	0.39	0.99	3.3

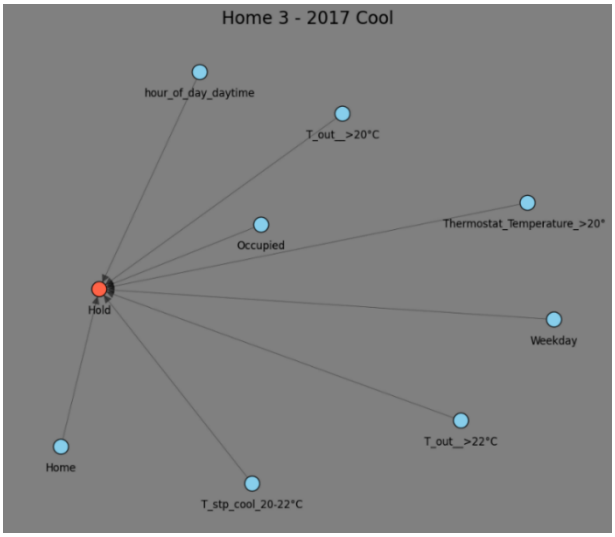

Figure 4-12: Schematic of Rules-Home 3 -2018-Cool

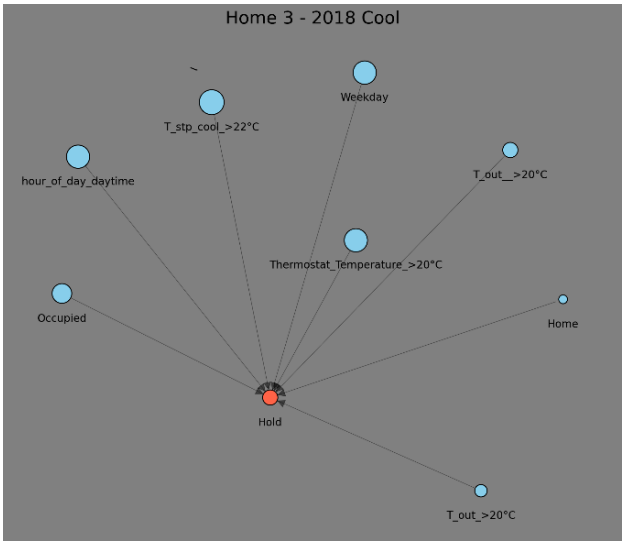

Figure 4-13: Schematic of Rules-Home 3 -2017-Cool

Table 4-9: 2017-Cool-Home 3

Home 3 2017 Cool					
Index	antecedents	consequents	support	confidence	lift
1	{'hour_of_day_daytime', 'T_stp_cool_22-24°C', 'T_out__>24°C', 'Occupied', 'Thermostat_Temperature_>22°C'}	{'Hold'}	0.53	0.95	3.8
2	{'T_stp_cool>24°C', 'Thermostat_Temperature_>22°C', 'hour_of_day_daytime'}	{'Hold'}	0.50	0.93	3.9
3	{'T_stp_cool_20-22°C', 'T_out__>20°C', 'Occupied'}	{'Hold'}	0.46	0.93	4
4	{'T_stp_cool_24-26°C', 'Occupied', 'hour_of_day_daytime'}	{'Hold'}	0.38	0.94	4.3
5	{"T_out__>24°C", 'hour_of_day_daytime'}	{'Hold'}	0.30	0.94	4.5

Chapter 5:

5. Conclusions, limitations, and future work

5.1. Conclusions

Automation has significantly reshaped the landscape of home environment control, offering myriad possibilities for enhancing comfort and energy efficiency. A key player in this transformation is the smart thermostat, which, through its data-driven approach, optimizes temperature regulation to an unprecedented degree. This research thoroughly examined a data-driven approach to support the automation of the smart thermostat for reducing human iterations, revealing the considerable potential for innovation and further refinement.

During the examination of collected data from two residential buildings over the heating and cooling seasons of 2017 to 2019, the study focused on the frequency and implications of occupancy overrides. The methodology entailed preprocessing the data, examining factors such as HVAC mode, occupancy state, temperature settings, and event schedules, and aggregating the dataset to hourly intervals. The structured approach and diverse dataset enabled us to implement our method independently for each building's data to understand if it could be generalized to households with varying characteristics.

Some exciting trends surfaced after analyzing the average setpoint and thermostat temperatures during various events and schedules. A discernable impact of occupant override behavior on the temperature settings was noticed, with a slightly warmer level set during the heating seasons and a cooler temperature during the cooling seasons across the homes and years studied. The difference in temperature during the 'Hold' and 'No_Hold' modes mainly indicated occupants' comfort preferences. Monthly fluctuations, differences between weekdays and weekends, and daily variances during weekdays were considered in our analysis of occupancy trends and setpoint temperatures. With this comprehensive approach, patterns of occupancy and setpoint temperatures can be vividly depicted, revealing user habits and preferences.

Notable differences were observed when comparing the average occupancy and setpoint temperatures for the heating and cooling seasons. Intriguing shifts in occupancy patterns were noted between the two years, with the highest average occupancy being seen on different days of the week and hours of the day. Furthermore, the month with the highest average occupancy and heating setpoint varied between the two years, revealing the fluid nature of occupant behavior and

temperature preferences. Concerning the cooling season, the average cooling setpoint temperature and overall occupancy were higher in 2017 than in 2018. Peak occupancy hours and highest cooling setpoint times also varied between the two years, underlining the dynamic interplay of multiple factors in shaping these trends.

Users with the Ecobee smart thermostat can customize the daily temperature schedule according to their preferences. The temperature can be automatically adjusted based on occupancy and manually changed through the 'hold' feature. In this study, the hold feature and the duration of the temperature hold are examined.

Understanding how the override feature works for personalized comfort and energy efficiency is essential. While holds can help users quickly adapt to changing temperature needs, optimal energy usage can be disrupted by frequent holds. Therefore, balancing comfort with energy conservation is necessary for users. To better observe the override feature, hold cycles are identified, their duration is calculated, unfinished cycles are handled, and cycle counts and durations are recorded for further analysis. Temperature differences during the hold cycle between control and setpoint temperatures and indoor and outdoor temperatures are also measured to understand user comfort levels and interactions with the thermostat.

Finally, hold cycles are categorized based on the number of daily overrides, ranging from cycles with no adjustments to cycles with three or more adjustments. This categorization further highlights user behavior and the operation of the thermostat.

The usage patterns of the 'hold' cycles vary significantly across different homes and over different years. It could highlight a diverse range of user preferences and behaviors.

These findings underscore the importance of understanding individual user patterns and preferences regarding thermostat usage. Such insights are valuable in improving smart thermostat technology.

Furthermore, analyzing the importance of different features relating to the Ecobee thermostat overrides provides essential insights into occupants' override behavior and patterns.

Key features influencing thermostat overrides include outdoor temperature, thermostat temperature, setpoint temperatures, hours of the day, and occupancy levels. The outdoor

temperature is particularly notable as a significant factor, indicating that households frequently adjust their thermostat settings in response to changing outdoor conditions. The thermostat and setpoint temperatures also play a vital role as the reference point for household comfort levels.

Average occupancy and the day hours add another layer of complexity to thermostat overrides. Households may override thermostat settings to adapt to changing occupancy levels or varying temperature preferences among household members.

Contrarily, the schedule and occupancy state are less influential in thermostat overrides. While they contribute to temperature adjustments, their impact is negligible compared to other features.

The insights gained from these features can help households and energy management systems make more informed decisions regarding temperature settings and adjustments. Furthermore, it can aid in the development of more intuitive and personalized smart thermostat systems.

In the final step, the Apriori algorithm has helped uncover patterns in thermostat override behaviors for three households over two years. These patterns suggest that indoor and outdoor temperatures, setpoint temperatures, occupancy, and time of day play significant roles in determining when and why occupants override their thermostat settings.

Specific automated actions have been suggested for each household, customized according to their unique patterns. However, a generalized strategy for automation could be distilled from these findings:

1. Adjust default temperature settings to more accurately reflect the occupants' preferences, reducing the necessity for overrides. This is evident from the frequency of 'Hold' states when the thermostat setpoint is between 18-20°C for cooling and 20-22°C for heating.
2. Implement occupancy-based overrides, as several instances indicate a 'Hold' state when the house is occupied. Here, the system would adjust temperature settings when it detects someone is at home.
3. Consider the influence of outdoor temperatures. Many rules suggest overrides occur when outdoor temperatures cross certain thresholds, i.e., less than or equal to 0°C for heating and greater than 20°C for cooling.

4. Integrate time-based automation. Overrides occur more during the day and on weekdays, indicating a need for adaptive settings based on the time and day of the week.

This data-driven approach to understanding thermostat override patterns can make residential temperature control more intuitive. It is important to note that these rules are specific to the observed households and might not apply universally. It would be beneficial to collect data from a broader range of households over a more extended period to enhance the accuracy and relevancy of the findings and trends discovered. The efficacy of suggested automation techniques can be verified and improved by testing them through simulations or prototype implementations. Further research on additional factors that may impact override behaviors, such as the integration of utility rate data, can uncover new patterns and insights. Surveying users on their reasons for overrides and their satisfaction with current thermostat programmability can provide valuable subjective perspectives to complement data mining. Larger datasets can be analyzed using advanced machine learning techniques, like deep learning, to reveal new relationships and trends. Studying override behaviors and automation opportunities in commercial buildings can identify unique patterns and insights. Conducting interdisciplinary research that combines data science, human-computer interaction, and thermal comfort studies can lead to a more holistic understanding of user behavior. Building energy simulations can be used to quantify energy efficiency opportunities of tailored automation strategies. In summary, future work should focus on expanding data collection, conducting user studies, utilizing cutting-edge analytics, simulating proposed approaches, and exploring new interdisciplinary collaborations and domains.

6. Reference:

Aggarwal, C. C. (2015). *Data mining: the textbook* (Vol. 1): Springer.

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings, 147*, 77-89.

Cabrera, D. F. M., & Zareipour, H. (2013). Data association mining for identifying lighting energy waste patterns in educational institutes. *Energy and Buildings, 62*, 210-216.

Capozzoli, A., Piscitelli, M. S., Gorrino, A., Ballarini, I., & Corrado, V. (2017). Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings. *Sustainable cities and society, 35*, 191-208.

Casals, M., Gangolells, M., Forcada, N., & Macarulla, M. (2016). Reducing lighting electricity use in underground metro stations. *Energy conversion and management, 119*, 130-141.

Combe, N., Harrison, D., Craig, S., & Young, M. S. (2012). An investigation into usability and exclusivity issues of digital programmable thermostats. *Journal of Engineering Design, 23*(5), 401-417.

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.

Day, J. K., & Gunderson, D. E. (2015). Understanding high performance buildings: The link between occupant knowledge of passive design systems, corresponding behaviors, occupant comfort and environmental satisfaction. *Building and Environment, 84*, 114-124.

Day, J. K., McIlvennie, C., Brackley, C., Tarantini, M., Piselli, C., Hahn, J., . . . Kjærgaard, M. B. (2020). A review of select human-building interfaces and their relationship to human behavior, energy use and occupant comfort. *Building and Environment, 178*, 106920.

Deng, Y. (2021). *Investigating Occupants' Hold Behaviours on Smart Thermostats using Data Mining and Machine Learning.* University of Toronto (Canada),

DuToit, S. H., Steyn, A. G. W., & Stumpf, R. H. (2012). *Graphical exploratory data analysis*: Springer Science & Business Media.

Erickson, V. L., Carreira-Perpiñán, M. Á., & Cerpa, A. E. (2014). Occupancy modeling and prediction for building energy management. *ACM Transactions on Sensor Networks (TOSN), 10*(3), 1-28.

Fan, C., Xiao, F., Madsen, H., & Wang, D. (2015). Temporal knowledge discovery in big BAS data for building energy management. *Energy and Buildings, 109*, 75-89.

Fan, C., Xiao, F., & Wang, S. (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy, 127*, 1-10.

Funde, N. A., Dhabu, M. M., Paramasivam, A., & Deshpande, P. S. (2019). Motif-based association rule mining and clustering technique for determining energy usage patterns for smart meter data. *Sustainable cities and society, 46*, 101415.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM sigmod record, 29*(2), 1-12.

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.

Hosseinihaghighi, S., Panchabikesan, K., Dabirian, S., Webster, J., Ouf, M., & Eicker, U. (2022). Discovering, processing and consolidating housing stock and smart thermostat data in support of energy end-use mapping and housing retrofit program planning. *Sustainable cities and society, 78*, 103640.

Huchuk, B., O'Brien, W., & Sanner, S. (2018). A longitudinal study of thermostat behaviors based on climate, seasonal, and energy price considerations using connected thermostat data. *Building and Environment, 139*, 199-210.

Huchuk, B., O'brien, W., & Sanner, S. (2020). Exploring smart thermostat users' schedule override behaviors and the energy consequences. *Science and Technology for the Built Environment, 27*(2), 195-210.

Huchuk, B., Sanner, S., & O'Brien, W. (2019). Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data. *Building and Environment, 160*, 106177.

Huchuk, B., Sanner, S., & O'Brien, W. (2022). Evaluation of data-driven thermal models for multi-hour predictions using residential smart thermostat data. *Journal of Building Performance Simulation, 15*(4), 445-464.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association, 113*(522), 626-636.

Liu, W., Chen, G., Yan, B., Zhou, Z., Du, H., & Zuo, J. (2015). Hourly operation strategy of a CCHP system with GSHP and thermal energy storage (TES) under variable loads: a case study. *Energy and Buildings, 93*, 143-153.

Meier, A. (2010). How people actually use thermostats.

Meier, A., Aragon, C., Peffer, T., Perry, D., & Pritoni, M. (2011). Usability of residential thermostats: Preliminary investigations. *Building and Environment, 46*(10), 1891-1898.

Melfi, R., Rosenblum, B., Nordman, B., & Christensen, K. (2011). *Measuring building occupancy using existing network infrastructure.* Paper presented at the 2011 International Green Computing Conference and Workshops.

Méndez, J. I., Ponce, P., Meier, A., Peffer, T., Mata, O., & Molina, A. (2020). *S 4 product design framework: a gamification strategy based on type 1 and 2 fuzzy logic.* Paper presented at the Smart Multimedia: Second International Conference, ICSM 2019, San Diego, CA, USA, December 16–18, 2019, Revised Selected Papers 2.

Moon, J. W., & Han, S.-H. (2011). Thermostat strategies impact on energy consumption in residential buildings. *Energy and Buildings, 43*(2-3), 338-346.

Palensky, P., & Dietrich, D. (2011). Industrial Informatics. *IEEE Transactions on, 7*(3), 381.

Panchabikesan, K., Haghighat, F., & El Mankibi, M. (2021). Data driven occupancy information for energy simulation and energy use assessment in residential buildings. *Energy, 218*, 119539.

Pang, Z., Chen, Y., Zhang, J., O'Neill, Z., Cheng, H., & Dong, B. (2021). How much HVAC energy could be saved from the occupant-centric smart home thermostat: A nationwide simulation study. *Applied Energy, 283*, 116251.

Parker, D., Sutherland, K., Chasar, D., & Center, F. S. E. (2016). Evaluation of the space heating and cooling energy savings of smart thermostats in a hot-humid climate using long-term data. *ACEEE Summer Study Energy Eff. Build, 2016*, 15.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Peffer, T., Pritoni, M., Meier, A., Aragon, C., & Perry, D. (2011). How people use thermostats in homes: A review. *Building and Environment, 46*(12), 2529-2541.

Pigg, S., & Center ofWisconsin, E. (2000). *Programmable thermostats that go berserk? Taking a social perspective on space heating in Wisconsin.* Paper presented at the Proc. ACEEE Summer Study on Energy Efficiency in Buildings.

Policy, I. E. A. D. o. S. E. (2013). *Transition to sustainable buildings: strategies and opportunities to 2050*: Organization for Economic.

Pritoni, M., Meier, A. K., Aragon, C., Perry, D., & Peffer, T. (2015). Energy efficiency and the misuse of programmable thermostats: The effectiveness of crowdsourcing for understanding household behavior. *Energy Research & Social Science, 8*, 190-197.

Ren, X., Yan, D., & Hong, T. (2015). Data mining of space heating system performance in affordable housing. *Building and Environment, 89*, 1-13.

Ren, X., Zhang, C., Zhao, Y., Boxem, G., Zeiler, W., & Li, T. (2019). A data mining-based method for revealing occupant behavior patterns in using mechanical ventilation systems of Dutch dwellings. *Energy and Buildings, 193*, 99-110.

Sandels, C., Widén, J., Nordström, L., & Andersson, E. (2015). Day-ahead predictions of electricity consumption in a Swedish office building from weather, occupancy, and temporal data. *Energy and Buildings, 108*, 279-290.

Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests.

Shen, W., Newsham, G., & Gunay, B. (2017). Leveraging existing occupancy-related data for optimal control of commercial office buildings: A review. *Advanced Engineering Informatics, 33*, 230-242.

Stopps, H., & Touchie, M. (2021). Smart choice or flawed approach? An exploration of connected thermostat data fidelity and use in data-driven modelling in high-rise residential buildings. *Journal of Building Performance Simulation, 14*(6), 793-813.

Stopps, H., & Touchie, M. F. (2020). Managing thermal comfort in contemporary high-rise residential buildings: Using smart thermostats and surveys to identify energy efficiency and comfort opportunities. *Building and Environment, 173*, 106748.

Stopps, H., & Touchie, M. F. (2021). Residential smart thermostat use: An exploration of thermostat programming, environmental attitudes, and the influence of smart controls on energy savings. *Energy and Buildings, 238*, 110834.

Tomat, V., Vellei, M., Ramallo-González, A. P., González-Vidal, A., Le Dréau, J., & Skarmeta-Gómez, A. (2022). Understanding patterns of thermostat overrides after demand response events. *Energy and Buildings, 271*, 112312.

Tukey, J. W., & Tukey, P. A. (1985). *Computer graphics and exploratory data analysis: An introduction.* Paper presented at the Proceedings of the sixth annual conference and exposition: computer graphics.

Urban, B., Elliott, D., & Sachs, O. (2012). Towards better modeling of residential thermostats. *SimBuild2012, Madison, WI, USA*.

Urban, B., & Gomez, C. (2013). *A case for thermostat user models.* Paper presented at the 13th Conference of International Building Performance Simulation Association.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., & Ahrentzen, S. (2018). Random Forest based hourly building energy prediction. *Energy and Buildings, 171*, 11-25.

Yan, D., & Hong, T. (2018). Definition and simulation of occupant behavior in buildings. *International Energy Agency, EBC Annex, 66*.

Yan, L., Liu, M., Xue, K., & Zhang, Z. (2020). A study on temperature-setting behavior for room air conditioners based on big data. *Journal of Building Engineering, 30*, 101197.

Yu, Z. J., Haghighat, F., Fung, B. C., & Zhou, L. (2012). A novel methodology for knowledge discovery through mining associations between building operational data. *Energy and Buildings, 47*, 430-440.

Zhang, C., Xue, X., Zhao, Y., Zhang, X., & Li, T. (2019). An improved association rule mining-based method for revealing operational problems of building heating, ventilation and air conditioning (HVAC) systems. *Applied Energy, 253*, 113492.

Zhang, X. M., Grolinger, K., Capretz, M. A., & Seewald, L. (2018). *Forecasting residential energy consumption: Single household perspective.* Paper presented at the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA).

Zheng, Z., Kohavi, R., & Mason, L. (2001). *Real world performance of association rule algorithms.* Paper presented at the Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.

7. Appendix

Appendix A: Average Temperatures Difference for Hold and Non-Hold

Home 3 – 2017

Table 7-1: Average Temperature difference - Home 3-2017

		Hold			No Hold			Difference		
Season	**Schedule**	Setpoint Temperature(°C)	Thermostat Temperature(°C)	Outdoor Temperature (°C)	Setpoint Temperature (°C)	Thermostat Temperature(°C)	Outdoor Temperature(°C)	Setpoint Temperature (°C)	Thermostat Temperature(°C)	Outdoor Temperature(°C)
Heat	**Home**	22.28	22.12	-6.13	22.09	22.01	-5.01	0.19	0.11	-1.11
	Sleep	22.18	21.79	-14.05	19.96	20.46	-6.60	2.22	1.32	-7.45
Cool	**Home**	22.42	22.09	19.43	22.71	22.58	20.34	-0.29	-0.49	-0.90
	Sleep	22.64	21.49	18.00	21.47	21.37	17.59	1.17	0.12	0.41

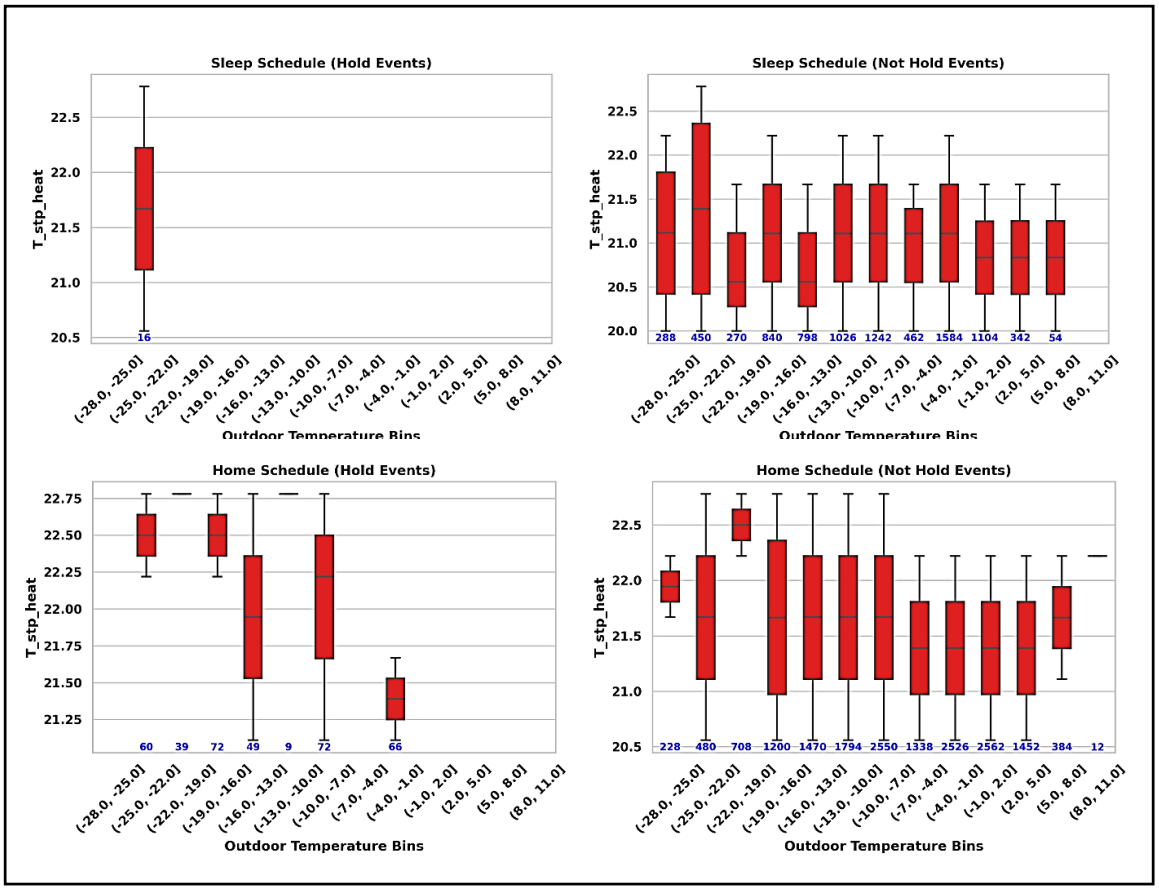

Figure 7-1: Average Temperature difference - Home 3-2017-Heating Season

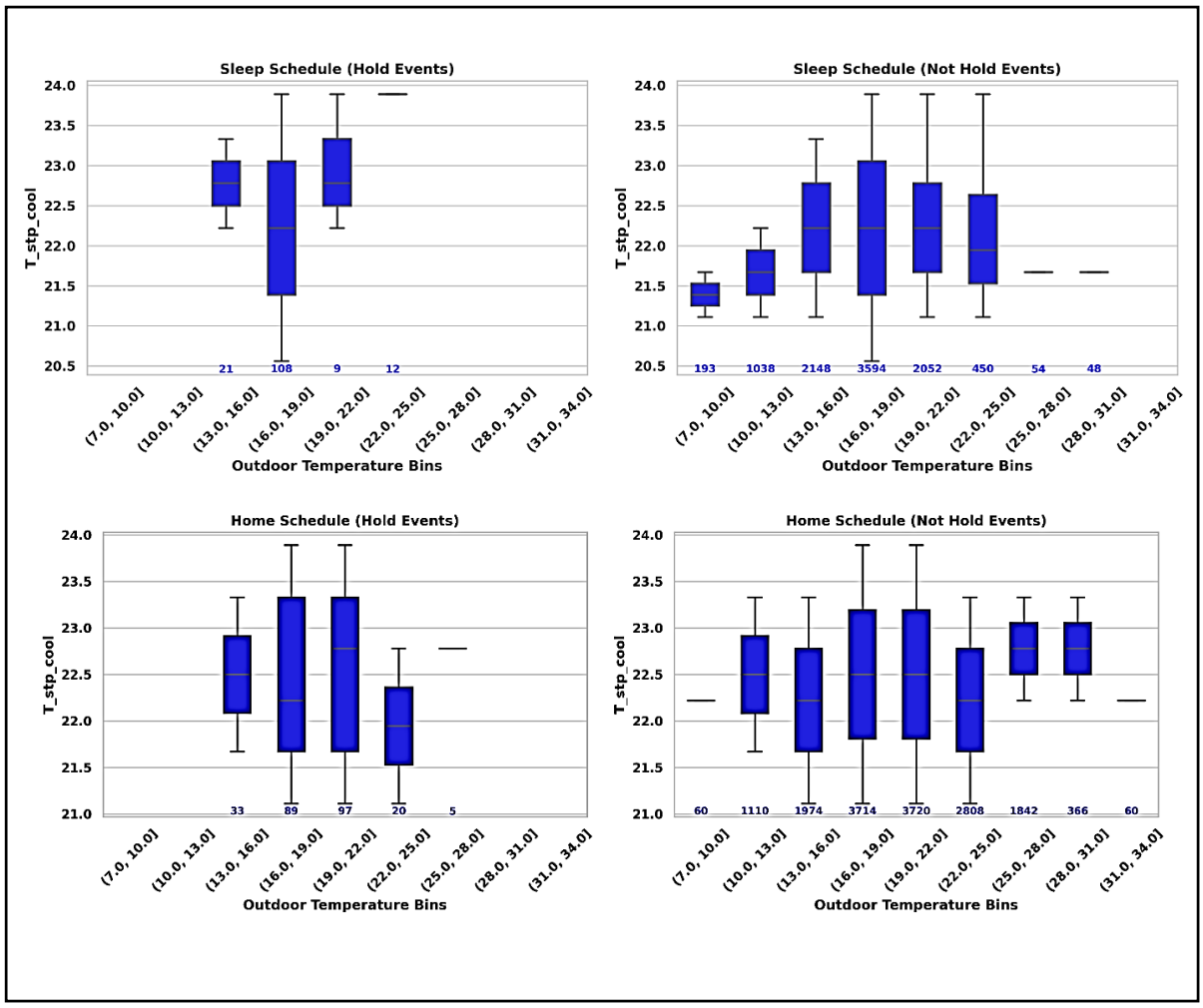

Figure 7-2: Average Temperature difference - Home 3-2017-Cooling Season

Home 5 – 2018

Table 7-2: Average Temperature difference – Home5-2018

		Hold			No Hold			Difference		
Season	**Schedule**	Setpoint Temperature(°C)	Thermostat Temperature(°C)	Outdoor Temperature (°C)	Setpoint Temperature (°C)	Thermostat Temperature(°C)	Outdoor Temperature(°C)	Setpoint Temperature (°C)	Thermostat Temperature(°C)	Outdoor Temperature(°C)
Heat	**Home**	2137	21.53	-11.33	20.40	21.05	-3.35	0.97	0.48	-7.97
	Sleep	21.13	21.31	-8.70	18.91	19.98	-4.79	2.22	1.32	-3.91
Cool	**Home**	21.68	21.30	22.25	23.22	21.57	19.99	-1.55	-0.28	2.22
	Sleep	21.31	20.67	17.59	24.50	21.04	15.50	-3.67	-0.36	2.29

Heating and Cooling Mode – 2018

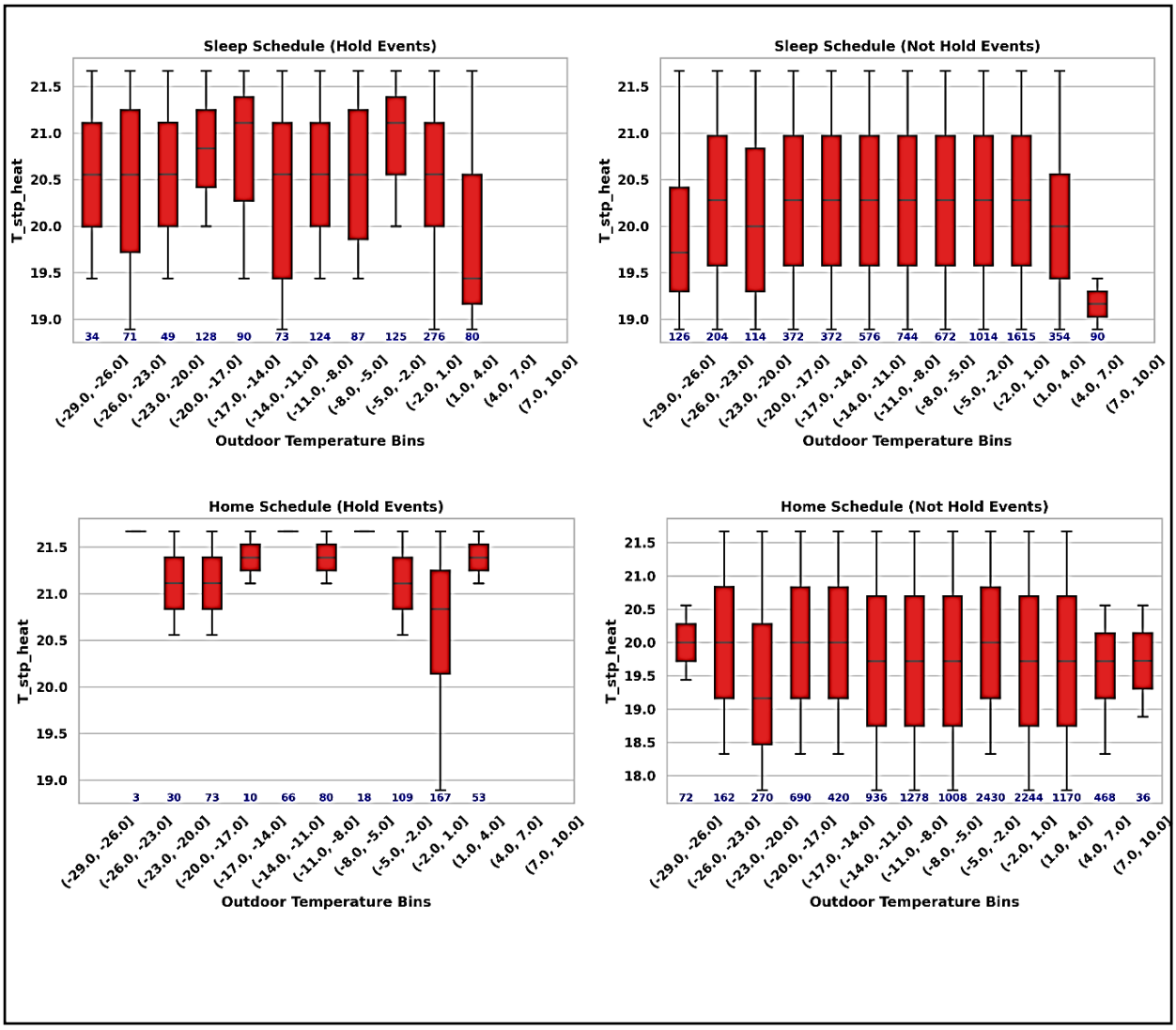

Figure 7-3: Average Temperature difference - Home 5-2018-Heating Season

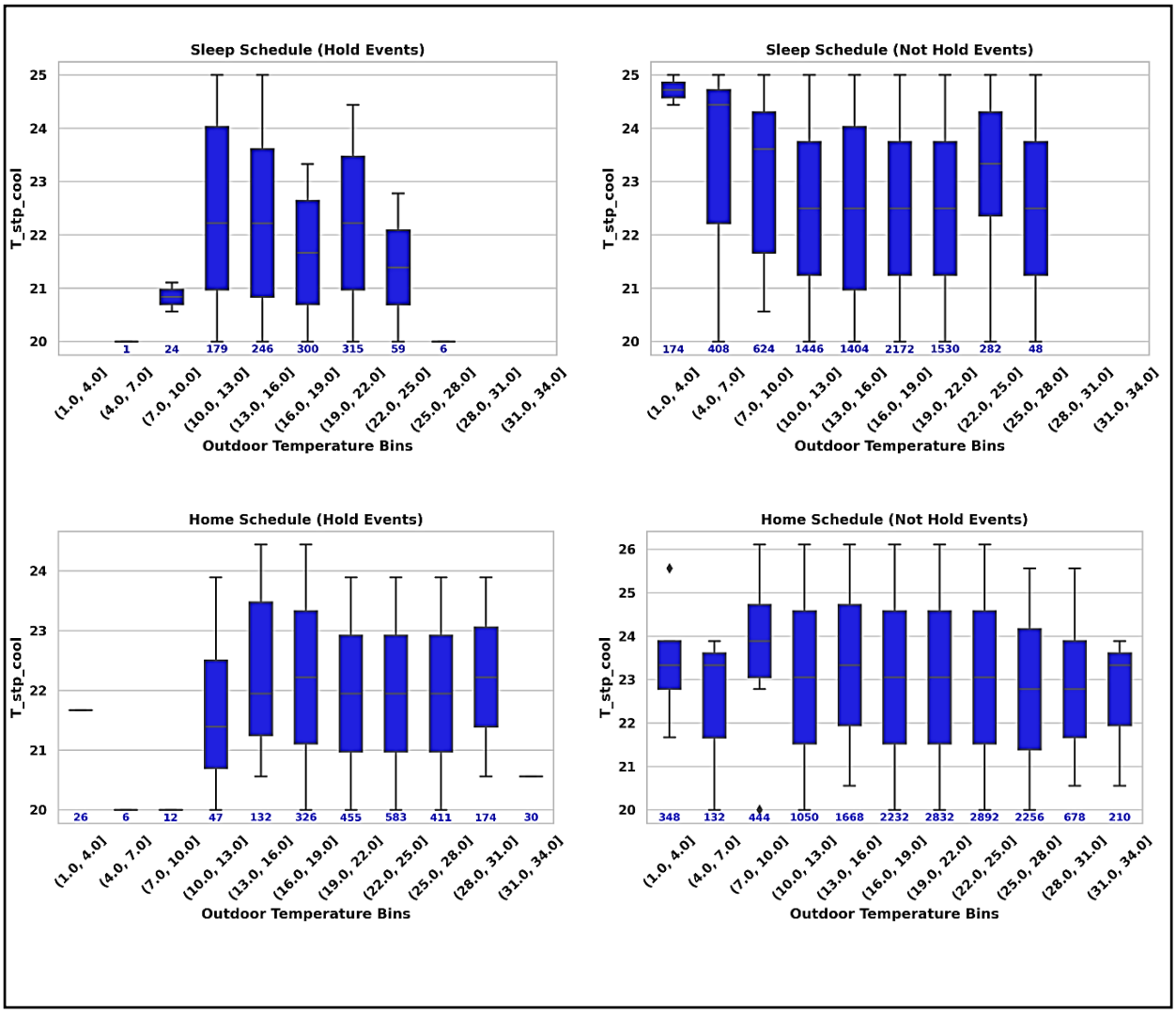

Figure 7-4: Average Temperature difference - Home 5-2018-Cooling Season

Home 5 – 2017

Table 7-3: Average Temperature difference – Home5-2017

		Hold			No Hold			Difference		
Season	**Schedule**	Setpoint Temperature (°C)	Thermostat Temperature (°C)	Outdoor Temperature (°C)	Setpoint Temperature (°C)	Thermostat Temperature (°C)	Outdoor Temperature (°C)	Setpoint Temperature (°C)	Thermostat Temperature	Outdoor Temperature (°C)
Heat	**Home**	21.18	21.47	-3.91	20.28	20.82	-2.89	0.89	0.65	-1.02
	Sleep	21.23	21.30	-6.87	18.79	19.88	-3.92	2.45	1.43	-2.96
Cool	**Home**	21.12	21.29	21.08	23.33	22.01	19.08	-2.21	-0.72	1.99
	Sleep	20.37	20.14	17.16	24.98	21.24	15.31	-4.60	-1.10	1.85

Heating and Cooling Mode – 2017

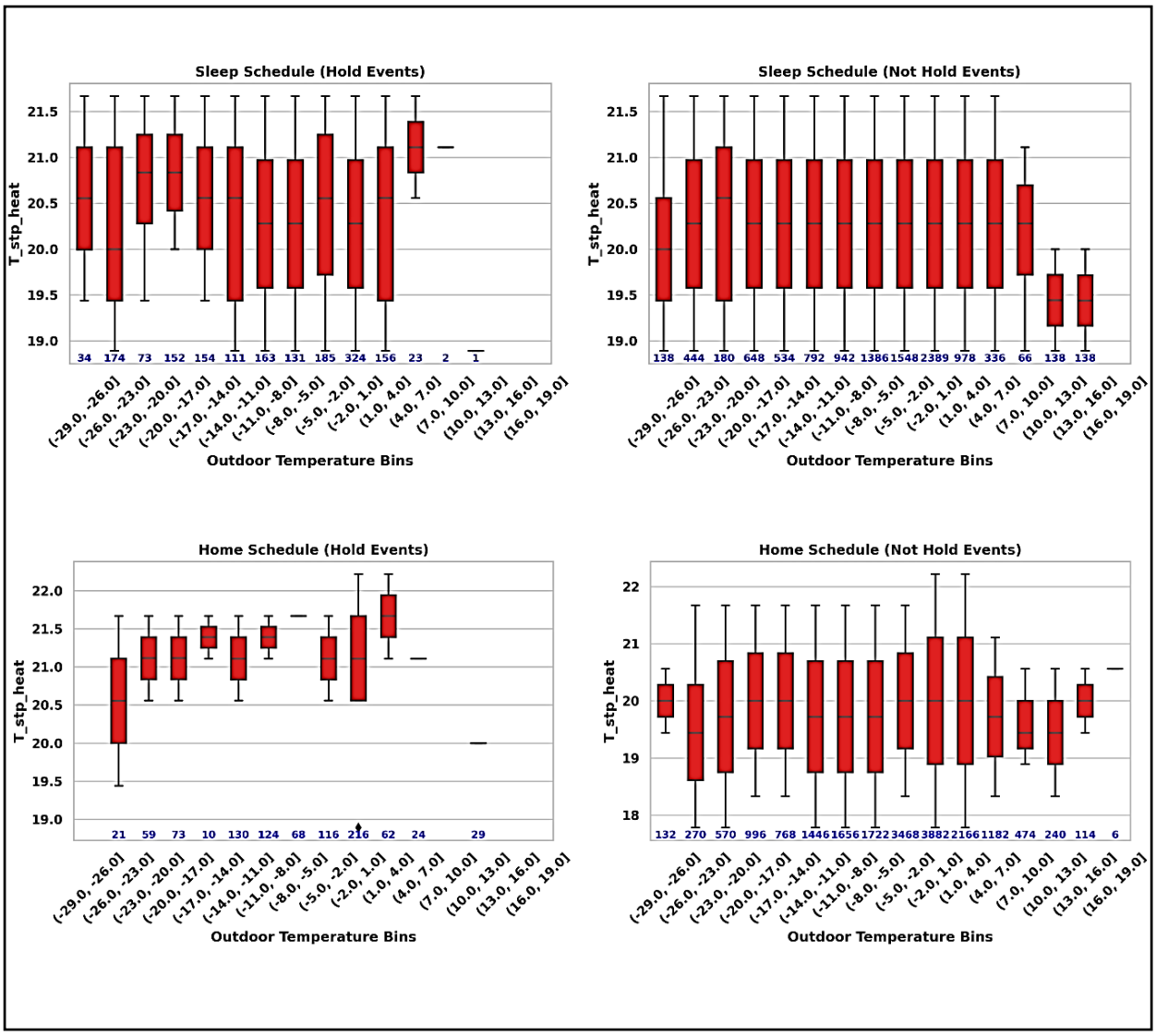

Figure 7-5: Average Temperature difference - Home 5-2017- Heating Season

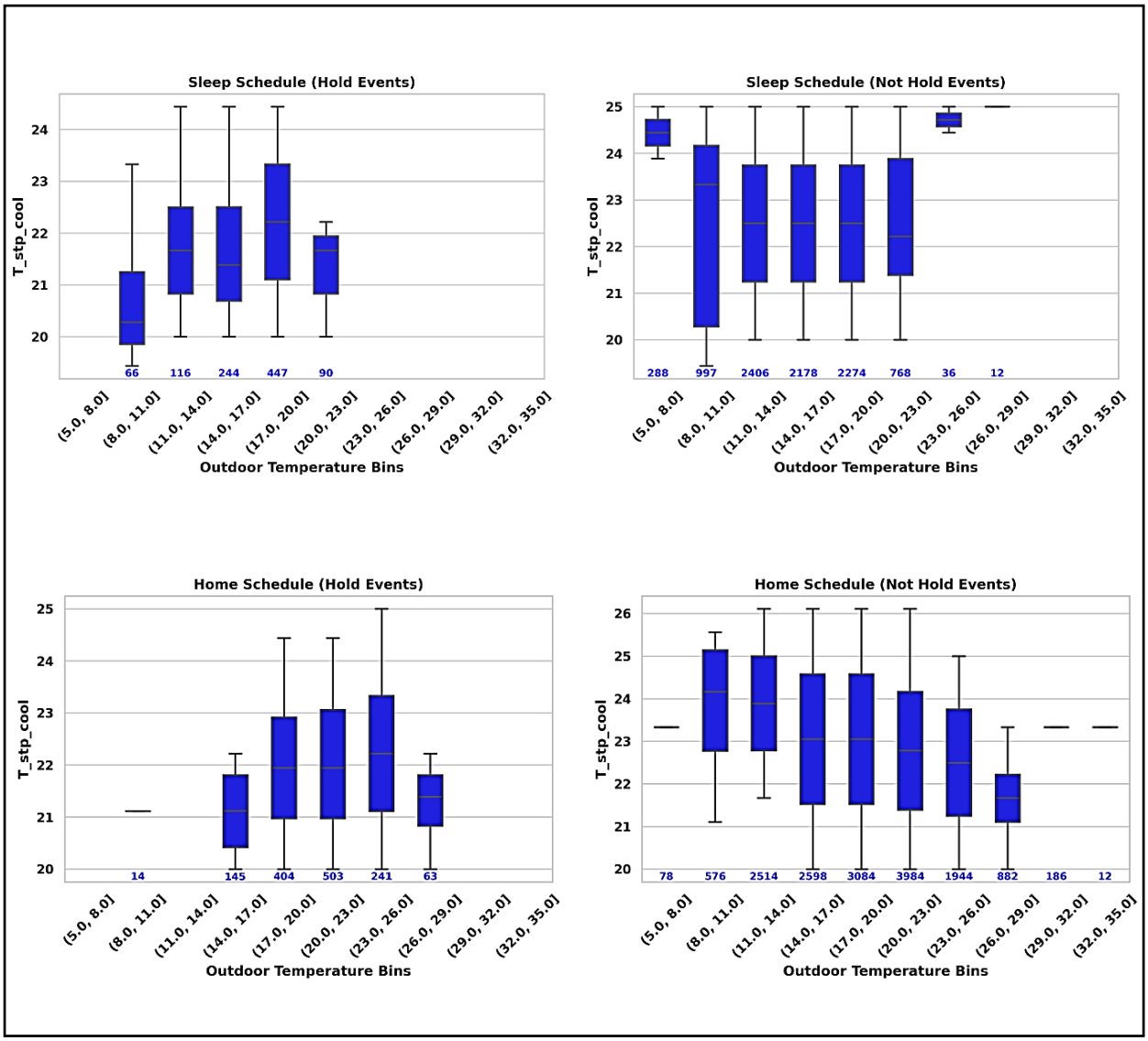

Figure 7-6: Average Temperature difference - Home 5-2017- Cooling Season

Appendix B: Average Occupancy and Setpoint Temperatures

Home 3 - 2017- Heat

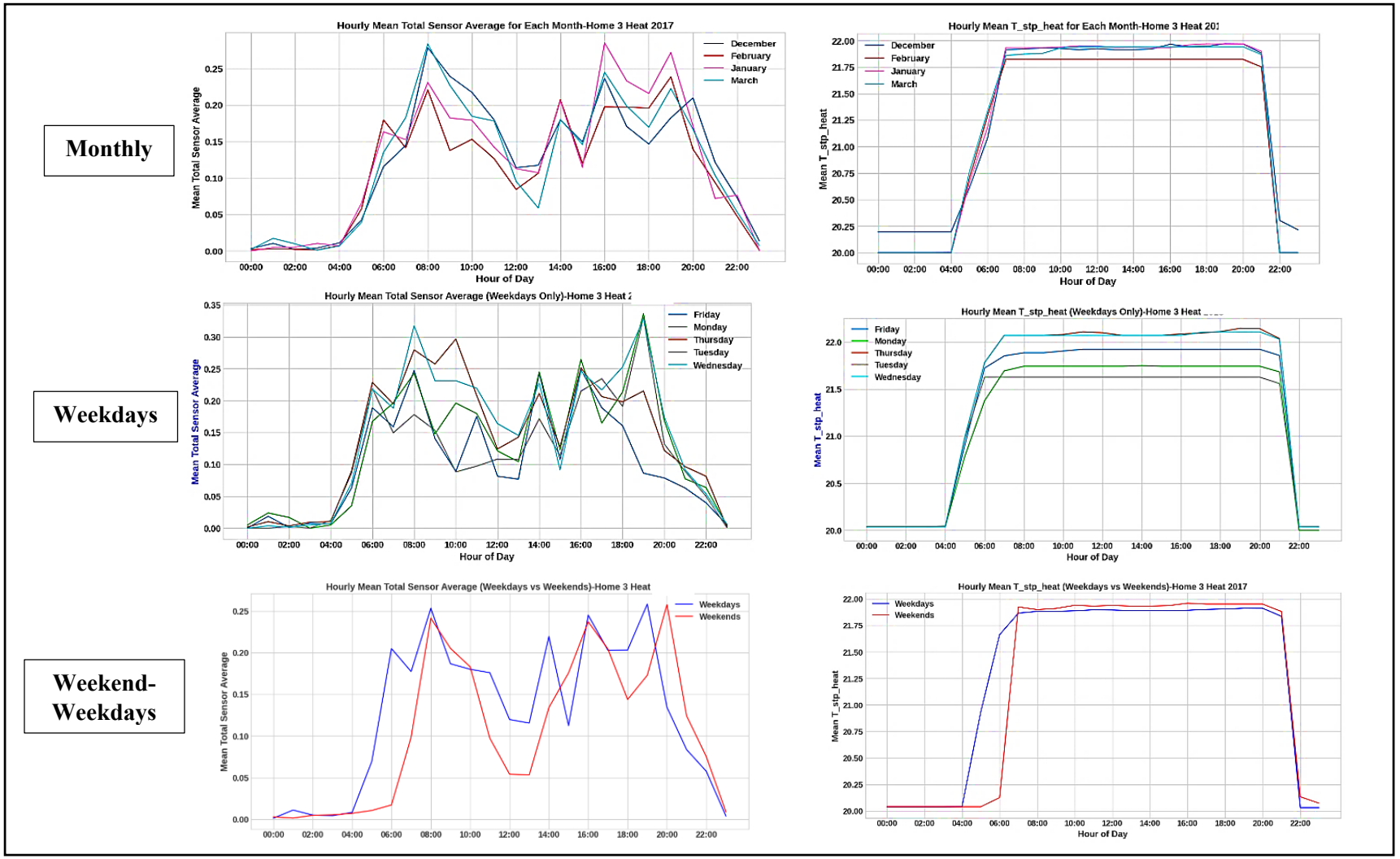

Figure 7-7: Average Occupancy and Setpoint Temperatures- Home 3-2017-Heating Season

Home 3 - 2017- Cool

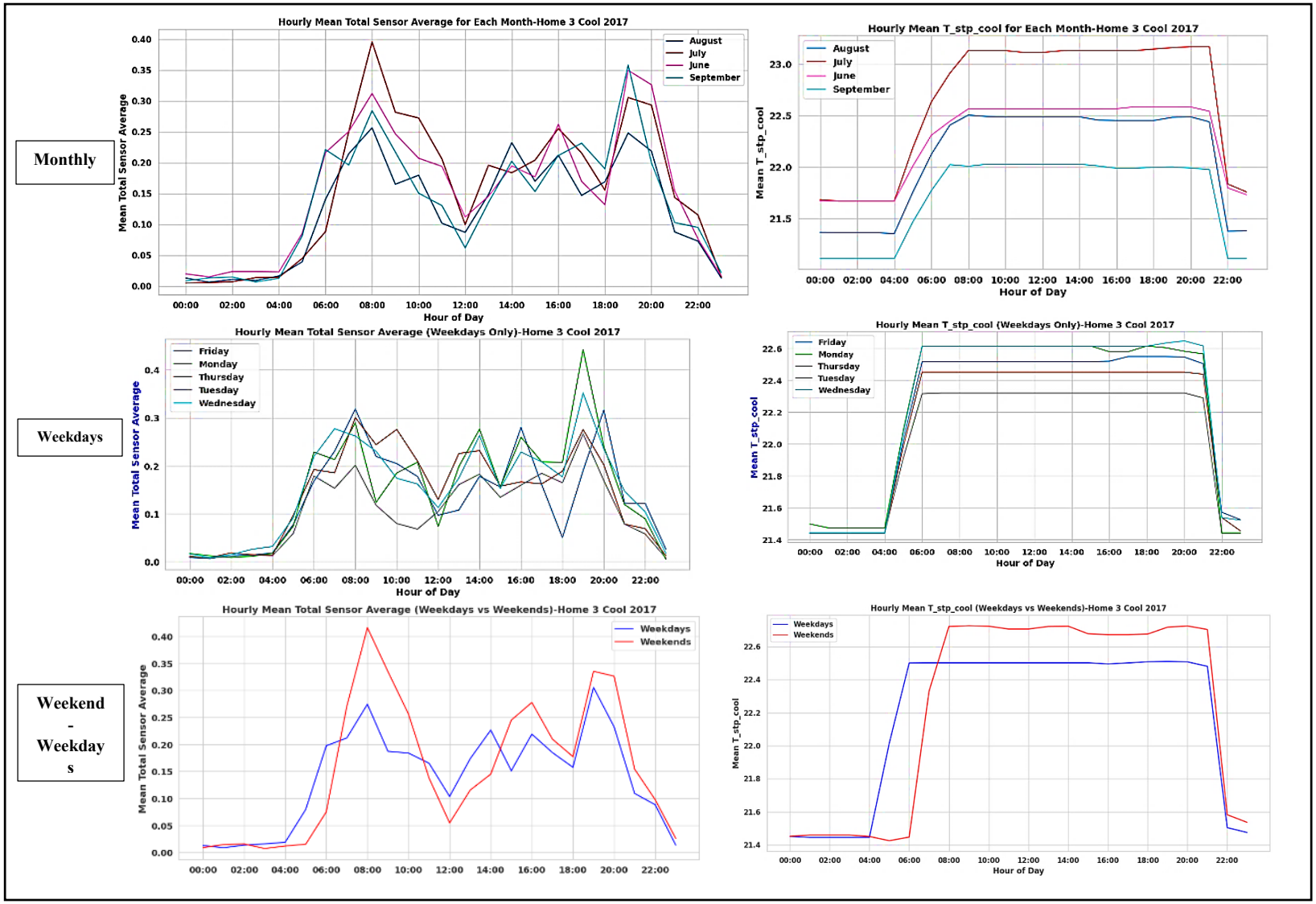

Figure 7-8: Average Occupancy and Setpoint Temperatures- Home 3-2017- Cooling Season

Home 5 – 2018 – Heat

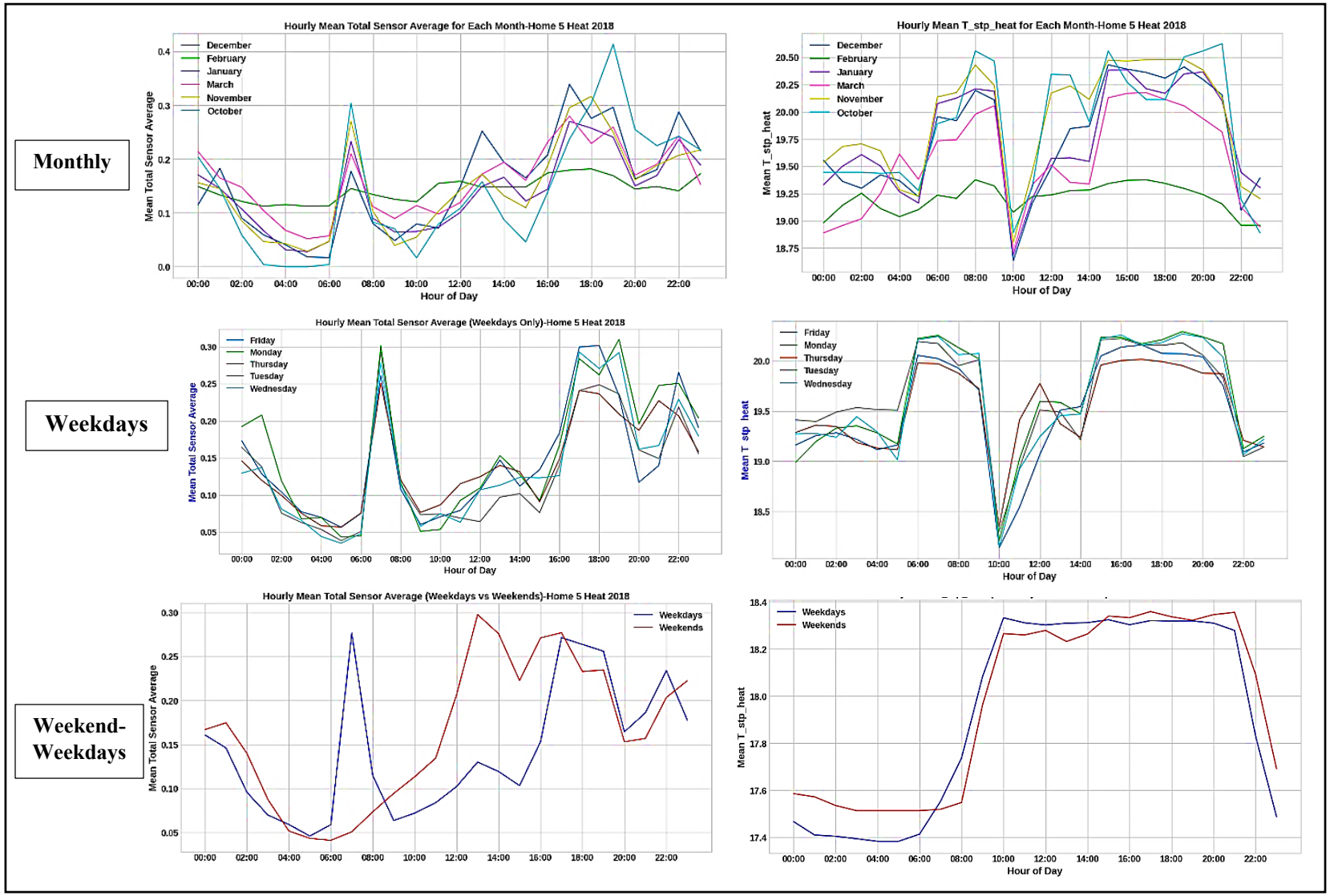

Figure 7-9: Average Occupancy and Setpoint Temperatures- Home 5-2018- Heating Season

Home 5 – 2018 – Cool

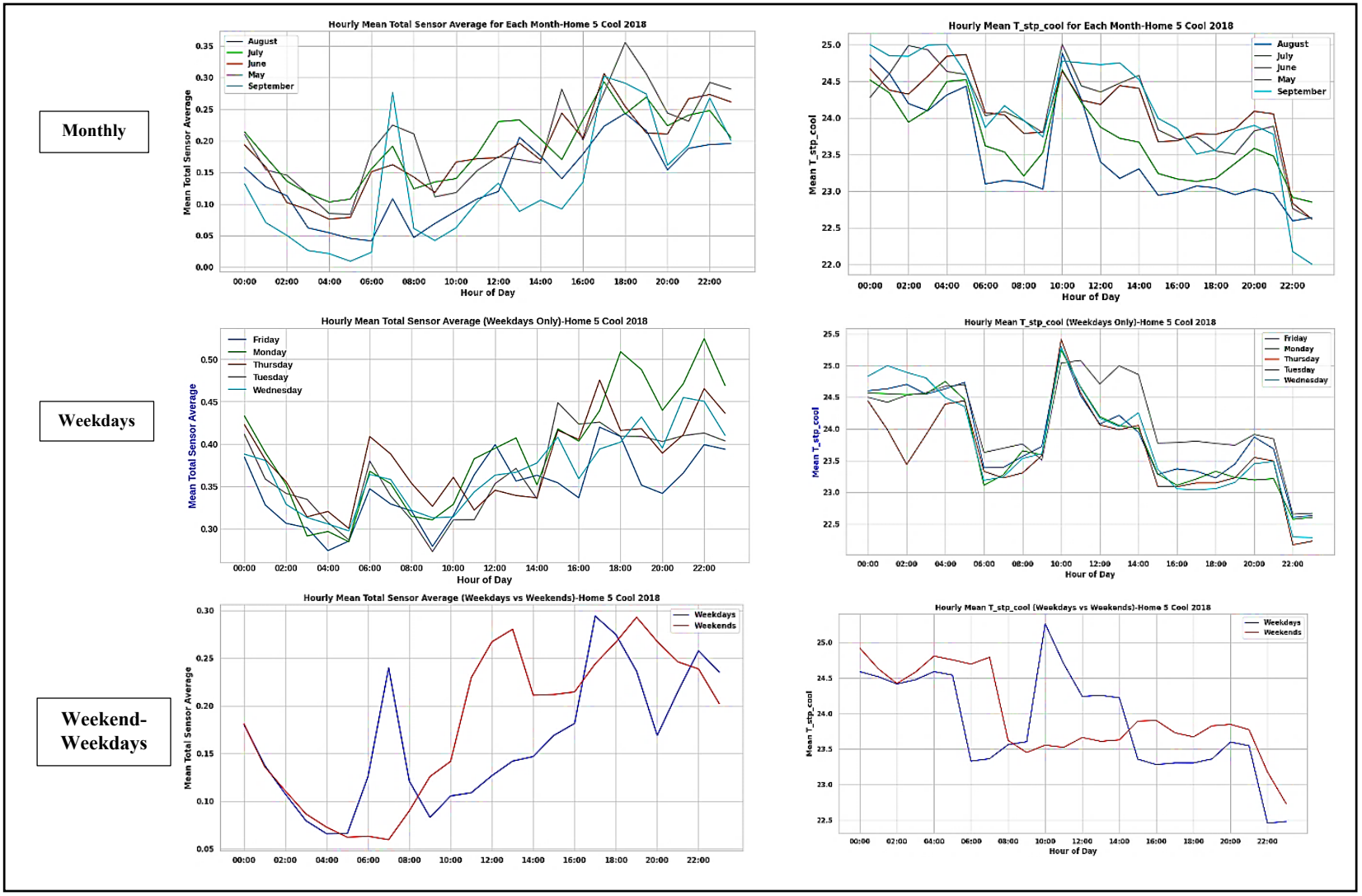

Figure 7-10: Average Occupancy and Setpoint Temperatures- Home 5-2018-Cooling Season

Home 5 – 2017 – Heat

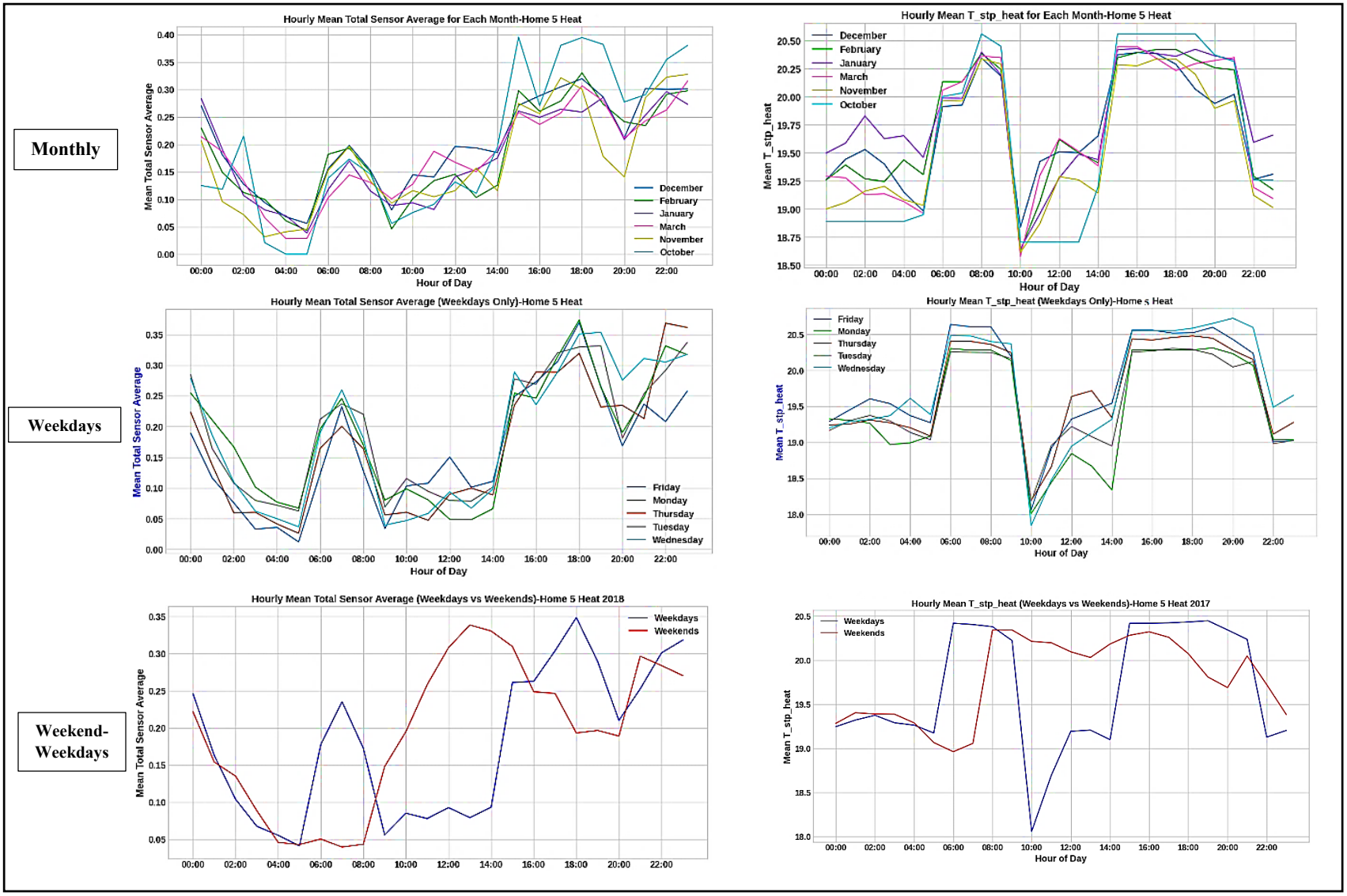

Figure 7-11: Average Occupancy and Setpoint Temperatures- Home 5-2017-Heating Season

Home 5 – 2017 – Cool

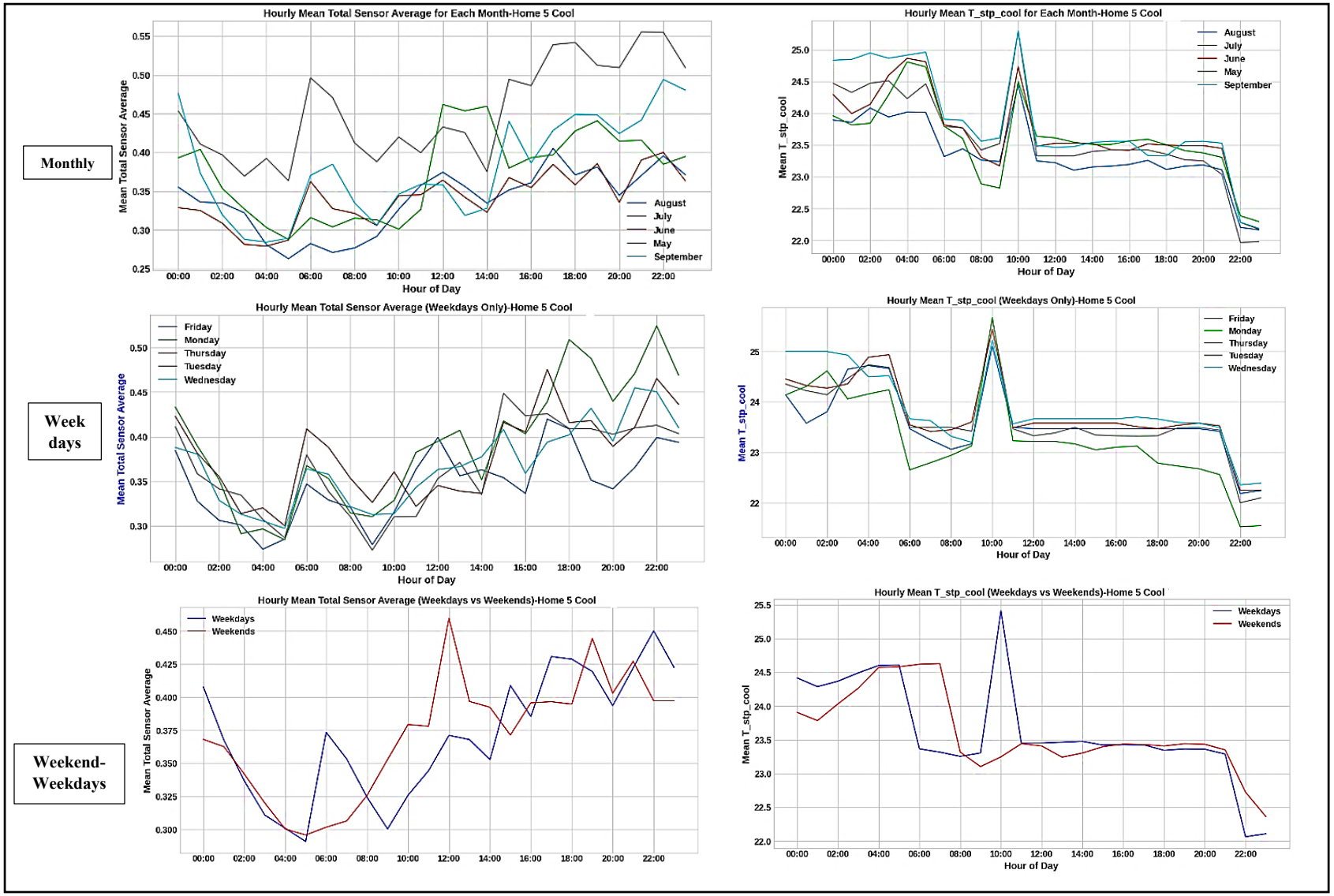

Figure 7-12: Average Occupancy and Setpoint Temperatures- Home 5-2017- Cooling Season

Appendix C: Association Rule Mining

Table 7-4 to 7-7 describe the provided rules by Home 5 for both years. 2017 and 2018.

Table 7-4: Home 5-Heat-2017

Home 5 2017 Heat					
index	antecedents	consequents	support	confidence	lift
1074	{'Thermostat_Temperature_>20°C', 'hour_of_day_nighttime', 'T_stp_heat_20-22°C'}	{'Hold'}	0.125	0.83	3
2904	{'Thermostat_Temperature_>20°C', 'hour_of_day_nighttime', 'T_out_≤0°C', 'T_stp_heat_20-22°C'}	{'Hold'}	0.135	0.84	3
221	{'hour_of_day_nighttime', 'T_stp_heat_20-22°C'}	{'Hold'}	0.145	0.85	2.9
1045	{'hour_of_day_nighttime', 'T_out_≤0°C', 'T_stp_heat_20-22°C'}	{'Hold'}	0.155	0.86	2.9

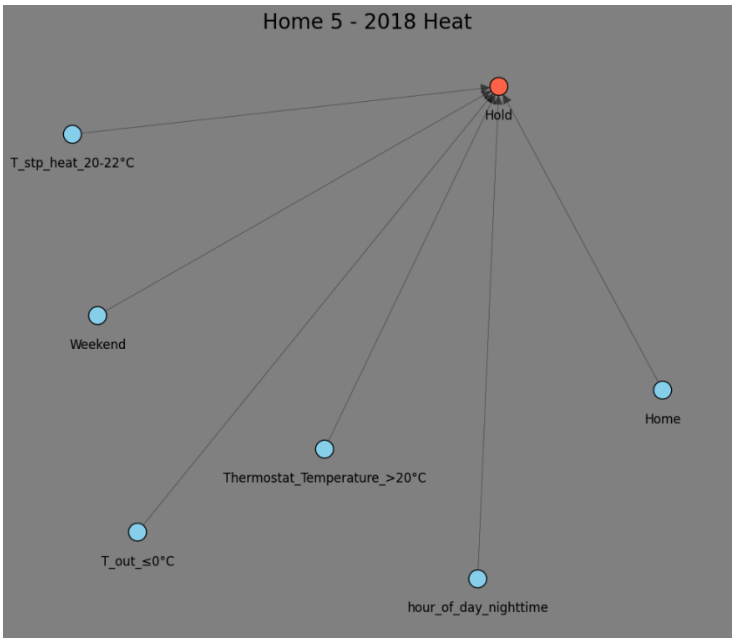

Figure 7-13: Schematic of Rules-Home 5-2018-Heat

Table 7-5: Home 5-Heat-2018

Home 5 2018 Heat					
index	antecedents	consequents	support	confidence	lift
1	{'T_out_≤0°C', 'Weekend', 'Home', 'T_stp_heat_20-22°C', 'Thermostat_Temperature_>20°C', 'hour_of_day_nighttime'}	{'Hold'}	0.55	0.86	2.7
2	{'T_out_≤0°C', 'Weekend', Home 'T_stp_heat_20-22°C', 'Occupied'}	{'Hold'}	0.50	0.94	3
3	{'Thermostat_Temperature_>20°C', 'Home', 'Weekday', 'T_stp_heat_20-22°C'}	{'Hold'}	0.42	0.94	3.1
4	{'Home', 'T_out_≤0°C', 'hour_of_day_nighttime', 'T_stp_heat_20-22°C'}	{'Hold'}	0.40	0.94	3.2
5	{'Home', 'T_out_≤0°C', 'Weekend', 'T_stp_heat_20-22°C',' T_out_≤0°C'}	{'Hold'}	0.37	0.94	3.3

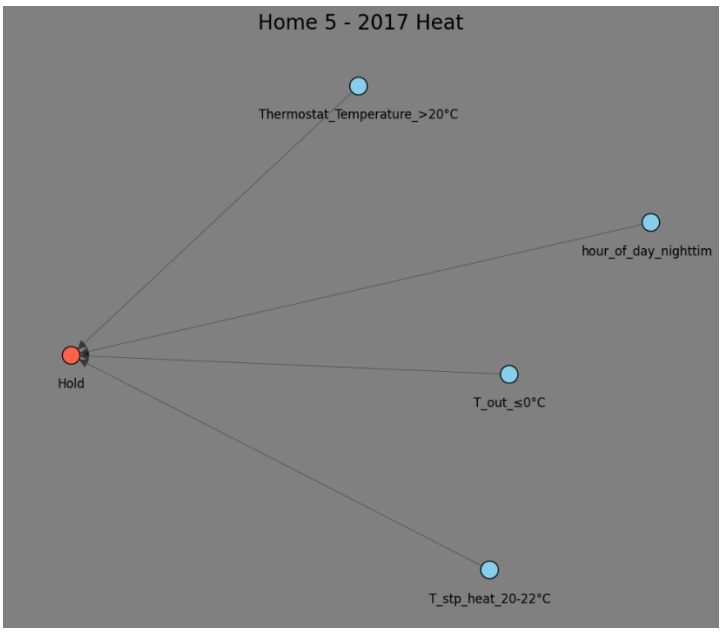

Figure 7-14: Schematic of Rules-Home 5-2017-Heat

Table 7-6: Home 5-Cool-2018

Home 5 2018 Cool					
Index	antecedents	consequents	support	confidence	lift
1	{'Home', 'T_stp_cool_24-26°C', 'T_out__>24°C'}	{'Hold'}	0.60	0.94	8.2
2	{'Home', 'Thermostat_Temperature_>20°C', 'T_stp_cool_20-22°C'}	{'Hold'}	0.51	0.95	8.1
3	{'Home', 'hour_of_day_nighttime', 'Occupied', 'T_out__>24°C'}	{'Hold'}	0.46	0.96	8.3
4	{'Weekday', 'hour_of_day_nighttime', 'T_stp_cool_22-24°C'}	{'Hold'}	0.35	0.95	8
5	{'Home', 'Thermostat_Temperature_>22°C', 'T_stp_cool_22-24°C', 'Occupied'}	{'Hold'}	0.30	0.96	8.1

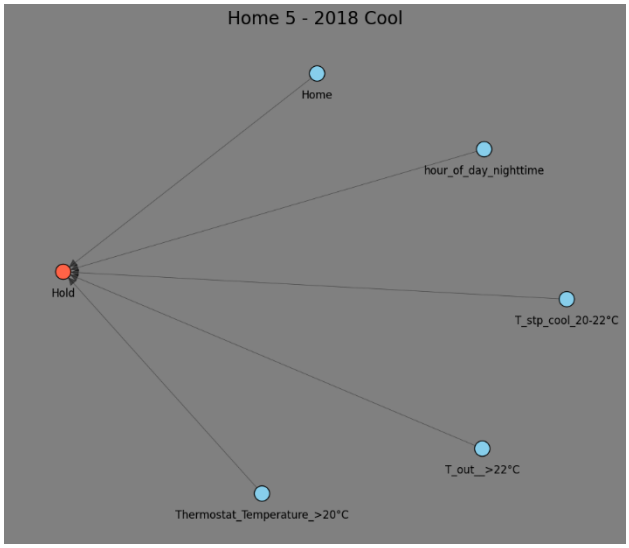

Figure 7-15: Schematic of Rules-Home 5-2018-Cool

Table 7-7: Home 5-Cool-2017

Home 5 2017 Cool					
index	antecedents	consequents	support	confidence	lift
1	{'Unoccupied', 'Weekend', 'T_stp_cool_24-26°C'}	{'Hold'}	0.50	0.91	3.4
2	{'Home', 'T_out__>24°C', 'T_stp_cool_22-24°C'}	{'Hold'}	0.45	0.92	3.3
3	{'Thermostat_Temperature_>22°C', 'Home', 'T_out__>24°C', 'T_stp_cool_20-22°C'}	{'Hold'}	0.40	0.92	3.2
4	{'Thermostat_Temperature_>20°C', 'Occupied', 'T_stp_cool_20-22°C'}	{'Hold'}	0.38	0.93	3.1
5	{'Home', 'Weekend', 'T_out__>26°C', 'T_stp_cool_22-24°C',"Unoccupied"}	{'Hold'}	0.30	0.93	3

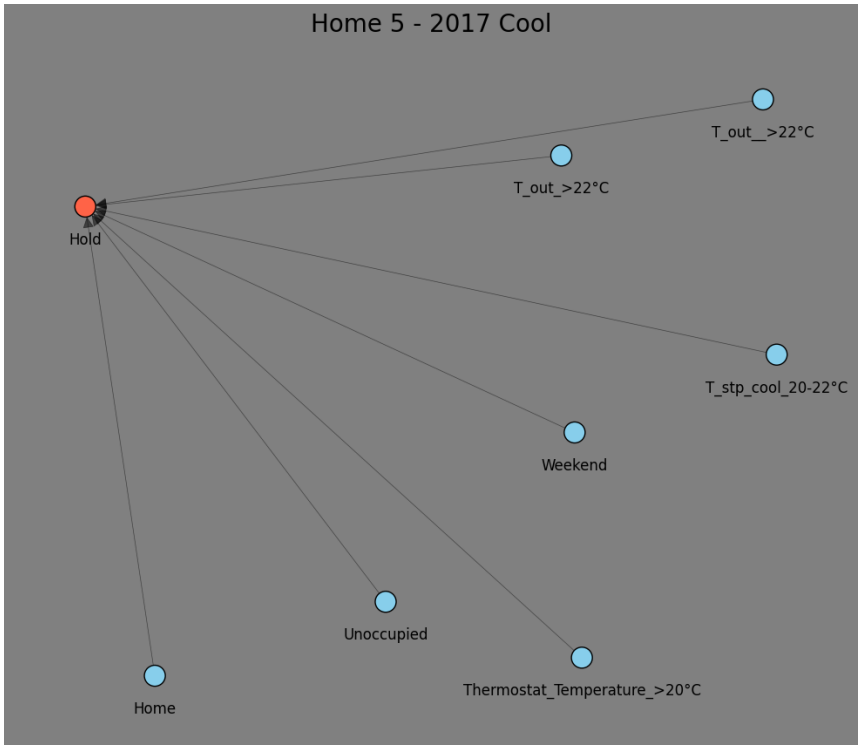

Figure 7-16: Schematic of Rules-Home 5-2017-Cool