

**Development and implementation of occupancy-based predictive
controls for modulation of air handling units' outdoor air dampers**

by

Brodie W. Hobson

A thesis submitted to the
Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Master of Applied Science

in

Civil Engineering

Department of Civil and Environmental Engineering

Carleton University

Ottawa, Ontario

April 2020

© Copyright

Brodie W. Hobson, 2020

Abstract

Current practices in the field of commercial and institutional building operation require ventilation to be provided to spaces assuming at or near full occupancy. However, full occupancy in the field of heating, ventilation, and air conditioning (HVAC) design is crudely estimated early in the design process and often does not reflect the actual maximum occupancy of the building. This results in the chronic overventilation of buildings, which wastes significant energy in heating climates where conditioning outdoor air is the main driver of HVAC energy use. Therefore, in forced air systems with air handling units, reducing outdoor air damper positions based on the predicted occupancy of buildings has the potential to generate significant energy savings. This research explores the potential of these occupancy-based predictive controls for outdoor air dampers in a case study building in Ottawa, Canada. Different sensors and model formalisms for estimating occupancy were evaluated. It was discovered that multiple linear regression models using Wi-Fi, plug-in equipment, and lighting load data produced accurate occupant-count estimates. A novel forecasting framework consisting of k-means clustering of occupancy data, motif and discord identification, and classification trees was employed to develop a rules-based method for occupancy prediction that could be trained offline and practically applied in existing commercial building automation systems. The results from a six-month long implementation of these occupancy-based controls in the outdoor air dampers showed that heating and cooling energy use were reduced by 38% and 10%, respectively, without any need for additional sensing or controls infrastructure. These findings indicate the potential benefit of data-driven building operations when occupants are accounted for directly in the sequence of operations and identifies areas that warrant further investigation.

Preface

This integrated thesis consists of two journal papers and a conference paper, either published or under review. Should readers wish to refer to materials from this thesis, the current thesis is required to be cited. The articles included in this thesis are as follow:

- **Article 1:** Hobson, Brodie W.; Lowcay, Daniel; Gunay, H. Burak; Ashouri, Araz; and Newsham, Guy R. Occupancy-count estimation using sensor fusion: A case study. *Building and Environment*. 2019; 159: 106154–106164.
- **Article 2:** Hobson, Brodie W.; Gunay, H. Burak; Ashouri, Araz; and Newsham, Guy R. Clustering and motif identification for occupancy-centric control of an air handling unit. *Energy and Buildings* [Under review].
- **Article 3:** Hobson, Brodie W.; Gunay, H. Burak; Ashouri, Araz; and Newsham, Guy R. Occupancy-based predictive control of an outdoor air intake damper: A case study. *Proc. of IBPSA-Canada's eSim 2020 Conf.* [Accepted].

The articles have been altered slightly to help the flow of this dissertation. Reference is made to figures from other chapters to avoid redundancy.

Use of copyrighted material from the published articles is acknowledged as per the corresponding publisher's permission guidelines with respect to the authors' rights.

In the aforementioned articles, Brodie W. Hobson was the principal contributor to the research methodology, data analysis, and preparation of written material and figures presented in the articles, under the supervision of Dr. H. Burak Gunay. Daniel Lowcay contributed to data acquisition and provided critical feedback and review along with Dr. Araz Ashouri and Dr. Guy R. Newsham on the manuscripts of their respective articles.

Acknowledgements

I extend an immeasurable amount of gratitude to my supervisor Professor Burak Gunay. Your unwavering guidance, support, and dedication has extended from my undergraduate degree through to my graduate studies and has brought me a fulfillment I never dreamt I could experience. Many thanks also go to my other co-authors and advisors, namely Dr. Guy Newsham, Dr. Araz Ashouri, and Daniel Lowcay.

I would also like to extend my appreciation to the other members of my committee, namely Professor Ian Beausoleil-Morrison, Dr. Araz Ashouri, and Professor Jean Duquette, for taking their time to provide feedback on the manuscript of this dissertation.

I would like to acknowledge the members of the Building Performance Research Centre, especially Professor Liam O'Brien and my office mates, for coming together to create such a supportive and knowledgeable team that I am proud to be a part of.

Funding for this research was provided by Natural Sciences and Engineering Research Council of Canada, Green Power Labs, Natural Resources Canada, CopperTree Analytics, Sensible Building Science, Delta Controls, Bentall GreenOak, and National Research Council Canada. I would like to acknowledge the support of Scott Macdonald, Gavin Symonds, and the entire team from Facilities Management and Planning at Carleton University; Trent Hilliard at Green Power Labs, and Ricardo Moromisato at CopperTree Analytics. I would also like to acknowledge the IEA-EBC Annex 66 and 79 researchers, whose expertise and discussions have benefitted me greatly.

Finally, I would like to thank all my friends and family, especially my parents, for their unwavering love, support, and sacrifice that has afforded me this incredible opportunity. I owe you a debt of gratitude that I can only hope I will someday be able to repay.

Table of Contents

| | |
|---|------|
| Abstract..... | i |
| Preface..... | ii |
| Acknowledgements..... | iii |
| Table of Contents..... | iv |
| List of Figures..... | vii |
| List of Tables..... | viii |
| 1. Introduction..... | 1 |
| 1.1. Research objectives and questions..... | 2 |
| 1.2. Document structure..... | 4 |
| 2. Opportunistic occupant-count estimation..... | 7 |
| 2.1. Introduction..... | 7 |
| 2.2. Literature review..... | 8 |
| 2.3. Methodology..... | 14 |
| 2.3.1. Building and dataset..... | 14 |
| 2.3.2. Model formalisms..... | 18 |
| 2.4. Results..... | 21 |
| 2.4.1. Multiple linear regression models..... | 21 |
| 2.4.2. Artificial neural network models..... | 22 |
| 2.5. Discussion..... | 24 |
| 2.5.1. Model formalism comparison..... | 24 |
| 2.5.2. Sensor utility comparison..... | 26 |
| 2.5.3. Model selection criteria..... | 29 |
| 2.6. Closing remarks..... | 32 |
| 3. Rules-based occupancy prediction..... | 37 |
| 3.1. Introduction..... | 37 |
| 3.2. Literature review..... | 41 |

| | | |
|--------|---|----|
| 3.3. | Methodology | 44 |
| 3.3.1. | Building and dataset..... | 45 |
| 3.3.2. | Occupant-count estimation | 46 |
| 3.3.3. | Clustering and motif identification | 47 |
| 3.3.4. | Online forecasting..... | 50 |
| 3.4. | Results and discussion | 50 |
| 3.4.1. | Occupant-count estimation | 50 |
| 3.4.2. | Clustering and motif identification | 52 |
| 3.4.3. | Online forecasting..... | 57 |
| 3.4.4. | Proposed intervention | 60 |
| 3.5. | Closing remarks | 62 |
| 4. | Air handling unit implementation..... | 66 |
| 4.1. | Introduction..... | 66 |
| 4.2. | Methodology | 68 |
| 4.2.1. | Damper profiles | 68 |
| 4.2.2. | Occupancy forecast..... | 72 |
| 4.2.3. | Implementation | 73 |
| 4.3. | Results and discussion | 75 |
| 4.3.1. | Forecast accuracy..... | 77 |
| 4.3.2. | Energy savings | 79 |
| 4.3.3. | Comfort impact | 80 |
| 4.4. | Closing remarks | 83 |
| 5. | Conclusions..... | 86 |
| 5.1. | Summary..... | 86 |
| 5.1.1. | Opportunistic occupant-count estimation | 86 |
| 5.1.2. | Rules-based occupancy prediction..... | 87 |
| 5.1.3. | Air handling unit implementation..... | 89 |

| | |
|--|----|
| 5.2. Research contributions..... | 90 |
| 5.2.1. Occupant-count estimation | 90 |
| 5.2.2. Rules-based occupancy prediction..... | 91 |
| 5.2.3. Air handling unit implementation..... | 91 |
| 5.3. Recommendations for future work | 92 |
| Appendix A..... | 96 |
| References..... | 97 |

List of Figures

| | |
|--|----|
| Figure 2.1: Floor and sensor configuration..... | 15 |
| Figure 2.2: Sample combinations of data streams, zones, and model formalisms. | 19 |
| Figure 2.3: Range of R^2 values produced by multiple linear regression. | 22 |
| Figure 2.4: Range of R^2 values produced by artificial neural networks. | 24 |
| Figure 2.5: Ground truth and predicted occupant count over time. | 31 |
| Figure 3.1: Wi-Fi device counts and ground truth occupancy..... | 47 |
| Figure 3.2: Average weekly occupancy profiles..... | 51 |
| Figure 3.3: Caliński-Harabasz index for clustering techniques. | 53 |
| Figure 3.4: Representative occupancy profiles for each clustering technique..... | 54 |
| Figure 3.5: Classification tree for forecasting occupancy day type..... | 56 |
| Figure 3.6: Cluster membership (a) using Wi-Fi data and (b) electricity use..... | 58 |
| Figure 3.7: Difference between occupant-count estimate and occupancy assumptions... | 59 |
| Figure 3.8: Outdoor air damper position vs. outdoor air fraction. | 61 |
| Figure 4.1: Occupancy profiles developed for the case study building..... | 69 |
| Figure 4.2: Damper position and air fraction for opposed blade dampers..... | 71 |
| Figure 4.3: Damper position profiles for (a) AHU1 and (b) AHU2. | 71 |
| Figure 4.4: Plug-in equipment and lighting load profiles. | 72 |
| Figure 4.5: Forward stepwise selection of timesteps for GCL+ forecast. | 74 |
| Figure 4.6: Outdoor air damper positions from before and after implementation..... | 75 |
| Figure 4.7: (a) Wi-Fi measured occupancy and (b) electrical load-based forecasts. | 77 |
| Figure 4.8: Change point model for baseline year. | 80 |
| Figure 4.9: Change point model for implementation year. | 80 |
| Figure 4.10: Changes in CO_2 concentration. | 81 |
| Figure 4.11: Changes in indoor air temperature. | 82 |

List of Tables

| | |
|---|----|
| Table 2.1: Floor size and use breakdown..... | 15 |
| Table 2.2: Floor sensor type and count breakdown. | 16 |
| Table 2.3: Data collection period(s) for each floor (2018). | 17 |
| Table 2.4: Selected sensor combinations for occupant-count estimation. | 30 |
| Table 4.1: Area outdoor air rate for the case study building. | 70 |
| Table 4.2: Confusion matrix for occupancy forecast..... | 78 |

1. Introduction

The operation of commercial and institutional buildings requires a significant amount of resources. In Canada, this sector of the building industry accounts for approximately 15% of all secondary energy use, with nearly half of this attributable to heating, ventilation, and air conditioning (HVAC) loads [1]. These HVAC services are provided to spaces primarily for the purpose of ensuring adequate indoor air quality for occupant health and comfort. Operating procedures such as ASHRAE Guideline 36-2018 [2] set forth ventilation rates to maintain this level of service which are based primarily on the number of occupants within the space under consideration. However, decisions about sizing HVAC equipment to meet the occupancy of a space are made early on during the design process, which can result in broad and highly conservative assumptions about the expected occupancy of the building [3], or worse yet, conflicting assumptions about occupancy between design stakeholders [4, 5]. This problem is further compounded by the evolving landscape of the modern workplace as many employers moving towards flexible working hours, hotdesking, and advances in telecommuting [6], reducing the number of rigid ‘nine-to-five’ workers seen more commonly in previous decades. Despite this, many of these operating procedures still state that system-level HVAC equipment should be operated assuming full or near full occupancy during working hours, while there exists a vast body of literature which has shown that occupancy in commercial and institutional buildings is often a fraction of this assumption.

The result of these crude occupancy assumptions and operational practices is the chronic overventilation of many commercial and institutional buildings. In buildings with centralized forced-air systems, this problem manifests itself immediately at the interface

between the HVAC system and the fresh outdoor air: the outdoor air dampers of their air handling units. In the heating and cooling season, the conditioning of outdoor air to meet supply air temperature setpoints is the primary driver behind HVAC loads. During these seasons (which constitute the vast majority of the year in many Canadian climate zones), the outdoor air dampers operate in a minimum-position mode; the dampers are opened just enough to provide fresh air for the buildings *expected occupancy*, while keeping the outdoor airflow as low as possible to remain economical. This expected occupancy in current practice is the single conservative assumption about the building's occupancy made early in the design phase.

Therefore, if the occupancy of the building can be estimated accurately enough for HVAC purposes, outdoor air damper minimum positions could be reduced based on the actual occupancy to yield significant energy savings without impacting indoor air quality. However, estimation alone is not adequate for operating system level equipment such as air handlers due to the transient nature of buildings' thermal and air quality conditions; it may take several hours from equipment actuation to reach desired setpoints or contaminant levels. Instead, developing forecasts for occupancy levels within a building may allow this goal to be realized. A practical and low-cost way to implement these occupancy-based predictive controls in existing commercial building automation systems is needed if industry is ever to see the potential widespread benefit of these or similar approaches.

1.1. Research objectives and questions

Given the environmental and economic impact of operating commercial and institutional buildings indoor climate control systems and the inadequacies of current practices, occupancy-based predictive controls for air handling units' outdoor air dampers are the

subject of this research. The development and implementation of these controls is undertaken for a case study academic office building in Ottawa, Canada. The major inquiries of this integrated thesis are broken down by chapter:

- Opportunistic occupant-count estimation (chapter 2):
 - What already installed sensing infrastructure can be leveraged for estimating occupant count implicitly?
 - What combinations of these sensors can be used in sensor fusion models to enhance occupant-count estimation accuracy?
 - What model formalisms are suited for this purpose?
 - How transferable are these occupant-count estimation models from one space to another?
- Rules-based occupancy prediction (chapter 3):
 - How can the occupancy in the case study building be characterized and how does this differ from the assumptions used in operation?
 - What are the constraints of commercial building automation systems that should be considered when developing a forecasting approach?
 - What statistical and machine learning techniques can be employed to forecast building occupancy one-day in advance given these constraints?
 - How accurate are the predictions generated from this approach and what implications does this potentially have on operations?
- Air handling unit implementation (chapter 4):
 - How can rules-based occupancy predictions be integrated into outdoor air damper controllers?

- What magnitude of energy savings can be achieved using this approach?
- What implications does the implemented control scheme have on indoor environmental quality?
- What are some of the barriers or considerations that arose during the implementation and how might these be addressed?

1.2. Document structure

The remainder of this integrated thesis consists of three main body chapters on (1) opportunistic occupant-count estimation, (2) rules-based occupancy forecasting, and (3) the air handling unit implementation, followed by the conclusion. Each chapter is outlined briefly below:

Chapter 2: This chapter presents the analysis of cost-effective, opportunistic data streams from the case study building to develop occupant-count estimations for HVAC control purposes. Implicit occupancy sensing via sensor fusion is conducted using available data from Wi-Fi access points, CO₂ sensors, PIR motion detectors, and plug-in equipment and lighting load submeters, with over 200-hours of concurrent ground truth occupant counts. Multiple linear regression and artificial neural network model formalisms are employed to blend these individual data streams in an exhaustive number of combinations over four floors of the building that have been studied. Insights into the relative utility of each sensor combination and formalism for occupant-count estimation are discussed.

Chapter 3: This chapter discusses the constraints of commercial building automation systems for occupancy forecasting and proposes a methodology that can be implemented in a controller given these constraints. Seven months of Wi-Fi, plug-in equipment and lighting load data were collected from the case study building. Different clustering

techniques were explored for generating representative daily occupancy profiles. Corresponding electrical load profiles for these representative occupancy profiles were extracted. A classification tree was developed to determine motif weekly sequences of these profiles for day-ahead forecasting based on the rules extracted during this process. These rules-based forecasts' accuracy and practicality for implementation purposes are evaluated, and an implementation framework is developed.

Chapter 4: In this chapter, outdoor air damper profiles are developed for each representative occupancy profile from the previous chapter, and the rules-based forecast is translated into the building automation systems control language (i.e., GCL+). Heating and cooling energy use, as well as zone-level temperatures and CO₂ concentrations, are monitored during a 24-week period where this program is given control of the outdoor air dampers. Three-point univariate changepoint models are employed to estimate energy savings. Changes in indoor air quality are evaluated using data from the same 24-week period from the previous year. The rules-based prediction accuracy and the efficacy of this approach are examined and discussed.

Chapter 5: This chapter summarizes the findings and conclusions from each chapter. The contributions of each portion of this research are also outlined. Recommendations for future work are developed.

Chapter 2

This chapter has been published as:

Opportunistic occupancy-count estimation using sensor fusion: A case study.

Hobson, B.W.; Lowcay, D; Gunay, H.B.; Ashouri, A.; and Newsham, G.R. *Building and Environment*. 2019; 159: 106154–106164.

2. Opportunistic occupant-count estimation

2.1. Introduction

It is estimated that 15-30% of the energy used by heating, ventilation, and air conditioning (HVAC) systems in commercial and institutional buildings is wasted due to inefficiencies in their operation [7, 8]. Considering that HVAC systems in said buildings account for ~15% of the secondary energy use and ~10% of greenhouse gas (GHG) emissions in Canada [1], improvements to their operation represent a key opportunity to reduce global energy consumption and CO₂ emissions.

As stated by Shen *et al.* [9], the key to improving the operation of HVAC systems is to provide heating, cooling, and ventilation only when and where they are needed, in the amount that they are needed. Notably, this is inherently dependant upon having access to high-resolution occupancy information. By knowing the first arrival and the last departure times, one can schedule the HVAC start and stop times and/or apply temperature setbacks. By knowing the number of occupants, one can adapt the ventilation rates (i.e., demand control ventilation). Implementation of such occupancy-centric control strategies are reported to achieve energy savings of 18% [10] to 55% [11] depending on the building archetype and climatic conditions. Further, occupancy-centric HVAC controls are reported to reduce the peak energy demand by diversifying the HVAC needs of individual zones [12].

Despite the importance of high-resolution occupancy information for energy efficient HVAC operation, there is no perfect, cost-effective sensor to count the number of occupants in a building accurately and non-intrusively. However, sensor fusion is a promising statistical approach to combine imperfect information from disparate data

sources [13]. Rather than deploying sensors dedicated for occupancy detection, sensor fusion can take advantage of pre-existing data streams (e.g., Wi-Fi access points, CO₂ sensors, motion detectors, electrical load meters) that are commonly available in existing buildings and expected to be correlated with occupancy, allowing for cost-effective occupant-count estimations. This opportunistic approach is referred to as implicit occupancy sensing [14].

This chapter focuses on using implicit occupancy sensing via sensor fusion models to develop occupant-count estimations for HVAC control purposes. By using the Wi-Fi device count, CO₂ concentration, passive infrared (PIR) motion detector trigger, plug load, and lighting load data from an academic office building, multiple linear regression and artificial neural network (ANN) models are developed to estimate occupant counts. These estimates are compared to ground truth data. The effectiveness of different sensors, sensor combinations, and the model formalisms for the purpose of occupant-count estimation is discussed. The transferability of models developed in one floor and used in another is explored in the case study building throughout. Criteria are put forward for selecting the most suitable model for occupant-count estimation, and these models are identified for the case study building given the data examined.

2.2. Literature review

Traditional occupancy sensing technologies have numerous shortcomings. The majority of occupancy sensors are motion detectors (e.g., PIRs or ultrasonic motion detectors); thus, they can only detect presence, not the number of occupants in a space. These sensors are also unable to distinguish between a lack of movement and an absence of occupants, which can lead to false vacancy detections if the time-out period is too short [15–18]. Increasing

the time-out period of 5 to 60 minutes after the last detected movement could reduce the frequency of false vacancy detections [19]. However, increasing the length of this time-out period also increases the possibility of false occupancy detections; this trade-off reduces the savings potential. Additionally, if the sensor detects a motion and activates building services such as lighting, the occupants may deem this undesirable and resort to methods of disabling the motion detector, such as intentionally blocking the sensor to obstruct its line-of-sight [20, 21]. This will either diminish or completely negate the sensors' functional purpose. Furthermore, because motion detectors require a direct line-of-sight to detect occupants, they need to be deployed in a fine grid to ensure their field-of-view is not obstructed by architectural partitions or furniture; this dense configuration incurs additional cost both for the sensors and labour. To address the binary nature of PIR sensors, Wahl *et al.* [16] proposed a system whereby PIR sensors were installed on either side of a door at the lintel level. This configuration produced an analogue output that could be used to determine if an occupant was entering or leaving the studied space. The system kept track of the net occupancy within the space of interest. However, this system suffered from inaccuracies when multiple occupants move through the door simultaneously, if two or more occupants moved in opposite directions, or if individuals or crowds gathered on either side of the door within the sensors' range. Liu *et al.* [17] also used PIR sensor data combined with a hidden Markov model to predict room level occupancy with accuracies ranging from 81% to 97%. However, this required a dense sensor grid and computationally expensive computer algorithms that are only applicable to the studied space. Similarly, Duarte *et al.* [18] proposed a network of PIR sensors to establish occupancy patterns in an office building testbed. While this system was able to establish high-level occupancy

patterns, the resolution was not high enough to develop zone-level HVAC control strategies and required a dense sensor grid. The first cost of installing such a granular PIR system has deterred implementation of these systems in spaces outside those required by codes – such as the National Energy Code for Buildings in Canada and ASHRAE 90.1-2019 [22]. This shows that the method by which motion detectors are implemented currently is not appropriate for occupant-count estimation for HVAC control purposes. As such, some researchers have turned to the use of CO₂ sensors for inferring occupancy levels [11, 23]. However, problems have been noted with calibrating the models to account for the latency between occupants' arrival/departure and the corresponding increase/decrease in CO₂ concentration, as well as the sensitivity of this technology to sensor position and reading accuracy.

As such, significant research over the past decade has focused on occupancy-detection studies using new sensor technologies explicitly built for occupant counting. Many of these technologies involve optical [10, 24–29] and/or infrared [10, 26, 27] camera-based technologies coupled with complex image recognition algorithms. However, many of these studies are limited to a single room office-type occupancy [10, 24, 26] with a limited number of occupants to monitor. Even with low occupancy, Cao *et al.* [10] specifically note that significant computational resources are required for this type of image processing approach. The first cost attached to these sensors and the related infrastructure, as well as the labour required to integrate them into an existing building automation system (BAS) cannot be neglected. Additionally, the algorithms used to extract the occupant count from the camera-based sensors can suffer from inaccuracies when numerous occupants enter the frame, or if occupants are moving too quickly to be captured by the camera's framerate,

which is typically reduced to a frame every couple of seconds to reduce the computational resources needed for real-time processing [10, 24–29]. Like motion detectors, this system cannot detect occupants that are outside the cameras line-of-sight, necessitating numerous well-placed sensors if adequate occupant counting is to be expected. Cameras also inherently raise questions about privacy. Although these studies make use of methods to pixilate and analyze images without compromising occupant identity, it has been shown that even a perceived threat to privacy associated with monitoring technologies such as cameras can deter the implementation of said systems [21]. Due to these many challenges and shortcomings, Petersen *et al.* [27] acknowledge that the applications of these technologies should be limited to gathering ground truth data until more robust and privacy conscious occupant counting methods are developed.

Another technology used in occupancy-detection studies is radio-frequency identification (RFID) tags. Carreira *et al.* [30] used RFID tags and a web-based application to track occupants and their vote to raise or lower the room temperature by 1°C. Li *et al.* [31] also used RFID tags to gather information about occupants' location and identity. These studies showed that RFID devices do not suffer from the same line-of-sight pitfalls as motion and camera-based sensors, however, their direct tracking of individual occupants' movements only exacerbate security and privacy concerns. The cost of implementing these sensors and their integration into the BAS are similarly a barrier to their deployment and implementation when compared to camera-based sensors, although Li *et al.* [31] note that sensor cost has decreased substantially over the past several years. Other proposed proprietary sensing methods include using electrical load data to characterize occupancy. Akbar *et al.* [32] applied k-nearest neighbor (KNN) and support vector machine (SVM)

machine learning classification algorithms on data from a ZigBee smart energy unit, which submeters electrical loads down to the appliance used at the desk level. It was found that this method could detect whether an occupant was present or absent from their desk with 94% accuracy. Labeodan *et al.* [33] deployed an office chair with various sensing technologies that allowed it to detect when the occupant was seated. However, in both these cases, the scalability of this method is questionable as every workstation would require an additional sensing unit; this method would also only work in cases where every occupant has a desk or workstation with plug-in equipment. Therefore, these kinds of systems may not be practical for many occupancy types where non-habitual users or the broader public represent a large fraction of the buildings occupants.

To address the weaknesses of any individual sensor technology, numerous studies have been conducted that take advantage of sensor fusion. Sensor fusion is a technique that combines data collected from multiple sensors and sensor technologies to provide enhanced performance over any single sensor. These studies have looked at different combinations of sensors and data, ranging from indoor air temperature [26, 34–40], indoor relative humidity [24, 35, 37–42], CO₂ levels [24, 34–41, 43], volatile organic compound (VOC) concentrations, [34, 37, 38, 42], PIR trigger events [34, 36–39, 41, 42, 44], lighting loads [34, 36–39, 41, 43], noise levels [36–39, 42, 45], Wi-Fi device counts [36, 40, 46–50], and other sensors such as door triggers [36, 41, 44] and Bluetooth device counts [46, 47, 51]. This data is then analyzed using different machine learning algorithms to generate occupant-count estimations with a high enough resolution to attain HVAC energy savings anywhere from 18% [10] to 55% [11]. However, many buildings do not monitor parameters such as VOC concentration or noise levels. Door entry and exit activity is not monitored

in many buildings with public or semi-public access. Bluetooth has also been shown to be less prevalent than Wi-Fi enabled devices [52] and therefore less useful for occupant-count estimation [53]. As Bluetooth and Wi-Fi depend on signal propagation, both suffer from potential dead-zones or poor coverage in buildings with large amounts of internal mass. Privacy concerns with the use of these signals remains a tangible barrier to their widespread use. These studies were also performed either in an experimental testbed or using ad hoc sensor arrays in a pre-existing built environment; this additional hardware drives up the cost of detecting occupants within the building. It would be advantageous to explore leveraging existing data streams within buildings using in-situ sensors further, as these opportunistic data streams would reduce the cost needed to facilitate occupant-count estimation. Additionally, the models in these studies were developed for a single space over the measured occupancy period, without consideration for how these models will perform if implemented in a different context. Overall, these studies showed that a limitation of sensor fusion is that models can differ vastly in terms of their predictive accuracy due to a variety of factors [21]. The complexity of the model formalism chosen can substantially affect the model accuracy. Similarly, the data streams, sensors, and sensor combinations available for analysis will have different levels of utility for occupant-count estimation purposes. Furthermore, the calibration of a model with space-specific ground truth data will impact its ability to determine occupancy levels. As the ground truth data collection process is invasive and resource/time intensive, it would be advantageous to explore the transferability of models to another context. This might include if a model trained and validated in one space at a certain time can be applied in an entirely different temporospatial context.

This study uses in-situ commercial sensors already installed in the case study building and ground truth data, without any other ad hoc data recorded. A dataset with over 200 hours' worth of manual ground truth occupant counts is used to assess the potential of sensor fusion-based implicit occupancy sensing. Models are developed for all possible sensor-type combinations. These models are implemented on different floors of the building to explore their transferability, which – to the authors' knowledge – has not been the subject of any significant study to date.

2.3. Methodology

2.3.1. Building and dataset

The analysis was carried out on four floors of an academic office building in Ottawa, Canada. Each floor varies slightly in configuration and size, see Figure 2.1 and Table 2.1, respectively. Note that the PIR and CO₂ sensors are commercial grade wall units mounted at approximately 1.6 meters in height, with the exception of the ceiling mounted PIR units in the washrooms of floor A and C.



Figure 2.1: Floor and sensor configuration for (a) floor A, (b) floor B, (c) floor C, and (d) floor D.

Table 2.1: Floor size and use breakdown.

| Occupancy Type | Approximate Area (m ²) | | | | |
|----------------------|------------------------------------|-------------|------------|-------------|-------------|
| | Floor A | Floor B | Floor C | Floor D | Σ |
| Private/Semi-private | 176 | 160 | 129 | 250 | 715 (17%) |
| Laboratory/Classroom | 427 | 618 | 377 | 389 | 1811 (43%) |
| Utility | 98 | 80 | 144 | 96 | 417 (10%) |
| Communal | 290 | 281 | 294 | 417 | 1283 (30%) |
| Total | 991 | 1139 | 944 | 1152 | 4226 |

All floors contain a mixture of private and semi-private office spaces, as well as classrooms, study spaces, washrooms, and space for utilities/IT services. These office spaces and some laboratories are used by professors and graduate students during the entire year, whereas the classrooms are typically only used during the academic year (i.e., from September to mid-April). The sensor grid density and granularity also vary between floors; sensor counts are shown in Table 2.2. On floor B, the sensors are disproportionately concentrated in the office spaces on the west side of the floor. The sensor counts are nearly double that of floor B on floors A and C, and the sensors on these floors are distributed over a larger proportion of the floor area.

Table 2.2: Floor sensor type and count breakdown.

| Sensor Type | Number of Sensors | | | |
|-----------------|-------------------|---------|---------|---------|
| | Floor A | Floor B | Floor C | Floor D |
| Wi-Fi | 3 | 5 | 2 | 4 |
| CO ₂ | 26 | 13 | 20 | 16 |
| PIR | 26 | 10 | 21 | 19 |
| Plug Meter | 1 | 1 | 1 | 1 |
| Light Meter | 1 | 1 | 1 | 1 |

The data from these sensors are accessible through the BAS, including individual CO₂ sensor readings, individual PIR activations, and instantaneous floor-level lighting and plug loads, all provided in a concurrent five-minute timestep. Individual Wi-Fi-enabled device counts are also provided in the same five-minute timestep from the IT network for each Wi-Fi access point. These counts include all mobile devices and ambient devices such as wireless printers and laptops connected to or attempting to connect to the Wi-Fi access

points, referred to as associated devices. Note that these data were anonymized before being provided, and therefore no MAC addresses were monitored or stored during the data collection campaign. The CO₂ sensors’ readings were verified using a reference CO₂ logger and were found to have a mean absolute error of 96 ppm. The functionality of each PIR sensor was also individually verified. Coverage maps were provided by IT services to verify that the Wi-Fi access points did not have any dead-zones on each floor. Models were developed using these data as-is. The data were selected for the concurrent times that ground truth was available.

Table 2.3: Data collection period(s) for each floor (2018).

| Floor | Date(s) & Time(s) | Peak Occupancy |
|--------------|--|-----------------------|
| A | Fri, 2-Mar 7:56-12:56 | 26 |
| | Sat, 3-Mar, 7:31-12:56 | |
| | Wed, 15-Aug, 17:25 to Fri, 17 Aug, 15:15 | 38 |
| B | Tue, 13-Feb, 7:18-11:08 | 72 |
| | Fri, 16-Feb, 7:33-11:13 | |
| | Mon, 26-Feb, 7:28-12:38 | |
| | Mon, 30-Jul, 16:00 to Wed, 1-Aug, 14:00 | 37 |
| C | Tue, 15-May, 15:31 to Thu, 17-May, 14:31 | 21 |
| D | Mon, 18-Jun, 14:05 to Wed, 20-Jun, 11:40 | 67 |

In order to verify and compare the models developed using the sensor data, ground truth occupant counts were collected for each floor. Three cameras were used, with one at the elevator lobby and one at each stairwell to ensure all entrance and exit points were monitored. This footage was then manually reviewed by a researcher and the timing of the entry and exit events were manually logged. Note that for privacy reasons, the cameras

used did not record any audio and the footage was deleted immediately after being logged. The methodology was reviewed and approved by the institution's research ethics board. Data collection was performed for floors A, B, C, and D between the periods shown in Table 2.3. The data from these observation periods was then grouped to match the same five-minute timestep as the BAS data. Overall, 208.5 hours of ground truth observations were recorded.

2.3.2. Model formalisms

The data collected was reduced to a simplified dataset with the following six common datatypes for each floor: 1) the number of associated Wi-Fi devices, 2) the ensemble mean of the CO₂ concentration (ppm), 3) the fraction of PIR sensors triggered, 4) the plug load (kW), 5) the lighting load (kW), and 6) the aggregated electrical load (i.e., the sum of the plug and lighting loads). This final datatype was added to explore the value of having the lighting and plug loads on one submeter for the purpose of occupant-count estimation versus having these load types aggregated. Note that in the cases where either the plug or lighting load was considered as a predictor of occupancy, the electrical load was not considered and vice versa, as this would be redundant. To further facilitate comparison, the refined dataset underwent feature scaling in the form of a min-max normalization before analysis. The focus was on the subset of data collected between 6 am and 10 pm, as this is when occupants were generally observed on all floors. Including data from the overnight period in model development could artificially inflate the predictive accuracy of the model; a model that can predict that a building is unoccupied overnight (a relatively trivial exercise) should not be conflated with a model that can accurately predict occupancy throughout the day. Such a model is not necessarily useful in commercial and institutional

buildings where overnight occupancy is minimal. A similar approach is adapted by Wang *et al.* [40] in their model development.

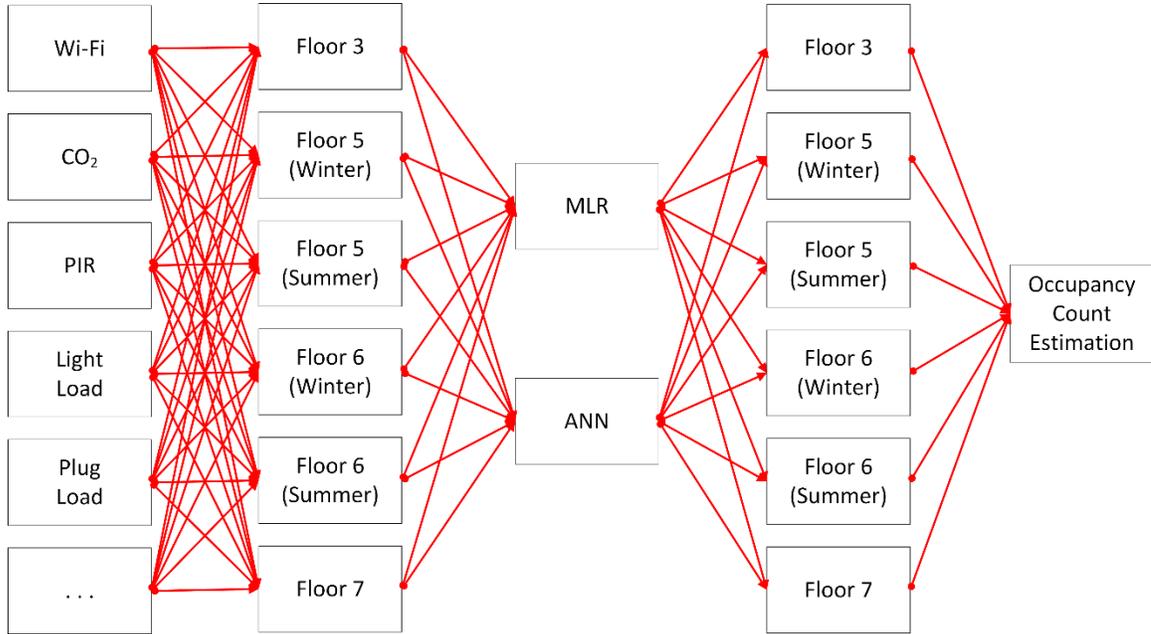


Figure 2.2: Sample combinations of data streams, zones, and model formalisms.

Using enumerative combinatorics, all 39 possible unique combinations of the six datatypes were identified. The effectiveness of each combination for occupant count-estimation is explored. Note that for floors A and B, separate models were developed for each of the winter and summer semester to explore how different occupancy levels in the same space would affect the quality of the resulting model (e.g., will the trends observed in the summer models hold true in the higher occupancy winter models, and vice versa?). Each of the 39 datatype combinations were then modelled using multiple linear regression and ANN formalisms for each floor, with additional models for floors A and B as mentioned. The data were randomly split into 70% and 30% for training and validation, respectively, and 10-fold cross-validation was performed. The models were then fitted to the datatypes for the floor they were generated from, as well as all other floors. This allows for the predictive

accuracy of a model developed in one space to be evaluated when applied to another space. The resulting fits were then compared to the ground truth data. This means that 39 datatype combinations were considered over six different floor scenarios, with 234 unique models developed for each model formalism, referred to as the ‘intra-floor’ fits. Then, each of the models developed for a single floor were tested on all other floors, for a total of 1170 fits for each model formalism, referred to as the ‘inter-floor’ fits. In total, there were 1404 fits for each model formalism, for a total of 2808 fits developed over the course of this study, see Figure 2.2.

The ANN models are two-layer feed forward neural networks developed and applied using R [54], namely the *neuralnet* package [55]. The single hidden layer uses sigmoid activation functions and contains a varying number of neurons depending on the number of sensors considered in the model. This scales a neural network’s complexity to match the amount of data considered. This approach is adapted from Heaton [56] and involves sizing the hidden layer to be two-thirds the size of the input layer (i.e., the number of sensor-types considered) plus the size of the output layer (i.e., one or the occupant-count estimation). The output layer uses a logistic function. The parameters of the ANN are then determined using the resilient backpropagation method with weighted backtracking. Training of the network was terminated when the partial derivatives of the error function reached 10^{-6} or 1000 iterations were performed without convergence. The multiple linear regression models were developed and fitted using the R’s *base* package [57], with the intercept coerced through the origin to account for the data normalization (i.e., minimum observed values should correspond to minimum observed occupancy). The coefficient of determination (R^2) and normalized root mean square error (NRMSE) were computed for

the fits produced by these models to assess their predictive accuracy relative to the ground truth data and to facilitate comparison between each other.

2.4. Results

2.4.1. Multiple linear regression models

Figure 2.3 presents the distribution of the R^2 for the 1404 fits developed from the multiple linear regression models (i.e., 36 for each of the 39 sensor combinations). The R^2 values reported in this figure are those of both the intra-floor and inter-floor fits resulting from the multiple linear regression models. Generally, it is observed that the models developed from sensor combinations that contain floor-level associated Wi-Fi device counts have a considerably higher R^2 and lower NRMSE than those without. The maximum and mean R^2 for all Wi-Fi only fits were 96.1 and 74.1%, respectively. The mean R^2 is 80.1% if only the intra-floor fits are considered. These fits have the highest mean R^2 and lowest mean NRMSE when only a single sensor-type is considered in all six floor scenarios studied. The highest observed R^2 is 96.5% with a NRMSE of 5.9%, corresponding to the sensor fusion of the Wi-Fi, CO₂, PIR, plug and lighting load data from the intra-floor fit for floor B (i.e., the model from floor B used to predict floor Bs occupant count). The highest observed R^2 and the corresponding NRMSE resulting from a model used from one floor to another (i.e., inter-floor model) was 94.7% and 7.5%, respectively, for the sensor fusion of Wi-Fi, PIR, and plug load data from the inter-floor fit between floor D for floor B in the winter semester. The lowest observed R^2 is 11.7% with a NRMSE of 26.4%, corresponding to the sensor fusion of the PIR and plug load data from the intra-floor fit for floor C (i.e., the model from floor C used to predict floor Cs occupant count). This range of accuracies

suggest that multiple linear regression models are suited to producing occupant-count estimations using sensor fusion, depending on the sensors available.

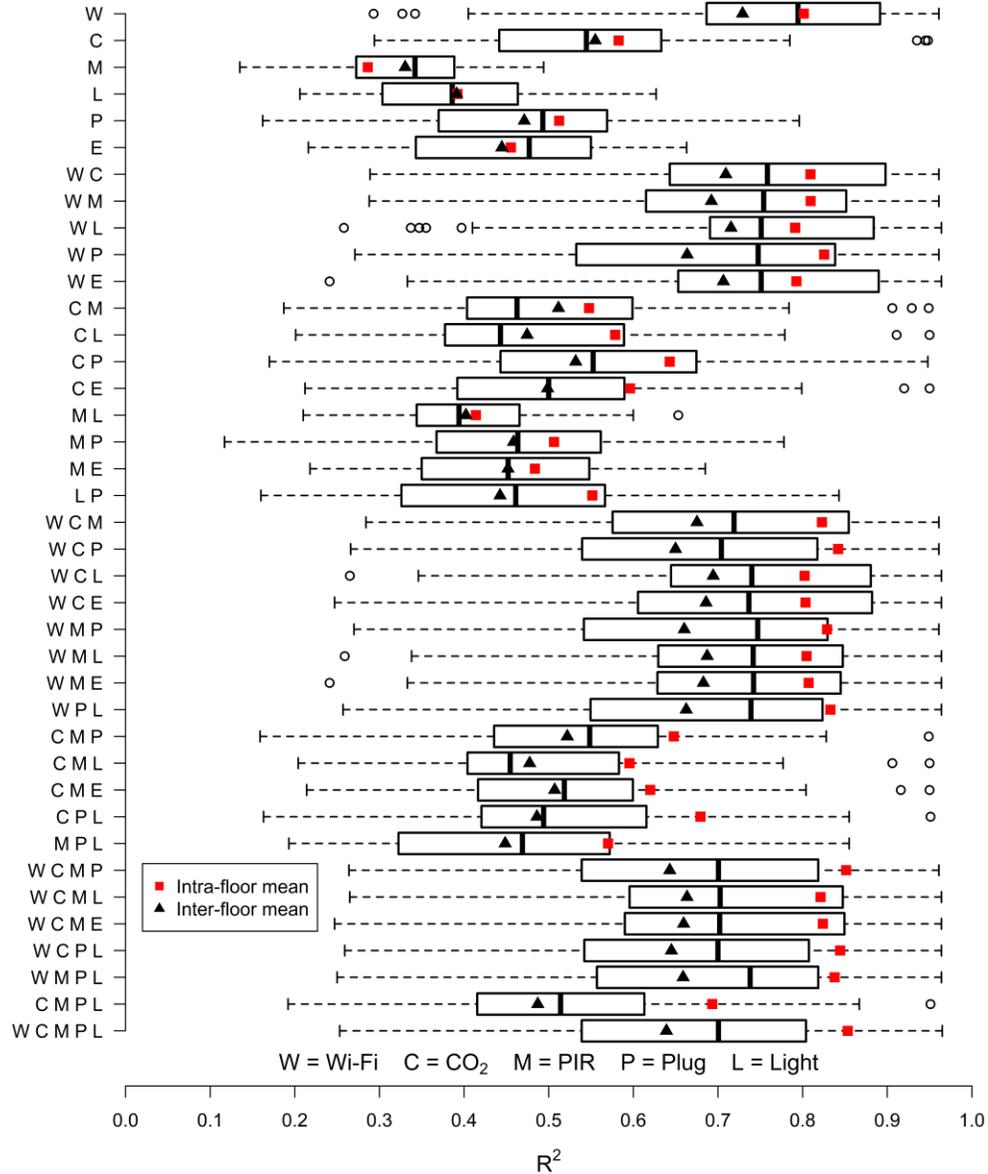


Figure 2.3: Range of R^2 values produced by multiple linear regression.

2.4.2. Artificial neural network models

Figure 2.4 shows the distribution of the R^2 of the 1404 fits for the ANN models. The coefficients of determination reported in these figures are those of both the intra-floor and inter-floor fits resulting from the ANN models. It remains the case that the models which

contain associated Wi-Fi device counts have a higher mean R^2 and lower NRMSE than those without. The maximum and mean R^2 for Wi-Fi only fits were 97.4 and 71.4%, respectively. The mean R^2 is 83.0% if only intra-floor fits are considered. These fits have the highest mean R^2 and lowest mean NRMSE when only a single sensor-type is considered in all six floor scenarios studied. The highest observed R^2 is 97.5% with a NRMSE of 5.0%, corresponding to sensor fusion of the Wi-Fi, PIR, and lighting load data from the intra-floor fit for floor B during the winter semester (i.e., the model from floor B used to predict floor Bs occupant count). The highest observed R^2 and corresponding NRMSE resulting from a model used from one floor (i.e., inter-floor fit) to another was 94.5% and 8.2%, respectively, for the sensor fusion of Wi-Fi and plug load data from the inter-floor fit between floor D for floor B in the winter semester. The lowest observed R^2 is less than 0.1% with a NRMSE of 31.8%, corresponding to the PIR data from the intra-floor fit for floor B (i.e., the model from floor B used to predict floor Bs occupant count). This range of accuracies suggest that ANN models are similarly suited to producing occupant-count estimations using sensor fusion, depending on the sensors available.

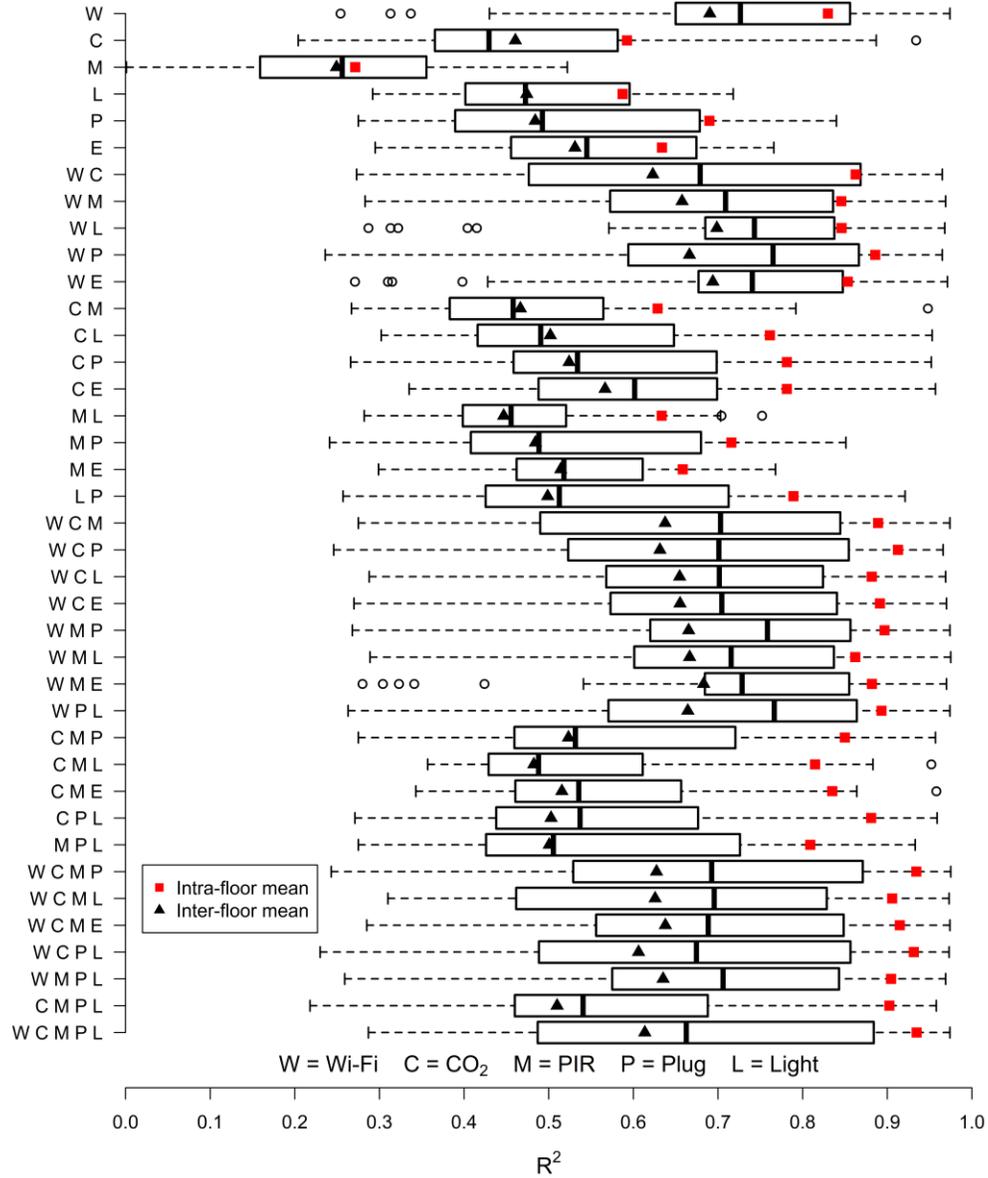


Figure 2.4: Range of R^2 values produced by artificial neural networks.

2.5. Discussion

2.5.1. Model formalism comparison

By visual inspection, the trends in the coefficient of determination for both model formalisms bear a resemblance to one another when the same floor is considered. Of the 1404 fits developed for each formalism, ANNs had a higher R^2 on a fit-by-fit basis for 56% of cases. The mean difference in R^2 and NRMSE between the fits using ANNs and the fits

using multiple linear regression is 1.7% and 0.8%, respectively. This suggests that ANN models have comparable accuracy to their multiple linear regression counterparts in the case study building. However, if the intra-floor fits are examined in isolation (i.e., the models developed and used on the same floor), there is a shift in favor towards ANN models for both the coefficient of determination and normalized root mean square error. In this case, ANN models had a higher R^2 and lower NRMSE on a fit-by-fit basis for 93% of cases, with a mean change relative to their multiple linear regression counterparts of 11.8% and -2.7%, respectively. The opposite is true when the inter-floor fits are examined (i.e., the models developed on one floor and used on another). In this case, multiple linear regression models had a higher R^2 and lower NRMSE on a fit-by-fit basis for 50.9% of cases, with a mean change relative to their ANN counterparts of 0.4% and -1.6%, respectively. This suggests the following regarding the use of either model formalism to produce sensor fusion models for occupant-count estimation:

- ANN models appear to offer higher accuracy over multiple linear regression models in 93% of the cases where the model is fitted to its own floor, making them suited to situations where ground truth data is available for calibration in the case study building; and
- Multiple linear regression models offer comparable accuracy to their ANN counterparts when models are fitted to a different floor, making them a candidate for transferability between floors and therefore suited to situations where ground truth data is not available for calibration for that floor in the case study building.

Therefore, for this building and the suggested application of these models, multiple linear regression is the superior model formalism. This is due to their comparable accuracy to

ANNs with reduced complexity, and the ease of implementation associated with the limited need for expensive, time consuming, and invasive ground truth data collection.

2.5.2. Sensor utility comparison

The intra-floor fits (i.e., the models developed and used on the same floor) generally benefit from including as many sensor datatypes as possible in their model. However, it is evident in both Figure 2.3 and Figure 2.4 that increasing the number of sensor typologies considered in the model formalism does not always equate to a higher R^2 ; a similar trend is observed in the NRMSE. Therefore, additional sensor data are advantageous if ground truth data are available, whereas less sensor data (i.e., one to two sensor datatypes) typically produce inter-floor fits with a higher mean R^2 value. The utility of sensor data produced by sensors for occupant-count estimation is linked to the sensors' configuration and/or density. Note that the magnitude of this sensitivity depends on the sensor-type considered. If this configuration and/or density varies significantly between spaces, then models developed with these sensors' data between spaces will suffer from inaccuracies. Including numerous sensor-types in models developed between spaces may therefore result in a compounding of these inaccuracies, which is why fewer sensor datatypes typically result in fits with a higher mean R^2 and lower NRMSE between floors.

By visual inspection of Figure 2.3 and Figure 2.4, it is apparent that associated Wi-Fi device counts are beneficial for floor-level occupant-count estimation. Recall that sensor fusion models that contain associated Wi-Fi device counts have a higher mean R^2 and lower mean NRMSE than those without. The fraction of PIR sensors triggered alone offers the lowest potential for occupant-count estimation. However, the range of R^2 values produced by the model which uses this datatype alone is narrowest in four out of six cases for both multiple

linear regression and ANN formalisms. This suggests that the fraction of PIR sensors triggered is not strongly linked to the sensors' configuration as models developed on one floor performed similarly on others. The utility of these PIR data increased when included in sensor fusion models with other datatypes, especially when ANN models are used. This may be attributable to the fact that the fraction of PIR sensors triggered has a non-linear relationship with occupant count (i.e., a mean R^2 of 32.3% using linear regression), whereas the non-linear functions used in computing ANN models are able to account for other, non-linear relationships these data have with occupant count. The mean CO_2 level also increased in utility when included in sensor fusion models and had the second highest mean R^2 of the stand-alone sensor datatypes for predicting occupant count. However, the range of values predicted from models using only the mean CO_2 readings across different floors was large, with a 65% and 70% difference between maximum and minimum values for multiple linear regression and ANN formalisms, respectively. This suggests that the quality of the resulting fit is highly sensitive to sensor configuration and/or density, and therefore mean CO_2 data do not lend themselves well to transferability from one space to another.

Floor-level plug, lighting, and aggregated electrical load also exhibited varying levels of utility in sensor fusion models. In general, floor-level lighting load data were more indicative of occupancy than floor-level plug load data. This may be because light levels fluctuated by a factor of more than four on average, whereas plug loads exhibit a fluctuation by a factor of less than two on average. Floors B and D has manual on and automatic off lighting controls, whereas floor A and C has automatic on and off controls. Interestingly, this did not necessarily equate to a stronger correlation between lighting load levels and

occupancy for floor A and C that might be expected from an automatic on/off control scheme. With regards to plug load data, it is suspected that submetering of these loads down to the room or workstation level would likely increase their individual contribution to the sensor fusion models, as was noted by Akbar *et al.* [32]. Models with aggregated electrical load data as opposed to separately submetered plug and lighting load data had a mean R^2 that was 0.7% and 2.6% higher for multiple linear regression and ANN model formalisms, respectively. The range of R^2 values was also 24.6% and 19.3% smaller for multiple linear regression and ANN model formalisms on average, respectively. This suggests that aggregated electrical data has higher utility for occupant-count estimation purposes. Ryu *et al.* [43] also found that electrical load data had higher information gain for occupancy-detection purposes than lighting or plug loads metered separately.

As the sensor configuration on each individual floor did not change between the summer and winter semester, it was hypothesized that the model from one semester should transfer well to the other semester. However, the fits developed from the summer semester models generally have lower mean R^2 values when compared to their winter counterparts. Floor A had a mean reduction in R^2 of 13.4% and 14.0% for multiple linear regression and ANN models between summer and winter, respectively. Recall from Figure 2.1 that the private offices on floor A are concentrated to one side, with laboratory and classroom spaces on the other. Sensors are distributed fairly evenly between these spaces. However, occupancy patterns in the private offices exhibit far less fluctuation between winter and summer compared to laboratories and classrooms. Therefore, because the sensor data have been combined into parameters – such as fraction of PIR sensors triggered across the entire floor, or the whole floors mean CO_2 concentration – the lack of activity on one half of the floor

biases these models, producing fits with a lower mean R^2 . This reduction may therefore be attributable to a change in occupancy patterns between semesters. Floor B had a mean reduction in R^2 of 8.9% and 7.5% for multiple linear regression and ANN models between summer and winter, respectively. One possible explanation is that the large computer lab on floor B was used over the summer to host a camp program for youth. It is likely that these youth did not have any Wi-Fi enabled devices, thus reducing the utility of the fits that contain Wi-Fi enabled device counts as a parameter. Therefore, if only the fits containing Wi-Fi enabled device counts are examined, there is a 10.4% and 13.9% reduction in the mean R^2 value for multiple linear regression and ANNs, respectively. If the fits developed from models that do not contain Wi-Fi enabled device counts are considered, there is only a 3.7% and 0.8% reduction in the mean R^2 value for multiple linear regression and ANNs, respectively. This suggests that Wi-Fi enabled device counts are useful for occupancy sensing only in a context where Wi-Fi devices are prevalent. Furthermore, the layout of the building or floor should also be considered in the cases where occupancy may be concentrated in a particular location during a particular time. This consideration is especially important if ground truth collection overlaps with an atypical day, as was the case with floor B during the summer semester.

2.5.3. Model selection criteria

The following criteria were developed for selecting the ‘best’ sensor combination for occupant-count estimation using sensor fusion in the building studied:

- 1) High lower quartile (Q1) R^2 values (see Figure 2.3 and Figure 2.4) – to ensure that the models are transferable and can be used in other spaces without recalibration to new ground truth data.

- 2) High R^2 and low NRMSE of the model fit – to ensure that the models are accurate.
- 3) Sensor configuration available and cost.

Table 2.4: Selected sensor combinations for occupant-count estimation using linear regression.

| Ranking (out of 39) | Sensor(s) | R^2 Q1 (%) | Mean R^2 (%) |
|------------------------|-------------------------|--------------|----------------|
| 1 | Wi-Fi + Light | 69.1 | 72.8 |
| 2 | Wi-Fi | 68.7 | 74.1 |
| 3 | Wi-Fi + Electrical | 67.1 | 72.1 |
| 4 | Wi-Fi + CO ₂ | 65.3 | 72.6 |
| 35 | Electrical | 34.5 | 44.6 |

Using criterion 1 and 2, the top three models are identified in Table 2.4. In addition, Wi-Fi alone is examined for its novelty and aggregated electrical loads are examined for their widespread availability and ease of access. As it has been established that multiple linear regression is the preferred model formalism for our purposes, the R^2 of these models are reported. Illustrative examples of the fits developed from associated Wi-Fi device count and electrical load data are provided Figure 2.5. Note that in Table 2.4, associated Wi-Fi device count alone ranks higher than the Wi-Fi and electrical as well as the Wi-Fi and CO₂ combinations. This further reinforces the findings that associated Wi-Fi device counts are beneficial for occupant-count estimation, and that a multiple linear regression formalism can be sufficient to develop a useful model for our purposes. Fits containing only aggregated electrical load data are observed to have a lower R^2 and higher NRMSE than the models with associated Wi-Fi device counts; regardless, in the absence of other sensors data, this data type can still be used as a regressor for occupant-count estimation.

The third criterion noted but not included in this table is sensor configuration availability and cost. This criterion represents an area where significant differences may be observed between buildings, or even floors of the same building as in this case study. For example, while Dong *et al.* [21] note that Wi-Fi is considered an ‘expensive’ sensor for occupant-count estimation, there is virtually no additional hardware cost if these sensors are already installed in the building for other purposes (i.e., access points for internet connectivity). This is the crux of implicit occupancy sensing. Metering of electrical loads is already common practice at the building level; however, this chapter examines submetering at the floor level. This raises the question: if installing submeters at this kind of resolution is only a moderate cost [21], can this cost be justified given the added benefit it may have for occupant-count estimation? As this chapter examines every sensor combination possible given the sensor-types examined, facility managers in the case study building could use these findings to determine the expected level of accuracy they can achieve for occupant sensing given the sensors or data available to them.

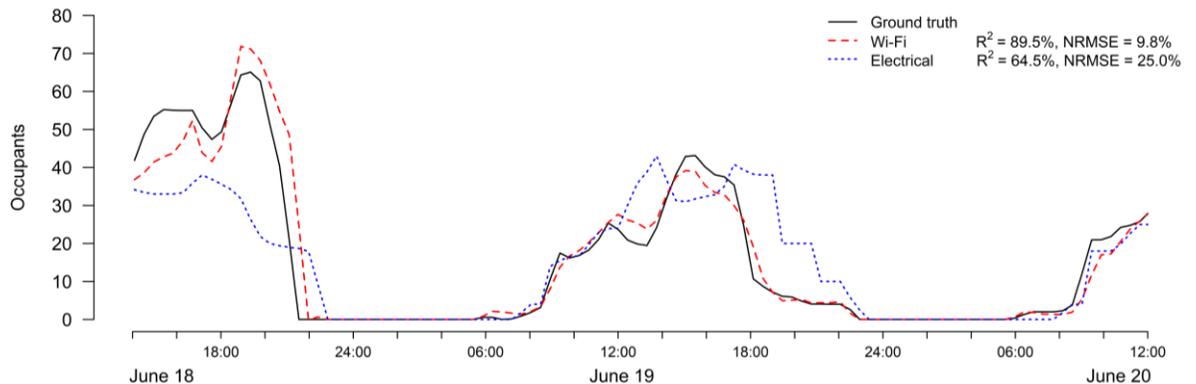


Figure 2.5: Illustrative example of ground truth and predicted occupant count over time, for Wi-Fi enabled device count and electrical load data for floor D using linear regression.

2.6. Closing remarks

In this chapter, sensor fusion of opportunistic data streams from Wi-Fi access points, CO₂ sensors, PIR sensors, and plug and lighting load meters for occupant-count estimation were explored on four floors of a case study academic office building in Ottawa, Canada. Multiple linear regression and artificial neural network model formalisms were trained and tested on 208.5 hours of ground truth data. Models from each floor were tested on the other floors to explore model transferability and sensitivity in the temporospatial context.

Key findings from this chapter are summarized as follows:

- Wi-Fi enabled device count data have the strongest relationship with ground truth occupant counts, and inclusion of these data in sensor fusion models has the single greatest impact on model accuracy.
- PIR data alone have the weakest relationship with ground truth occupant counts. However, these data have utility in sensor fusion (i.e., the inclusion of these data in models with other data sources was generally observed to improve accuracy).
- Multiple linear regression models were the preferred model formalism as they were found to have comparable accuracy to artificial neural networks for the proposed purpose, while having higher transferability and less complexity.
- Multiple linear regression models that include Wi-Fi enabled device count data can produce accurate occupant-count estimates (mean R^2 of 74.1% up to R^2 of 96.1%).
- Aggregating floor-level plug-in equipment and lighting load data to a single electrical load increased the utility of these two data in sensor fusion models.

While this chapter has shown that implicit occupancy sensing can be accomplished using sensor fusion, there are several research questions and issues to be resolved, namely:

- *How can real-time access to Wi-Fi counts be facilitated?* While Wi-Fi shows great promise for occupant-count estimation purposes, there is often a gap between facility managers and IT services when it comes to integrating this information into the BAS. As modern buildings have their BAS and Wi-Fi networks on the same local area network (LAN), increased communication between these parties may facilitate future use of Wi-Fi access points as implicit occupancy sensors for HVAC control purposes. However, addressing this problem will still not benefit buildings where occupants do not have Wi-Fi enabled devices. Privacy concerns remain an important issue. There are also considerations that need to be made for the expected prevalence of Wi-Fi enabled personal mobile devices across different cultures, socio-economic backgrounds, and over time. These differences are expected to have a great impact on the usefulness of Wi-Fi for occupant-count estimation depending on these contextual factors.
- *How much ground truth data is needed to develop an adequate model?* As ground truth collection is a labourious, expensive, and invasive, it would be beneficial to establish the threshold of data needed to develop a sufficiently accurate model. This may prove to enhance the quality of models or reveal the point at which the extra data start to have diminishing returns.
- *To what degree do occupants need to be counted to realize HVAC savings?* The ventilation in buildings is typically controlled by dampers with a discrete number of positions. As such, an absolute occupant count may not be necessary to realize HVAC savings. For example, it may be sufficient to group the occupant count into discrete states such as high, medium, or low. A worthy research question would be

how accurate these models must be to consistently identify the correct occupancy state, and the relationship between the number of states, model complexity, and the resulting HVAC savings.

- *How does sensor grid density and data timestep affect model quality?* While the accuracy of models developed in one space and applied to another was examined for the case study building, it is impossible to quantify the proportion of the change in the models' accuracies that can be attributed to different sensor grid densities between floors, as this change is a result of numerous confounding factors. Additionally, the data used in this chapter were taken at five-minute timesteps, which may be higher resolution than many traditional BAS systems are capable of logging. The effect of changing these parameters would be a worthy research question.
- *Where and when are these specific models applicable?* While this chapter showed that models could be developed in one floor and used in another in the case study building, the transferability of these models was still within a single building with an academic office-type occupancy across all floors. A worthy research question would be how these models would perform in a completely different building of the same occupancy type, or a building of a different occupancy type altogether.

It should be noted that the findings presented in this chapter are explicitly related to the case study building. This research highlighted that the performance of Wi-Fi enabled device count as a lone data stream was only exceeded by models which included Wi-Fi data in conjunction with additional sensor data streams; however, the marginal improvements in the prediction accuracy of these models may not justify the additional

capital cost in buildings where these sensors are not already installed. The cost-effective, non-intrusive, accurate and simple nature of Wi-Fi device count alone make it the best candidate for opportunistic occupant-count estimation in the case study building. As these data are isolated from the BAS, aggregated plug-in equipment and lighting load data could be used to facilitate the use of these occupant-count estimation for practical controls applications.

Chapter 3

This chapter has been submitted for publication as:

Clustering and motif identification for occupancy-centric control of an air handling unit.

Hobson, B.W.; Gunay, H.B.; Ashouri, A.; and Newsham, G.R. *Energy and Buildings*.

3. Rules-based occupancy prediction

3.1. Introduction

As building heating, ventilation, and air-conditioning (HVAC) and envelope technologies improve, the influence of occupants on commercial building energy use increases [58, 59]. This influence stems from occupants' presence and energy-related behaviours [60]. To improve the operation of HVAC systems, services must be provided to the building when and where they are needed, in the amount that they are needed [9]. Currently, most commercial buildings operate using a set of fixed schedules for this purpose, which assume full or near full occupancy during working hours [22, 61]. In reality, buildings rarely reach this level of occupancy. For example, a study by Barbour *et al.* [62] showed that the average occupancy of commercial buildings in downtown Boston was 20% of the designed occupant capacity. There is therefore a significant mismatch between the services being provided and the actual occupancy of many commercial buildings, resulting in overventilation and energy waste. Considering that HVAC systems account for 15% of the secondary energy use, and 10% of greenhouse gas emissions in commercial buildings in Canada [1], the use of accurate occupancy-count estimates in commercial building HVAC operations is pertinent to reducing global energy use and CO₂ emissions. However, reactive controls based on instantaneous occupancy-count estimates alone are insufficient for optimal operation of system-level equipment, such as air handling units (AHUs). This is due to the transient nature of buildings' thermal and air quality conditions; it may take several hours for equipment actuation to achieve the desired setpoint or contaminant level. Instead, forecasting occupancy levels can allow a building automation system (BAS) to make proactive and informed decisions about AHU operation. Implementation of these

controls in a BAS to reduce the AHU's outdoor air (OA) fraction based on forecasted occupancy may realize significant energy savings by reducing chronic overventilation [63, 64].

A variety of technologies for occupant counting have been previously developed and studied. Jung and Jazizadeh [65] and Saha *et al.* [66] recently conducted comprehensive literature reviews exploring technologies and data analytics approaches for occupant-count estimation, respectively. Optical and/or infrared camera-based technologies (e.g., [10, 24–29]) coupled with image recognition algorithms can be deployed at entrances and exits to determine occupancy at the room, floor, or building level. These people-counting cameras are becoming relatively popular in large commercial buildings as a part of security networks. Another common source of occupancy data from security networks is card access (e.g., badge in/out) data in buildings with controlled access. Radio-frequency identification (RFID) technologies (e.g., [30, 31, 67, 68]) can be used to monitor occupants' distributions within a space. Bluetooth (e.g., [69–71]) and, to a greater extent, Wi-Fi (e.g., [36, 40, 46–50, 72]) data have been used as a simple and accurate source of occupant-count data. Wi-Fi is advantageous due to the prevalence of Wi-Fi enabled devices, with sensing infrastructure such as access points (APs) already installed in many buildings for IT purposes. The occupant-count data produced from each of these technologies can generate insights into building occupancy patterns that can be used to improve building operations. However, security and privacy concerns coupled with technological limitations remain a tangible barrier to the integration of these standalone sensors or sensor networks (e.g., IT services and security networks) into BASs. This puts explicit occupant-count data out-of-

reach for most commercially available BASs; their direct use for occupancy-centric HVAC controls is therefore limited given these constraints.

Indeed, currently installed BASs vary considerably in capability to acquire or store data related to occupancy. This range of capability can be attributed to a lack of standardization of BASs [73, 74] and the wide-ranging ages of buildings within the commercial building stock [75]. Regardless, data logging is becoming a core functionality of many BAS service frameworks [76, 77] (e.g., BACnet, BACnet/WS, OPC UA, oBIX). A commonly logged data source in the BAS is electrical loads. Electrical load has been shown to exhibit a strong relationship with building occupancy (e.g., [37–39, 41, 43]), especially energy-use patterns that are directly caused by occupants, such as plug-in equipment and lighting loads [78]. Therefore, trends in recorded occupant-count estimates from standalone data sources can be analyzed and compared to trends in recorded plug-in equipment and lighting electricity use. Predictable relationships between occupancy and lighting/plug loads can be identified and trained offline. The electricity use for plug loads and lighting can then be used for online forecasting of occupancy in the absence of the explicit occupant counts from standalone data sources, which are traditionally isolated from the HVAC control network (e.g., Wi-Fi, people counting camera networks, card access records, etc.). The forecasting technique employed must be simple and robust enough to operate with the relatively limited computational power of a BAS. In buildings with this metering, leveraging two disparate data streams *Wi-Fi* and *electricity use from plug loads and lighting* would result in no additional capital cost for sensing or control infrastructure.

This chapter presents a clustering and motif identification-based approach for day-ahead forecasting of building occupancy patterns, using an academic office building as a case

study. Seven months of concurrent Wi-Fi device count, plug load, and light load time-series data were collected. The Wi-Fi data were processed through an occupant-count estimation function to create estimated occupancy profiles for each day. To determine a handful of representative profiles, the seven-months of daily occupancy profiles were clustered using three different methods, namely agglomerative hierarchical clustering (AHC) with correlation dissimilarity or Euclidean distance, and k-means clustering. The cluster's quality was analyzed and quantified, and the preferred clustering technique was selected. Representative occupancy and corresponding lighting and plug load day types for each cluster were produced. Each representative occupancy day type was quantized using alphabetic characters and a character was assigned to each day. Rule extraction was employed on the sequence of over 200 characters corresponding to each unique day in the study period to identify common occurrence patterns. These recurring sequences of day types, called *motifs*, were used in a classification tree to generate rules for day-ahead forecasts of occupancy day types. With the offline portion of model development and training complete, an online forecasting approach was implemented. Time-series plug-in equipment and lighting load data were compared to the representative lighting and plug load day types and the appropriate corresponding character was assigned to each day. This sequence of characters was then processed using the classification tree to predict which occupancy day type the next day will belong to. The applications of the insights generated by this approach, as well as the forecasting accuracy and error propagation, are discussed. The goal of the developed approach is to ultimately implement this forecasting technique in building AHUs to control the OA dampers based on occupancy, which will be the subject of the next chapter.

3.2. Literature review

Autoregressive integrated moving average (ARIMA) models are a class of statistical modelling that have been applied for occupancy forecasting with good accuracy (e.g., [49, 79]). ARIMA has also been used in energy use projections [80], and is used in other disciplines for many other applications. Liu *et al.* [81] popularized an online learning approach whereby predictions are replaced with measured values as the time-series progresses, reducing the model drift and leading to high accuracy in both linear and non-linear systems. However, due to signal noise, the accuracy of these algorithms still decreases significantly as the length of the prediction horizon increases. This is problematic when considering a day-ahead prediction horizon to modulate system-level equipment. Additionally, while implementation of online forecasting techniques such as ARIMA are technically possible in commercial BASs, they require real-time accurate occupant-count data such as Wi-Fi device counts, which are currently unavailable in almost all BASs as discussed previously.

To address these shortcomings, unsupervised machine learning and visual analytical techniques can be used to produce time-series forecasts. Where a supervised technique seeks to predict labels based on a set of features, unsupervised techniques seek to characterize these labels irrespective of any features. A few literature reviews have been completed previously on unsupervised machine learning techniques for analysis of building operations; these reviews have covered a wide swath of applications largely related to energy system load profiling rather than occupancy forecasting directly, such as the reviews conducted by Morais *et al.* [82] and Chicco [83]. Miller *et al.* [84] conducted a review of these previous studies that highlighted unsupervised machine learning and

visual analytical techniques. These techniques include clustering, motif and discord identification, and rule extraction.

The two most popular clustering methods in the literature are hierarchical clustering and k-means clustering. Hierarchical clustering depends upon a matrix of pairwise dissimilarities for a set of variables. The subset of hierarchical clustering methods whereby the two least dissimilar variables are grouped is called agglomerative hierarchical clustering (AHC) in the literature [85, 86]. Different dissimilarity metrics can be used for hierarchical clustering such as Euclidean distance, Manhattan distance, Minkowski distance, or correlation dissimilarity [87]. The choice of these metrics can have a significant impact on the shape, size, and number of clusters generated. The two least dissimilar variables are grouped and their centroidal value is taken; this process is repeated until a single cluster is formed for the entire set of observations. These relationships are typically represented by a dendrogram, which can be pruned to yield the desired number of clusters. K-means clustering, on the other hand, is a partition-based clustering method [88]. The Euclidean space in which the data points lie is partitioned, and the points are clustered in the partition (i.e., Voronoi cell) that minimizes the sum of squared distances between the centroid of the cell and said points [89]. This process is stochastic as the cell centres are randomly initialized. As this method does not use a pairwise dissimilarity matrix, no dendrogram is produced. Instead, the number of desired clusters k is specified when initializing this unsupervised algorithm. For building energy applications, clustering is typically applied to time-series data which has been segregated into 24-hour (i.e., daily) periods to determine typical day types. This approach is commonly referred to as load

profiling [84], and has been used widely in predicting heating or cooling (e.g., [90–97]) and electricity loads (e.g., [98–103]).

Motif and discord identification is an exploratory technique which seeks to characterize common trends (i.e., motifs) or outliers (i.e., discords) in time-series data [104]. This is often achieved by using symbolic quantization. Quantization is a technique by which a continuous or semi-continuous input data set is transformed into a smaller, discrete output data set to facilitate insights such as pattern recognition [105]. This is done by binning values at discrete intervals based on one or more criteria. In symbolic quantization, an alphabetic character or ‘symbol’ is then assigned to each bin. A collection of the resulting letters over a specified number of intervals creates a word. Frequently repeated words are referred to as *motifs*, and infrequent words are referred to as *discords*. Symbolic aggregate approximation (SAX) is a symbolic quantization method developed by Lin *et al.* [106]. This methodology directly quantizes time-series data into motifs and discords, and was applied by Ashouri *et al.* [107] for heating and cooling load forecasting. However, it should be noted that motif and discord identification is a broad term and refers to any approach that seeks to characterize patterns in time-series data. Therefore, clustering is considered a binning approach in the context of symbolic quantization and motif and discord detection. Rule extraction is a machine learning technique that aims to find relationships between variables [84]. Decision trees are a type of supervised rule extraction algorithm, with classification and regression trees (CART) [108] being a common form of decision trees in application. CARTs are used for prediction and forecasting of a variable based on maximum likelihood. Classification trees are used specifically when the labels and features are categorical variables, such as letters or motifs generated by symbolic quantization

methods. This algorithm produces a tree-like structure with a conditional logic. Each interior node of the tree represents a variable, and each leaf represents a decision based on the conditional response to said variable. This node-leaf structure is propagated through a series of cascading subordinate responses until a tree of the desired size is produced or another stopping criterion is met. This tree can be pruned at any arbitrary point to reduce complexity and overfitting.

The conditional responses at each node of a classification tree are binary in nature (i.e., a given condition is either satisfied or not). Given this structure, the underlying conditional logic of a decision tree can be converted to a simple rules-based Boolean logic which can be easily used in commercial building controllers. Therefore, by utilizing clustering, motif and discord identification, and rule extraction, the complexity of occupancy forecasting can be simplified to this Boolean logic for implementation purposes in existing BASs. This paper is focused on developing this novel forecasting technique for the purpose of practical controls applications in air handling units. As noted in the review by Park *et al.* [109], this study's focus on system-level equipment address a gap in existing occupancy-centric controls literature, which has focused almost exclusively on smaller scale zone-level applications to date.

3.3. Methodology

The methodology developed in this study is summarized as follows. Data from an academic office building were collected for this study. The proposed methodology consists of two major phases: offline training and online forecasting. During the offline training phase, clusters and motifs were identified. Clustering was employed to identify a handful of representative day types. Motif identification determined common occurrence patterns of

these day types over the sequence of days. This phase required archived lighting and plug load data from the BAS, as well as Wi-Fi data for occupant-count estimation from the isolated IT system (i.e., not connected to the BAS). Typical occupancy day types were generated via clustering of the Wi-Fi data. A classification tree was developed using the occupancy motifs. These day types were then associated with corresponding lighting and plug-in equipment electrical use patterns. This offline training phase was conducted in R [54] and was necessary as most commercially available BASs are not capable of handling the computational and analytical requirements of running these machine learning algorithms. The online forecasting phase then used real-time plug-in equipment and lighting electricity use data (already available in the BAS) and compared them to the corresponding lighting and plug load patterns from the offline training phase to determine which day type the day belonged to. These day types were recorded, and the resulting sequence of characters was analyzed using the logic from the classification tree to predict the day type one day in advance. Control decisions can then be executed to modulate the AHUs dampers and control outdoor airflow rates, thus reducing chronic overventilation.

3.3.1. Building and dataset

The case study building is a 6650 m² academic office building located in Ottawa, Canada. The building contains a mixture of office spaces, study spaces, classrooms, and laboratories. It is served by two AHUs (AHU1 and AHU 2) with a total airflow rate of 25,000 L/s. The minimum OA damper positions at the time of this study were set to 40% and 30% for AHU1 and AHU2, respectively. The estimated breathing zone outdoor airflow rate is 10,000 L/s. Assuming the common rule-of-thumb of 10 L/s-person per the default

combined outdoor airflow rates provided in ASHRAE 62.1-2019 [110], the building is constantly providing enough outdoor air for 1000 occupants during occupied hours.

The seven-month data collection period for the study extended from late January to late August. Lighting and plug load and Wi-Fi enabled device data were collected at concurrent five-minute timesteps for the entire study period. Wi-Fi enabled device data were collected by the Wi-Fi APs spread throughout each floor. Coverage maps were provided by IT services to verify that the APs did not have any dead zones. Anonymized Wi-Fi enabled device counts were provided by IT services; no MAC addresses were monitored or stored over the course of this study. Possible sources of error in the provided data include counting devices outside of the building within range of the APs, as well as double counting of devices if occupants' devices moved to a different Wi-Fi APs coverage zone within the five-minute timestep. The plug-in equipment and lighting electricity use data were accessible through the BAS in the form of floor-level lighting and plug loads. These data were aggregated to the building level and exclude fans, pumps, and other HVAC loads.

3.3.2. Occupant-count estimation

Eight-and-a-half weekdays of ground truth occupancy data were collected using cameras at the building's entrances and exits: four-and-a-half during the winter (i.e., academic) semester and four during the summer semester. Two continuous two-day recording sessions were used each semester. The footage was manually reviewed by a researcher to record the occupant-count at each timestep. For privacy reasons, the cameras did not record any audio and the footage was deleted immediately after review. This method was approved by the institution's research ethics board. Linear regression models were shown in the previous chapter to produce a strong correlation between Wi-Fi enabled device count

and the number of occupants for occupant-count estimation purposes [72]. The data from occupied hours were randomly partitioned into 70% training and 30% testing sets; a linear regression model was employed with a R^2 of 84.5% and normalized root mean square error (NRMSE) of 8.9% on the test data, indicating a strong positive linear correlation for occupant-count estimation purposes, see Figure 3.1. The normal distribution of the model residuals indicates that the occupant-count estimates produced are within $\pm 10.4\%$ of the building's true occupancy at 95% confidence. As ground truth occupancy data were only collected for eight-and-a-half days during the study, a fundamental assumption was that the nature of this relationship holds true for the entire study period. Artificial occupant-counts are then produced for the duration of the study using this model.

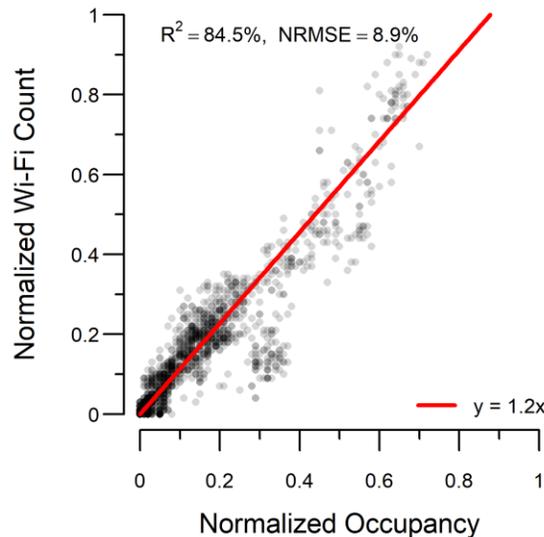


Figure 3.1: Relationship between Wi-Fi device counts and ground truth occupancy after min-max normalization.

3.3.3. Clustering and motif identification

The time-series occupant-count estimations underwent min-max normalization and were binned on a daily basis. After evaluating different dissimilarity metrics, the correlation dissimilarity ($1-\rho$) and Euclidean distance (d) metrics were selected for AHC. The pairwise

dissimilarity matrix between each daily profile for each distance metric was computed and AHC was performed using R's *clues* package [111] on the dissimilarity matrices. K-means clustering using Lloyd's algorithm [89] was performed directly on the normalized daily data using R's *stats* package [57].

In order to reduce the dimensionality of the occupancy day types which need to be programmed in the BAS, a handful of representative day types should be selected. Selecting too few will result in little to no insight into motifs (i.e., frequently repeated sequences of days) and discords (i.e., infrequently repeated sequences of days), whereas selecting too many may become cumbersome to implement and defeat the purpose of the clustering exercise. After experimenting with clustering parameters and the resulting model accuracies, the ideal number of clusters was limited to between two and ten for the case study building. Knowing the ideal number of clusters for each technique, the dendrograms for the AHC methods were cut to produce the desired number of clusters for each distance metric. Similarly, the number of clusters for k-means clustering was selected, and clustering was performed on the daily profiles to identify representative day types. For the purpose of this study, profiles were considered representative if they occurred five or more times during the seven-month study, whereas a profile was considered an outlier if it occurred less than five times. The day types were then assigned symbolic characters.

As the desired number of clusters is inherently not known, the Caliński-Harabasz (CH) index [112] was computed for each clustering technique and a varying number of clusters using R's *clues* package [111]. The CH index is the ratio of inter-cluster variance to the intra-cluster variance, with a multiplier to penalize the overfitting as the number of clusters increases. It should be noted that while a higher CH index indicates 'better' clustering, this

quantitative metric is meant to indicate the quality of the clustering irrespective of contextual or practical factors and should therefore be interpreted carefully before the desired number of clusters is selected. The number of clusters that yields the highest CH index were plotted for each clustering technique for visual inspection.

In order to capture a representative occupant-count estimate for each individual day type, the upper 95th percentile of values at each timestep for the daily profiles belonging to that day type was conservatively taken. This is because these occupant-count estimates are being used for ventilation purposes, and as such the upper values are used to avoid potential underventilation (i.e., it is conceivable that if the cluster centers were used as the representative occupancy day types, approximately half of the days would under-predict the occupancy and ventilation which could result in indoor air quality degradation). Based on visual inspection of the clusters and the CH index, the preferred clustering technique was selected.

With symbolic characters assigned to each occupancy day type, a string of approximately 200 characters was produced for the seven-month study period. This string of characters was binned into seven-character long ‘words’ or sequences which represent the pattern of daily occupancy day types in the week leading up to each day. These words were used as predictors, and the daily characters were used as the predictand in a classification tree to extract motif and discord words. The dataset was randomly partitioned into 70% training and 30% testing sets, and a classification tree was trained using R’s *caret* package [113]. The classification tree with the best model fit to the testing data was selected by performing 10-fold cross validation. The classification tree was pruned to the nodes which provided greater than 5% improvement to reduce both the size of the tree and overfitting.

Similar lighting and plug load day types are expected to correspond at a high-level to the occupancy day types (e.g., if occupancy is ‘high’ on one day, the electrical load is also expected to be ‘high’). Using the cluster membership from the preferred occupancy clustering technique, daily lighting and plug load profiles were grouped with the same membership to produce corresponding representative lighting and plug load day types. The representative profiles were used as the cluster centers for the plug-in equipment and lighting electricity use data.

3.3.4. Online forecasting

The sum of the square errors (SSE) was computed for each daily plug-in equipment and lighting electricity use profile and the representative lighting and plug load profiles, and the character for the representative profile with the lowest associated SSE was assigned to that day. This was repeated for each day of the seven-month study to produce a string of approximately 200 characters. These characters were then grouped into seven-character long sequences (i.e., in the same fashion as the occupancy data) and fed through the previously trained classification tree to produce day-ahead day type estimates. These lighting and plug load based day type forecasts and their associated occupant-count estimates were compared over the study period to the Wi-Fi based occupant-count estimates and the building’s current occupancy schedule to determine the accuracy of the occupant-count estimates, and by extension, the possible reduction in overventilation.

3.4. Results and discussion

3.4.1. Occupant-count estimation

The Wi-Fi enabled device data were normalized and fitted to the previously trained linear regression model to produce artificial time-series occupant-count estimates for the entire

seven-month study for the whole building. These occupant-count estimates were amalgamated into weekly profiles for each semester to broadly characterize the building’s occupancy. The average weekly occupancy profiles and the upper (Q3) and lower (Q1) quartile profiles for each semester can be seen in Figure 3.2.

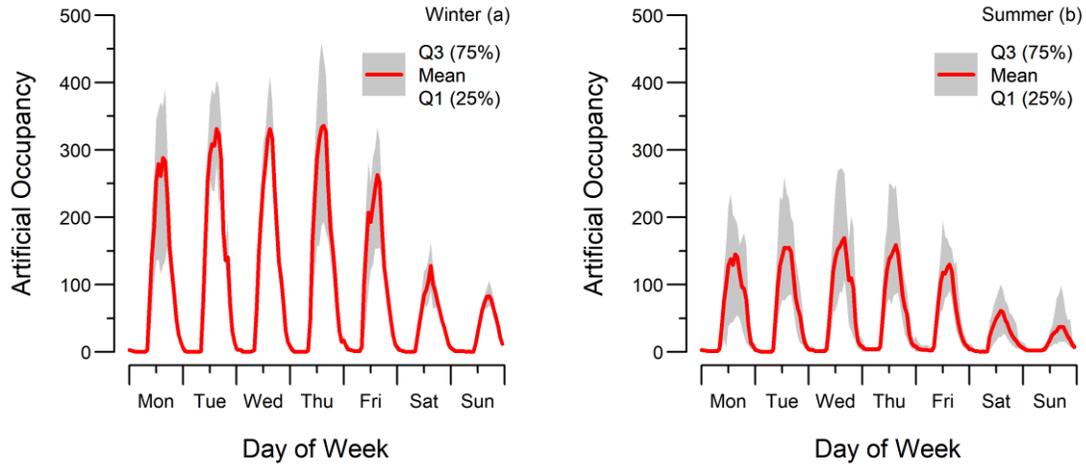


Figure 3.2: Average weekly occupancy profiles for (a) the winter (i.e., academic) and (b) the summer semesters in the case study building. Note that the values are calculated at each timestep to produce a representative, artificial profile.

Upon visual inspection of the artificial occupant-counts, high-level occupancy patterns of the case study building can be observed. The building follows a cyclic occupancy pattern, with higher occupancy during the winter academic semester (i.e., from January to late April) and lower occupancy during the summer semester (i.e., from May to the end of August). This is expected to be due to the comparatively limited number of classes and undergraduate laboratories being used over the summer semester. Generally, the occupancy during the summer semester had greater relative fluctuation on each day than the winter semester. Based on this analysis, the maximum estimated occupancy in the case study building was 515 ± 54 persons over the duration of the study, or approximately 46 to 57% of the 1000 occupant assumption used for current operations. The average peak occupancy during weekdays was between 189 ± 20 and 226 ± 24 persons, or approximately

17 to 25% of the assumed occupancy. Recall the study by Barbour *et al.* [62] found that the average peak occupancy in commercial buildings in Boston was similar at 20% of the designed maximum capacity. Based on this analysis, the building is being overventilated consistently throughout the study period, which has implications on energy use.

With knowledge of the seasonal occupancy patterns in the building, there is an opportunity to develop a simple ventilation schedule based on these occupant-count estimates. Reducing the minimum ventilation to the maximum estimated occupancy for an academic and summer weekday, as well as the lower occupancy weekend, could potentially reduce overventilation significantly from the 1000 occupant assumption currently used. This is referred to as the ‘naïve’ approach as it can be implemented at an early stage of analysis with a single data stream (e.g., offline Wi-Fi device counts or another occupant-count estimation). For the naïve approach, the data were partitioned into academic weekdays, summer weekdays, and weekends, and the 95th percentile of the maximum daily occupancy values was taken. The maximum occupancies were 403 ± 42 , 194 ± 20 , and 104 ± 11 persons for the academic weekday, summer weekday, and weekend, respectively. This approach will be used as a benchmark to quantify the additional energy savings possible through the proposed online forecasting technique.

3.4.2. Clustering and motif identification

The daily occupant-count profiles were clustered using k-means clustering, and AHC using correlation dissimilarity ($1-\rho$) or Euclidean distance (d). The number of clusters was varied from two to ten, and the CH index was computed to determine the number of clusters for each technique which produced the highest cluster quality, see Figure 3.3. Based on the results, the ideal number of clusters – or occupancy profiles – for AHC with correlation

dissimilarity ($1-\rho$) was six, with five profiles being ideal for AHC with Euclidean distance (d) and k-means clustering. The CH index indicated that the clusters produced by k-means were of higher quality than those of either AHC method. The Euclidean distance metric also produced higher quality clusters than the correlation dissimilarity distance metric.

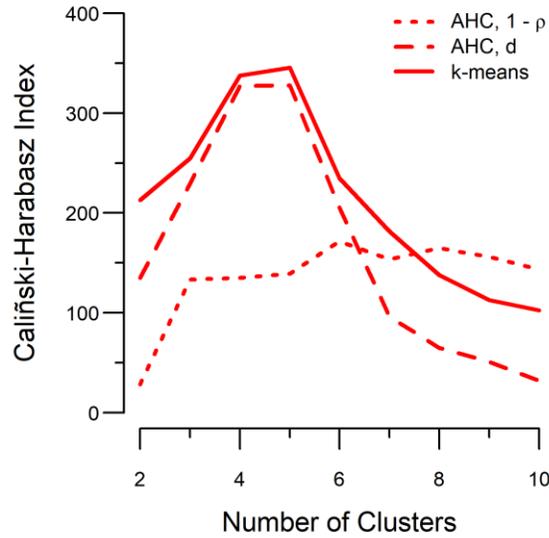


Figure 3.3: Caliński-Harabasz index for clustering techniques and varying number of clusters. A higher value indicates 'better' cluster quality.

While different numbers of occupancy profiles (i.e., clusters) were selected for the different clustering techniques, each method produced four representative (i.e., five or more occurrences) occupant-count profiles. Therefore, AHC using correlation dissimilarity ($1-\rho$) had two outlier profiles (i.e., six total profiles minus four representative), and AHC using Euclidean distance and k-means clustering had one outlier (i.e., five total profiles minus four representative). These representative day types were then quantized by assigning a symbolic character to represent the individual letter in the overall word (i.e., sequence of characters) in which the day occurs. A similar naming convention established by Ashouri *et al.* [114] was used, and the occupancy profiles with the highest to lowest peak estimated occupancy were assigned the letters 'h', 'm', 'l', and 'L', for 'high', 'medium', 'low', and

‘very low’, respectively. Outliers were assigned the characters ‘ x_1 ’ for the first outlier, and ‘ x_2 ’ was used for the second outlier in the case of AHC using correlation dissimilarity ($1-\rho$). Outlier ‘ x_1 ’ was a single outlying day for all three clustering techniques; this was why the CH index between four and five clusters for AHC using Euclidean distance and k-means clustering remains relatively unchanged, recall Figure 3.3. The representative occupancy profiles for each clustering technique and their family of curves are summarized in Figure 3.4a-c.

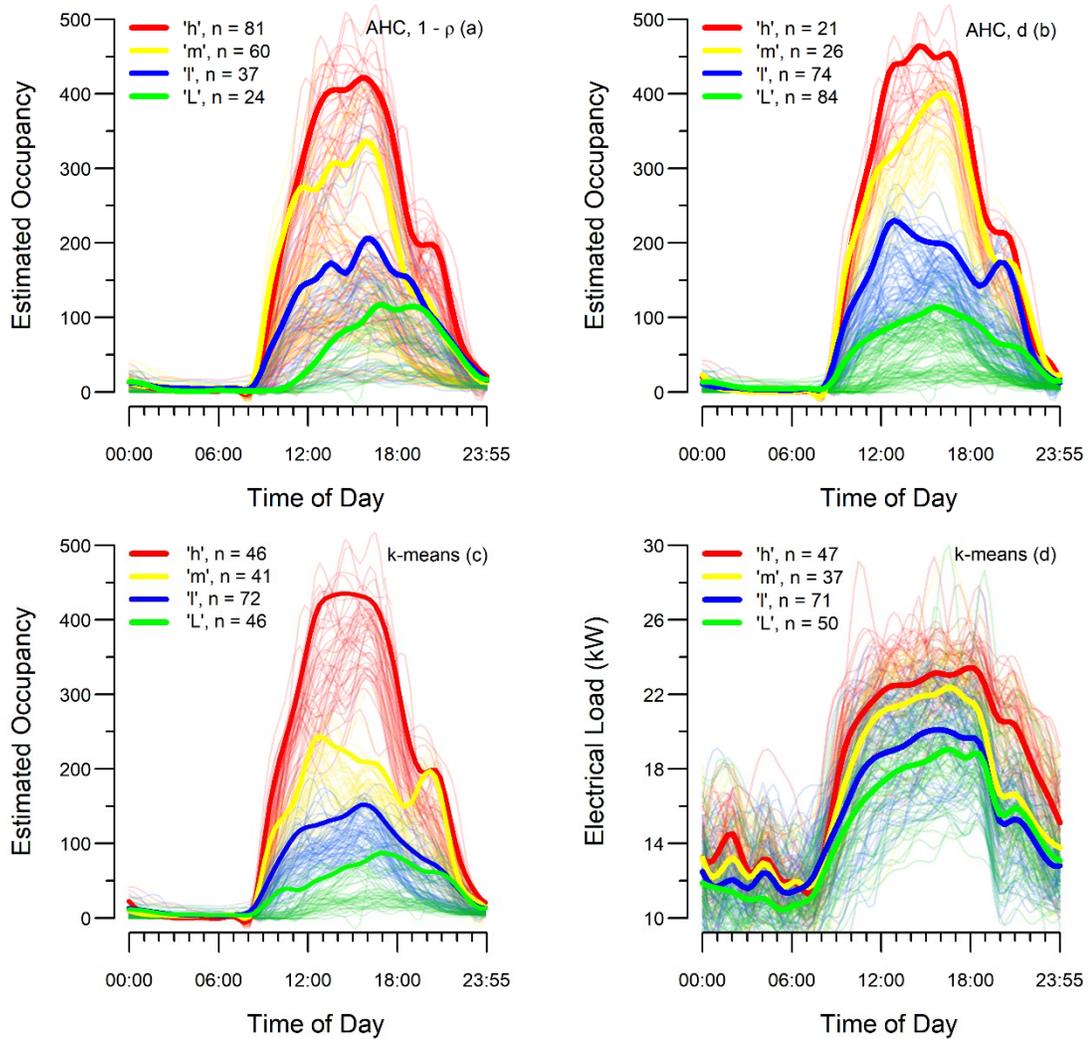


Figure 3.4: Representative occupancy profiles for each clustering technique (a) – (c), and representative lighting and plug load profiles (d), where n is the number of days belonging to each cluster.

Upon inspecting the occupancy profiles, it was evident that AHC with correlation dissimilarity ($1-\rho$) had the poorest cluster quality. As correlation is a measure of the linear relationship between two variables, the shape of the daily profile carries more weight than the magnitude; this means that days with low occupancy and high occupancy can be found to have a high correlation if the shapes of their occupancy patterns are similar. This results in distinctive shapes for each of the four representative occupancy profiles, but a wide spread of maximum occupancy values within each cluster, which does not lend itself well to calculating minimum required ventilation for the proposed application. As AHC with Euclidean distance (d) and k-means clustering are both geometric clustering methods, they perform well in clustering shapes and profiles that lie in a Euclidean plane. Therefore, both techniques are valid for the proposed application. However, as k-means clustering has a higher CH index between these two techniques, it was selected as the preferred clustering technique. The peak occupancy for the high 'h' occupancy cluster using k-means clustering was approximately 450 persons.

However, in order to forecast occupancy within the BAS, accessible lighting and plug load data must be used to determine the next day type. Representative lighting and plug load day types that correspond to each occupancy day type were produced. The average (i.e., mean) value at each timestep for the daily lighting and plug load profiles belonging to each cluster were used to create the cluster centers. The SSE between the daily lighting and plug load time-series data and the cluster centers was computed, and the corresponding character for each representative occupancy profile was assigned to that day. As lighting and plug load is not perfectly correlated with occupancy, there was some misclassification, such as when there was an abnormally high or low lighting and plug load on a low or high

occupancy day, respectively. This can be seen in the different number of days belonging to each cluster in Figure 3.4d (recall in Figure 3.4c ‘h’, ‘m’, ‘l’, and ‘L’ had 46, 41, 72, and 46 days, respectively). Overall, 84.5% of days were correctly reclassified (i.e., had the same membership as the occupancy profiles). Recall that this conversion to lighting and plug load was necessary as these data are monitored by the BAS, whereas Wi-Fi data are not.

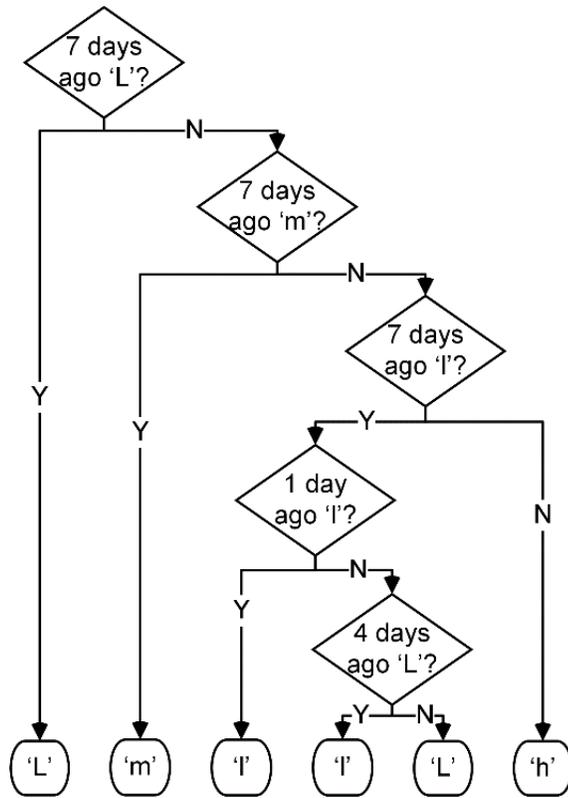


Figure 3.5: Classification tree for forecasting occupancy day type in the case study building one day in advance, based on the past 7 days of observation.

Using the characters assigned to each day from k-means clustering, the characters were binned into seven-character long words representing the previous seven days of occupancy day types. A classification tree was trained to determine motif and discord words using the randomized 70% training portion of the data, and nodes were removed that contributed to

less than a 5% increase in successful classification, see Figure 3.5. With the offline models and classification tree trained, the approach can transition into the online forecasting phase.

3.4.3. Online forecasting

The misclassification that occurs when transitioning to the use of lighting and plug load data was propagated through the classification tree, which contributes to forecasting error. This resulted in a net successful classification rate of 70.4% when using the lighting and plug load motifs in tandem with the classification tree, see Figure 3.6b, compared to the actual cluster membership from k-means clustering using Wi-Fi based occupant-count estimation data, see Figure 3.6a. The error in classification can be attributed to the classification tree depending heavily on the observations from seven days ago (i.e., on the same day the previous week). This occasionally resulted in a one-week ‘lag’ period where the forecast may over or underestimate the occupancy day type, such as the transition after the week of April 13th into exams and the summer semester where the buildings occupancy switches to a lower, less predictable occupancy pattern (recall Figure 3.2). This means that, during the academic semester when the occupancy motif was somewhat static (i.e., ‘hhhhhL’ from Monday to Sunday), occupancy forecasting may not be necessary to accurately predict the day type during the weekdays and weekends. However, once the less predictable summer semester has begun, the algorithm was able to respond to the evolving occupancy patterns as no single motif was able to fully capture the changes in occupancy during this semester. This suggests that the benefits of such a forecasting application may not be as pronounced in traditional occupancy scenarios, such as an office building with a rigid and predictable schedule. Indeed, commercial buildings with less variation between days may reduce the opportunity for savings from a dynamic ventilation schedule. But as

workplaces with rigid and predictable schedules are becoming less and less common in the modern landscape [115], many workplaces are transitioning to flexible working hours [6, 116]. Adapting occupancy schedules to match these flexible working hours could therefore offer increased HVAC savings over a static scheduling approach in these types of modern workplaces, with increased savings achievable in workplaces with larger fluctuations in occupancy [117, 118].

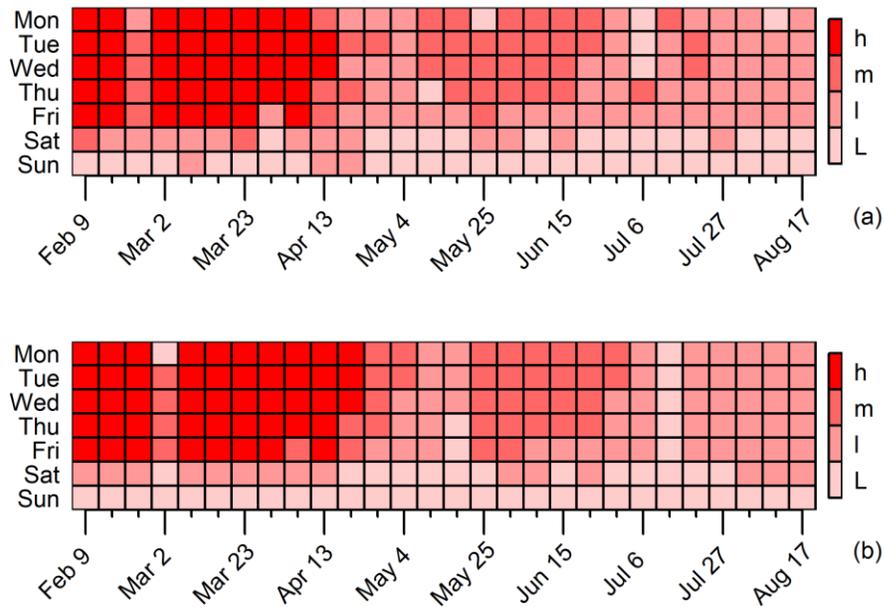


Figure 3.6: Cluster membership (a) using Wi-Fi data and k-means clustering and (b) lighting and plug-in equipment electricity use forecasting.

While the k-means clustering-based Wi-Fi occupancy profiles were compared to the lighting and plug load forecast to see the relative success of the forecasting algorithm, it should be noted some overventilation still occurred with both approaches. This was due to the conservative use of the 95th percentile for each representative day type to avoid underventilation and indoor air quality degradation. However, this slight overestimation of occupancy was comparatively small to the default assumption of 1000 persons, or even the naïve approach which reduced the minimum ventilation to the highest observed occupancy

for academic and summer weekdays and weekends Recall these peak occupancies were 403 ± 42 , 194 ± 20 , and 104 ± 11 persons ($\pm 10.4\%$), respectively, see Figure 3.7. While using a naïve scheduling approach significantly decreased the error to 150 ± 178 persons compared to the current default error of 882 ± 169 persons (i.e., 17% of the default mean error), there was still substantial benefit to employing the lighting and plug load-based forecasting method. This further reduced the error to 47 ± 69 persons at 95% confidence (i.e., 5% of the default mean error) and produced a narrower error range. Conversely, the benefit of using k-means clustering on the Wi-Fi occupant-count data directly was not as pronounced. While the mean error and range was lower (i.e., 41 ± 63 persons), this represented only an additional 0.7% reduction in the mean error compared to the lighting and plug load-based forecasting method.

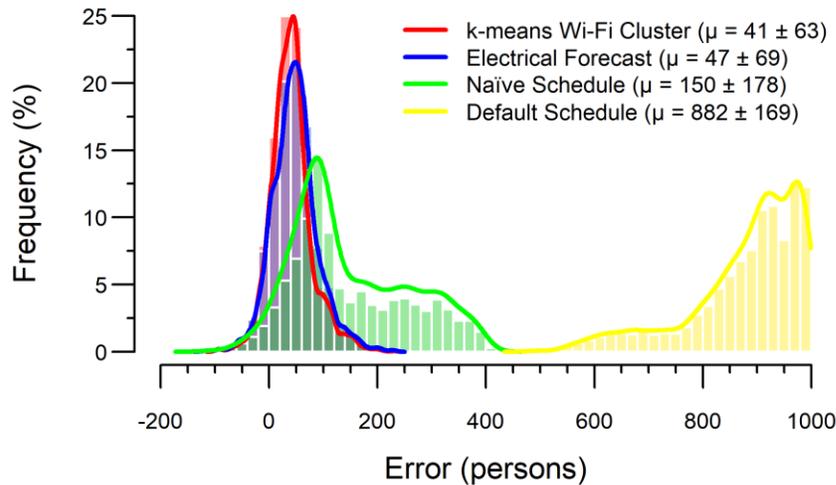


Figure 3.7: Difference between occupant-count estimate time-series data and different occupancy assumptions during occupied hours.

To this end, it can be said that the added complexity and accuracy of the forecasting approach can be justified over the relatively higher error and simplicity of the naïve schedule approach. However, the added cost of integrating Wi-Fi enabled device counts into the BAS would not justify the comparatively small increase in accuracy where the

electrical submetering infrastructure is already in place. Therefore, the use of lighting and plug load data for online forecasting of occupancy day types using a classification tree has the potential to reduce overventilation when compared to static ventilation schedules, with accuracy comparable to that of potentially costlier BAS integrated Wi-Fi enabled device counts.

3.4.4. Proposed intervention

With the offline model training completed and the classification tree for online forecasting developed, the outlined approach can be applied to the BAS to control the OA damper in the AHUs. Lighting and plug load data will be recorded by the BAS each day, and the SSE between the recorded data and each of the representative lighting and plug load day types will be computed at the end of the day. The day type and corresponding character with the lowest SSE will then be assigned to that day. This process will be repeated each day to develop an updated seven-character long word. Simultaneously, the updated word will be passed through the rules-based logic of the classification tree to predict the next occupancy day type one day in advance and modulate the OA damper appropriately. An OA damper position profile for each of the occupancy day types must be developed for the AHU to ensure the appropriate outdoor air fraction (OAF) is achieved for the required number of occupants. To determine this, the person outdoor air rate must be disaggregated from the area outdoor air rate. This can be achieved with the building floor area using minimum ventilation requirements set out in ASHRAE 62.1-2019 [110].

If a linear relationship between the damper position and OAF is assumed, the OA damper in the case study building's AHUs can be modulated in increments of 1% (i.e., from 0% when fully shut to 100% when fully opened), each 1% increase in damper opening beyond

the minimum area outdoor air rate corresponds to an approximate 40 person increase in occupancy. Recall that the occupancy error for the proposed forecasting model was 47 ± 69 persons at 95% confidence. Therefore, even at the upper threshold of confidence, the largest error in damper movement that could be achieved using the occupancy forecast is under 3% for overventilation, or under 1% in the case of underventilation.

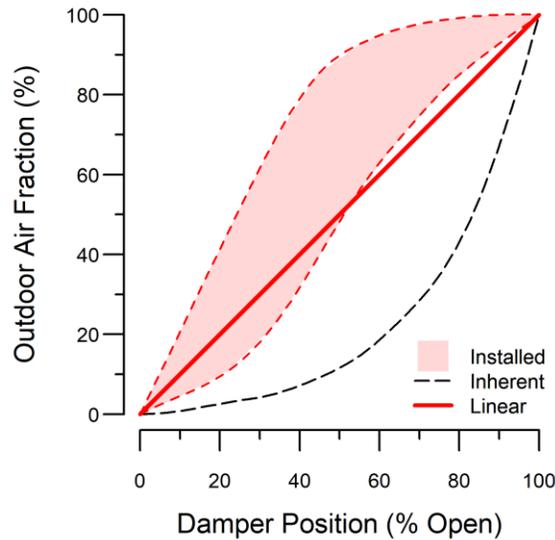


Figure 3.8: Outdoor air damper position vs. outdoor air fraction, adapted from ASHRAE Handbook.

However, the assumption of a linear relationship between the OAF and the damper position isn't necessarily accurate, especially at lower ventilation rates. As can be seen in Figure 3.8, installed OA dampers do not have a linear relationship with the OAF. While idealized laboratory tests have shown that the inherent relationship between the damper position and OAF is exponential, installed dampers have been experimentally found to exhibit a logarithmic relationship that varies based on the differential between atmospheric pressure and the pressure in the mixing plenum [119]. The actual performance of the building's OA damper lies somewhere within this family of curves, which will affect the damper position profile.

3.5. Closing remarks

In this chapter, historical Wi-Fi enabled device count and building level plug-in equipment and lighting loads were used to create a rules-based approach for day-ahead forecasts of occupancy-levels in an academic office building in Ottawa, Canada. Clustering and motif and discord identification were employed to extract representative daily occupancy profiles and corresponding electrical load profiles. A classification tree was trained using the electrical load data to predict occupancy one-day in advance with 70.4% accuracy. The majority of misclassification came from overclassified days, which is conservative for ventilation purposes.

Key findings from this chapter are summarized as follows:

- The building's maximum estimated occupancy was demonstrated to be only 46% to 57% of the 1000-person assumption used in its current HVAC operation, with a mean peak occupancy of approximately 17% to 25%.
- K-means clustering was found to be the preferred clustering method for extracting typical occupancy profiles compared to either of the agglomerative hierarchical clustering methods.
- Four representative daily occupancy profiles high 'h', medium 'm', low 'l', and very low 'L' were generated using the 95th percentile of the k-means clustering data.
- A classification tree was developed with a successful classification rate of 70.4% and error of 47 ± 69 persons at 95% confidence. This translates into 1% and 3% error in damper movement for underventilation and overventilation, respectively.

- Compared to an error of 882 ± 169 persons when assuming 1000 occupants during working hours as the building does presently, the proposed forecasting approach significantly reduces overventilation and has the potential to improve HVAC energy performance.

Issues that need to be resolved include those for immediate consideration to make implementation feasible, as well as long-term considerations for similar applications, include:

- *How do the occupancy profiles translate to OA damper movements?* As mentioned, the relationship between the OA damper position and the OAF is not linear, and the OAF is almost never measured in buildings. As such, the relationship between the OA damper position and OAF must be examined in the building before the final ventilation profiles can be developed for each occupancy profile.
- *How can the proposed methodology be translated into the BAS?* Translating the workflow presented into a control language will require writing a new program for a portion of the AHU. There are barriers associated with this that include the limit on program size (e.g., a couple thousand bytes), BAS network bandwidth and the effect of added network traffic on the controller scan rates, and ensuring that any intervention does not interfere with the hierarchy of the control routines already implemented in the building. Additionally, any proposed intervention must be checked and approved by the building operators. Solutions to these barriers are explored in the following chapter.
- *What happens if the building occupancy patterns or electrical loads change dramatically?* A large change in occupancy, such as a change of use, would

negatively impact the performance of the proposed technique. As the representative profiles are trained offline, the models would need to be recalibrated and reimplemented in the BAS periodically to address model drift, or when significant changes in occupancy-type or large amounts of additional equipment are added. As we move towards the Internet of Things (IoT), the number of Wi-Fi enabled devices per person is likely to increase. New techniques to automatically establish the relationships between IoT devices and occupancy need to be explored, such as the method proposed by Park *et al.* [120].

- *What effects will an implementation have on building performance?* Building performance simulations of occupancy-based ventilation can be undertaken to evaluate the impact on energy use and indoor air quality [115, 121]. However, simulating an existing large commercial building adds complexity beyond the traditional shoebox models used to evaluate these techniques. Alternatively, the magnitude of these savings can be demonstrated directly via implementation [122]. The additional savings from modulating the damper to even lower levels is a worthy research question, as more precise controls will expand the complexity of the control algorithm needed and has the potential to offer diminishing returns.

The methods used for this forecasting technique were selected to make implementation within a real-world AHU plausible with available technology. The next chapter will explore translating this method into the GCL+ [123] control language for implementation within the case study building's BAS. Long term comfort and energy performance of the control algorithm will be investigated after implementation.

Chapter 4

This chapter has been submitted for publication as:

Occupancy-based predictive control of an outdoor air intake damper: A case study.

Hobson, B.W.; Gunay, H.B.; Ashouri, A.; and Newsham, G.R. *Proc. eSim 2020 Conf.*

4. Air handling unit implementation

4.1. Introduction

As many modern workplaces begin to further accelerate towards trends such as hotdesking and flexible workspaces [116], commercial and institutional buildings are hosting a smaller and smaller fraction of their maximum occupancy on a typical workday (e.g., Barbour *et al.* [62]). Despite this, most system-level components of heating, ventilation, and air conditioning (HVAC) systems still operate on static pre-set schedules that assume the building is almost fully occupied during working hours (e.g., ASHRAE Guideline 36-2018 [2]), resulting in chronic overventilation. Considering that indoor climate control systems of these buildings account for up to 50% of the buildings' energy use in Canada [1], there is a growing opportunity to reduce energy use and greenhouse gas emissions by providing HVAC services at an appropriate level for the actual occupancy of the building. This is traditionally attempted with demand control ventilation (DCV), which depends on CO₂ sensors to control a variable air volume (VAV) system by increasing the airflow rate to zones with higher CO₂ levels and therefore higher occupancy. However, the accuracy of commercial-grade CO₂ sensors can often be poor [124]. A study by Fisk *et al.* [125] found commercial CO₂ sensors frequently had measurement errors in excess of 20%. Additionally, the lag time between occupant arrival and the build-up of CO₂ has been identified as a weakness to this approach, especially in large spaces with highly variant occupancy profiles [121]. If improvements are to be made to the operation of system level equipment, such as air handling units (AHUs), the actual occupancy of the space must be predicted beforehand to ensure adequate indoor air quality (IAQ) is maintained. Occupancy-based control of the outdoor air damper in AHUs has been shown in simulation

to realize significant energy savings [63]. Reducing the outdoor air fraction in climates with lengthy heating and/or cooling seasons could reduce the temperature differential between the mixed air and the desired supply air setpoint, lowering the heating and cooling loads.

In addition to occupancy forecasting and occupant-count estimation, an emerging research challenge is to implement occupancy-centric control algorithms in commercially available building automation systems (BASs) at the system level [109] due to the implementation challenges associated with the limited computational power of these systems. Additionally, while data sources like Wi-Fi have been identified as having strong correlations with occupant count (e.g., [50, 126–128]), the integration of these data streams into BASs is not prevalent due to technological limitations, and privacy and security concerns. However, a ubiquitous data stream within BASs is electrical loads [76] which – while having a weaker correlation with occupant count than Wi-Fi [72] – follow high-level trends with occupancy levels throughout the day [78]. Therefore, combining these two disparate data sources may allow for occupancy forecasting to be implemented within a commercial BAS without additional sensing or control infrastructure.

This chapter explores the implementation of an occupancy forecasting technique to control the outdoor air damper within a case study building in Ottawa, Canada. Occupant count estimates from Wi-Fi enabled device count data were used to identify daily occupancy patterns [72]. These patterns were compared to patterns in building-level plug-in equipment and lighting loads to determine representative electrical load profiles for high, medium, low, and very low occupancy days. A proposed rules-based occupancy forecasting approach based on these representative profiles [129] was implemented as a

program in the BAS. Forecast accuracy, energy savings, IAQ impacts, challenges and relevant anecdotes from the implementation are examined and discussed.

4.2. Methodology

Representative occupancy profiles from Wi-Fi data (isolated from the BAS) were used to calculate damper profiles. Representative electrical load profiles (available in the BAS) were used as a proxy for the occupancy level, and a classification tree was used to predict day-ahead occupancy levels. To overcome the limited computational power of the building controllers, these models were trained offline using the programming language R [54]. The resulting values and Boolean logic from this process were translated into the building's BACnet controls programming language, GCL+ [123], and implemented in the case study building. It is worth noting that this program only effects the damper position when the dampers are in their minimum position mode. If another operating or safety mode is enabled, the program rightfully loses authority over the outdoor air dampers.

4.2.1. Damper profiles

The case study building is a 6650 m² academic office building located in Ottawa, Canada. It is served by two AHUs (AHU1 and AHU2) with a combined outdoor airflow rate of 10,000 L/s at minimum damper positions of 40% and 30%, respectively. The outdoor air dampers are both opposed blade dampers. Assuming the common rule-of-thumb of 10 L/s·person for combined outdoor airflow rates from ASHRAE 62.1-2019 [110], the building is providing enough outdoor air for 1000 occupants during occupied hours.

The first chapter [72] explored different data streams for occupancy-count estimation to determine the actual occupancy of the case study building. Wi-Fi enabled device counts were collected over a seven-month study and the trends in occupancy were analyzed. Using

k-means clustering, it was found that four distinct daily occupancy profiles were commonly observed in the case study building, denoted ‘h’, ‘m’, ‘l’, and ‘L’ for high, medium, low, and very low occupancy, respectively [129], summarized in Figure 4.1 (recall Figure 3.4c). These profiles represent the peak occupancy expected at each timestep for a day that falls within each respective cluster. The overall peak occupancy rate of the building rarely exceeded 50%, while the average peak occupancy rate was between 17% and 25%. Recall the study by Barbour *et al.* [62], which found a similar peak occupancy rate of 20% in their case study buildings.

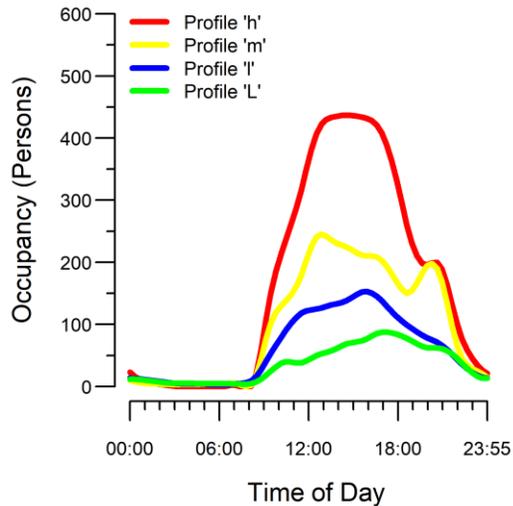


Figure 4.1: Occupancy profiles developed for the case study building.

Area (R_a) and person (R_p) outdoor air rate requirements set forth in ASHRAE 62-1-2019 [110] were used to calculate minimum breathing zone air rates (V_{bz}) and thus the outdoor air fraction for each of the occupancy profiles. As-built drawings of the case study building were used to determine the area fraction of the outdoor air requirements (Table 4.1), which was approximately 3000 L/s of the 10,000 L/s provided at the current minimum outdoor air dampers’ positions. The remaining 7000 L/s accounts for the person outdoor airflow rate requirements for 1000 persons, which can be increased or reduced depending on the

occupancy of the building (e.g., if the peak occupancy of the building is actually 500 persons, the breathing zone air rate can be reduced from 10,000 L/s to the 3000 L/s area outdoor air rate plus half of the 7000 L/s person outdoor air rate, or 6500 L/s total).

Table 4.1: Area outdoor air rate for the case study building.

| Occupancy Category | Area (m²) | Area Outdoor Air Rate, R_a (L/s) |
|-------------------------------|-----------------------------|---|
| Lecture classroom | 472 | 142 |
| University/college laboratory | 700 | 630 |
| Computer lab | 670 | 744 |
| Cafeteria/fast-food dining | 74 | 67 |
| Corridor | 1437 | 431 |
| Lobby/prefunction | 150 | 45 |
| Office space | 1486 | 446 |
| Storage/utility | 1092 | 327 |
| Total | 6650 | 2831 |

When in economizer mode (i.e., when the outdoor air fraction is at or near 100%), a one-to-one linear relationship between the outdoor air damper position and outdoor air fraction is used in the case study building. However, the relationship between the outdoor air fraction and the outdoor air damper position is not necessarily linear, especially at lower airflow rates. The ASHRAE Handbook of Fundamentals [130] specifies a family of curves for installed flow characteristics of different dampers with different control authority based on experimental data from the Air Movement and Controls Association [119]. Felker and Felker [131] note the actual installed characteristics of dampers typically fall in between

the 1% and 10% damper authority (DA_1 and DA_{10}) curves, see Figure 4.2. The DA_{10} curve was used in this study as it is a conservative estimation of the amount of airflow that will be achieved at each damper position.

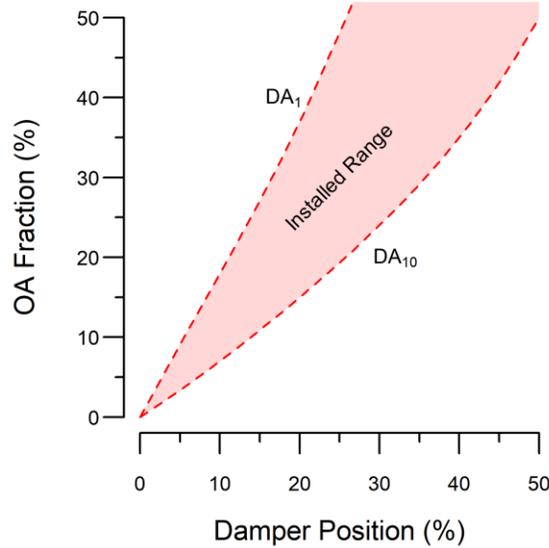


Figure 4.2: Damper position and air fraction for opposed blade dampers.

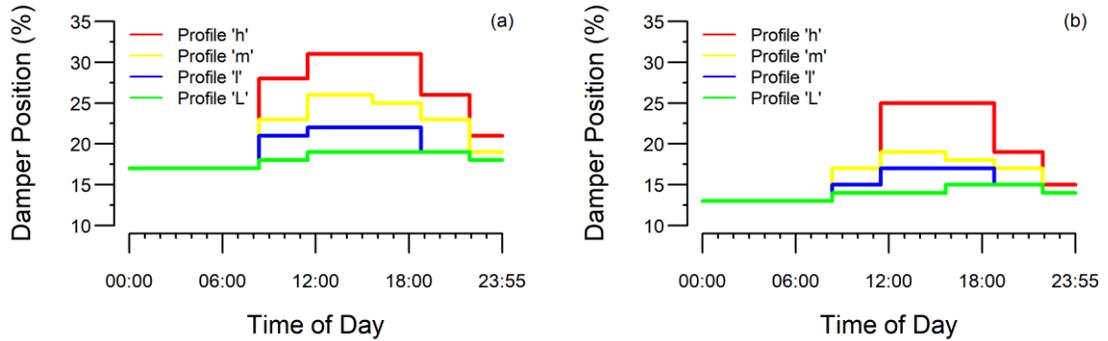


Figure 4.3: Damper position profiles for (a) AHU1 and (b) AHU2.

The outdoor air fraction for each occupancy profile was then translated into its respective damper position profile using this relationship, see Figure 4.3. The new minimum damper position is between 17% and 31% from zero to maximum observed occupancy for AHU1 and 13% and 25% for AHU2. Recall the previous minimum damper positions were 40% and 30%, respectively, regardless of occupancy levels. To reduce wear on the actuators

and network traffic on the BAS, the profiles were discretized into multi-hour increments, and the maximum damper position for each increment was used for the entire increment period. Higher occupancy increments were extended by an hour before and after as a conservative form of pre-conditioning.

4.2.2. Occupancy forecast

A forecast was developed to determine when each damper position profile should be used. Plug-in equipment and lighting load data available within the case study building's BAS were examined at the same time as the previous chapter [129]. Mean daily electrical load profiles were developed from the days clustered under each occupancy profile, see Figure 4.4 (recall Figure 3.4d). Each day of the study was assigned to an appropriate letter based on which daily profile had the lowest sum of square errors (SSE) compared to the BAS metered load.

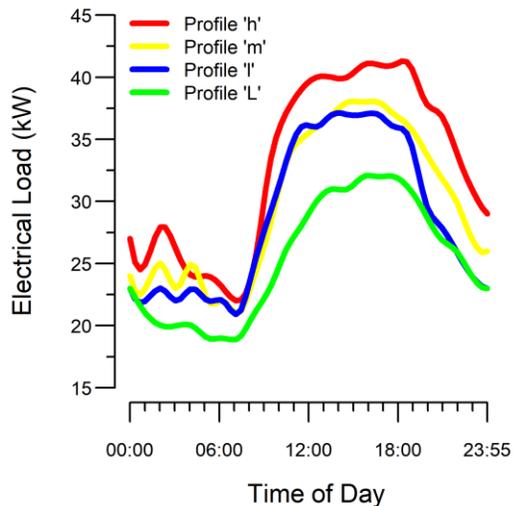


Figure 4.4: Plug-in equipment and lighting load profiles developed for the case study building.

Motif and discord identification were used on the string of letters to determine commonly repeated weekly patterns in these electrical load data as a proxy to occupancy. The letters

for these weekly patterns were analyzed to develop a classification tree following the same methodology of the previous chapter, recall Figure 3.5. Using this forecasting technique on the previously recorded data showed that days were successfully classified 70.4% of the time with an occupancy estimation error of 47 ± 69 persons at 95% confidence. Notably, the forecast was heavily biased towards over-classification of occupancy due to conservative assumptions made in the development of the occupancy and electrical load profiles.

4.2.3. Implementation

The electrical load profiles for each occupancy profile were hardcoded into a matrix in a GCL+ program. Electrical load data were recorded within the program at hourly intervals. The sum of the square errors (SSE) between the recorded electrical data and each column of the matrix were computed. The index of the column with the lowest SSE was taken as the representative profile for that day. Due to limited program size (i.e., 10,240 bytes) the number of timesteps that could be used to calculate the SSE was limited. To determine the number of timesteps and which timesteps to be used, the previously recorded data were analyzed using a forward stepwise approach. This revealed that the electrical loads at 1 pm, 5 pm, and 7 pm contributed significantly to forecast accuracy, with a successful classification rate of 67.7% using only the three timesteps compared to the 70.4% achievable when all 24 hourly timesteps were used, see Figure 4.5. This reduced the dimensions of the electrical load profile matrix to four by three (i.e., four profiles with three values) while simultaneously decreasing the polling rate for recording the electrical loads to only three times per day.

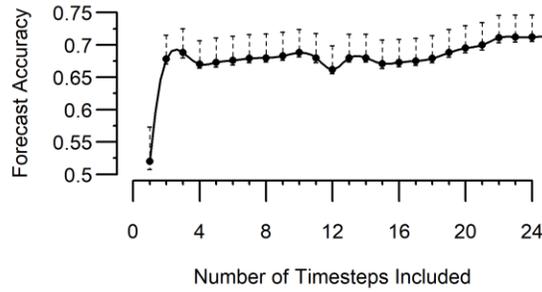


Figure 4.5: Forward stepwise selection of timesteps for GCL+ forecast.

Once the representative electrical load (and therefore occupancy) profile for the current day was determined, the index corresponding to the profile was recorded in a seven-character long vector. This vector was populated throughout the week and is the basis for the occupancy forecast. When the program is initialized, this vector is initially filled with dummy data that corresponds to five high ‘h’ occupancy weekdays and two low ‘l’ occupancy weekend days (i.e., a conservative weekly profile). A series of ‘IF’ statements developed from the classification tree (recall Figure 3.5) then examine the seven-character long vector once at 10 pm (i.e., once the electrical load data have been recorded and network traffic is low) and the corresponding damper profiles for the predicted occupancy level for the following day are selected. These damper positions are stored in a four by six matrix (i.e., the four profiles and damper positions for six discrete time increments illustrated in Figure 4.3). Note that, even if the program forecasts the incorrect occupancy for the next day, the program will still recalculate what the ‘measured’ occupancy was based on the electrical loads after they have been measured at the end of the day to reduce error propagation. One last time dependent ‘IF’ statement checks the time and specifies the correct row of the damper matrix at the appropriate time increment. The values from this matrix are then read into the damper control program. To see this program in its entirety, refer to Appendix A.

4.3. Results and discussion

Data for the case study building were collected from the third week of August until the beginning of February (i.e., 24-weeks) from the previous year (baseline year) and during the same period for the implementation year for comparison. These data include chiller electrical use data, AHU fan power data, as well as building level plug-in equipment and lighting loads and bulk metered electrical load data. Note that the outdoor air damper program was turned off in error from November 17th to December 20th during the implementation year. Operational data from this period are neglected from both the baseline and implementation year in the analysis. The building is served by a district heating system; steam flow and temperature data were available for the building’s heat exchanger. CO₂ sensors in 125 zones and temperature sensors in 138 zones, as well as AHU operational data, were also available. Wi-Fi enabled device counts were also collected for this study via IT services; no MAC addresses were monitored or stored over the course of this study as per the institution’s research ethics board.

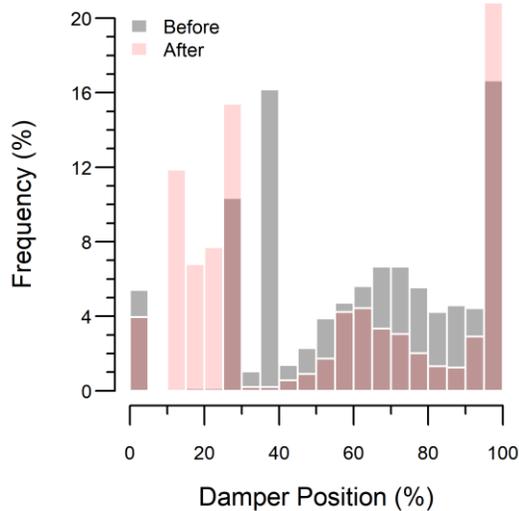


Figure 4.6: Outdoor air damper positions from the baseline (before) and implementation (after) period.

Accounting for the changes in climate between the baseline and implementation year is critical to ensure the appropriate levels of savings are attributed to the new outdoor air control program. Note that the implementation year had comparatively milder weather than the baseline year; the baseline monitoring period had 2454 heating degree days (HDD) and 157 cooling degree days (CDD), whereas the implementation year had 2131 HDD and 96 CDD during the 24-week study period. Switchover from heating to cooling occurred October 2nd and October 9th during the baseline and implementation year, respectively. Comparing the outdoor air damper positions from the baseline year and implementation year reflects the lower minimum damper positions, see Figure 4.6. While it may be expected that the building's AHUs spent more time in the minimum position mode during the colder baseline year, they actually spent 12.0% and 16.6% of hours collectively in minimum position mode during the baseline and implementation period, respectively. This was due to a soft fault in the economizer program of the AHUs that was discovered and altered during the implementation process. This fault was causing the units to go into the economizer mode when the return air temperature exceeded the outdoor air temperature, regardless if it was heating or cooling season. As such, the unit was often fully economizing at low outdoor air temperatures; this caused the supply air temperature to remain at the lower end of the temperature threshold (13°C), resulting in increased loads on the building's perimeter heaters to reach zone setpoints. On a particularly cold day during the implementation period when the program was erroneously turned off, the extreme volume of cold outdoor air caused a heating coil in one of the units to freeze. This fault was identified by facilities personnel who were able to manually reset the unit without incident, at which point the outdoor air damper program was reimplemented.

4.3.1. Forecast accuracy

After the implementation period, Wi-Fi enabled device count data were retroactively examined to determine a baseline for the occupancy forecast accuracy. The analysis showed that a handful of days exceeded the highest expected occupancy profile, with 12 total occupied hours spent above the high ‘h’ occupancy threshold over the entire 24-week study period. It was found 50, 22, 15, and 80 days during the study period could be classified as high ‘h’, medium ‘m’, low ‘l’, and very low ‘L’ occupancy, respectively, see Figure 4.7a. This was then compared to the electrical load forecast-based results from the implementation period. The program classified 101, 3, 13, and 50 days during the study period as high ‘h’, medium ‘m’, low ‘l’, and very low ‘L’ occupancy, respectively, see Figure 4.7b.

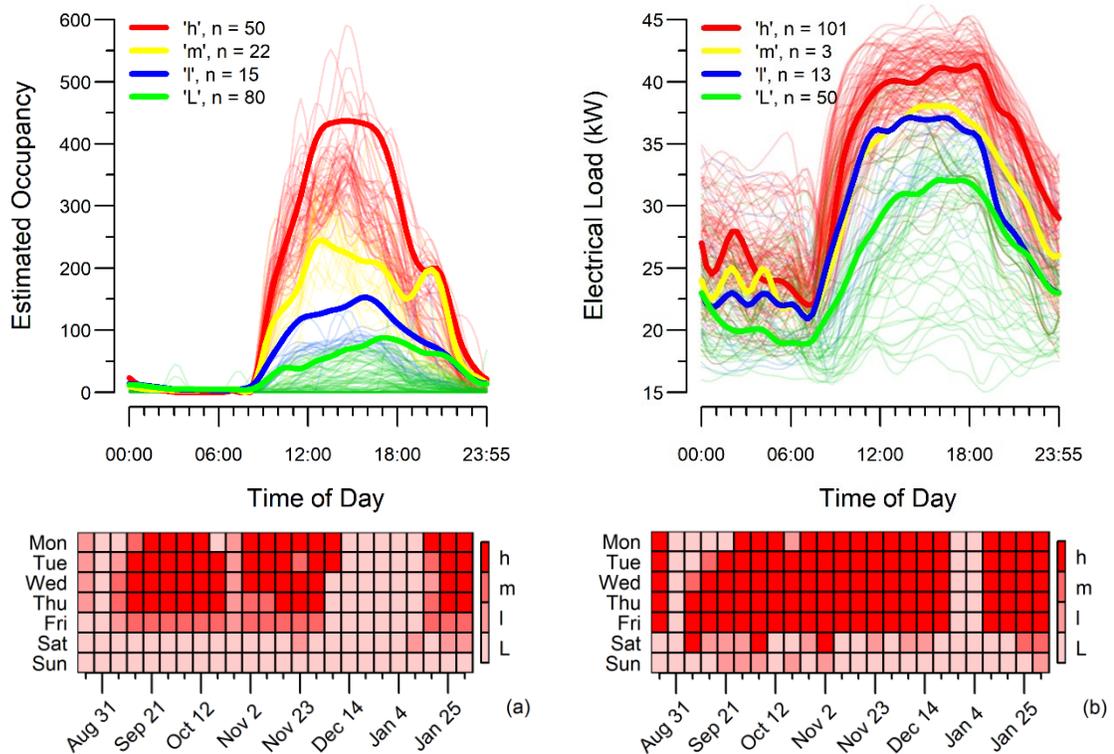


Figure 4.7: (a) Wi-Fi measured occupancy profiles and (b) electrical load-based forecasts from the implementation year.

The overall prediction accuracy for the forecast implemented was 58%. However, it should be noted that the forecasts were intended to be conservative; 97% of forecasted days had an observed occupancy that was at or below forecasted levels, see Table 4.2. Overall, the accuracy of the electrical load-based occupancy level forecast was lower than expected. This is in part due to the fact that the algorithm had to be initialized at the beginning of the study period and reinitialized during the week of December 20th as the outdoor air damper program was erroneously turned off; this led to the over-classification of ten days of low ‘l’ and very low ‘L’ occupancy days to high ‘h’ during the initialization process, reducing the apparent forecast accuracy. However, despite the highly conservative nature of the program, significant energy savings were still achieved, which may largely be attributed to the outright reduction in outdoor air fraction associated with even the highest new outdoor air damper profile.

Table 4.2: Confusion matrix for occupancy forecast.

| | | Wi-Fi measured | | | |
|---------------------|---|----------------|----|----|----|
| | | h | m | l | L |
| Electrical forecast | h | 48 | 20 | 11 | 22 |
| | m | 1 | 1 | 1 | 0 |
| | l | 0 | 0 | 1 | 12 |
| | L | 1 | 1 | 2 | 46 |

4.3.2. Energy savings

Energy use data for the implementation period were compared to energy use data from the previous year during the same period. Three-parameter univariate changepoint models [132] were used to account for the temperature differences between the two years. Note that the heating energy use from the steam loop data include all heating coils and perimeter heaters, but do not include any potential use of personal electric resistance heaters by individual occupants. Fitting the implementation year temperature data to the changepoint models for the baseline year (Figure 4.8) indicates that the building would have used an estimated 85 MWh_{eq} for cooling and 297 MWh_{eq} for heating if there was no implementation. Actual cooling and heating energy use during the implementation year were 76 MWh_{eq} and 183 MWh_{eq}, representing 10.3% and 38.4% savings, respectively. This change is reflected in the lower slopes of the changepoint models shown in Figure 4.9. Two distinct groupings of points emerge; the upper family of heating days is characteristic of the energy use on days following the high ‘h’ occupancy profile, whereas the lower family represents the heating days where low ‘l’ and very low ‘L’ occupancy was predicted. Medium ‘m’ occupancy days were not prevalent enough during the study period to attribute these days to any one changepoint model. Note that the fan energy profiles show a negligible decrease in fan energy consumption of 2.8% (from 71 MWh to 70 MWh using the baseline model and the implementation year itself, respectively). Additionally, while the CV(RMSE) of the baseline heating changepoint model is within the range recommended by ASHRAE Guideline 14-2014 [133], it should be noted that the cooling model falls outside of this range. Note that the changepoint models for the implementation year are for illustrative purposes only and are not used in the energy savings calculations.

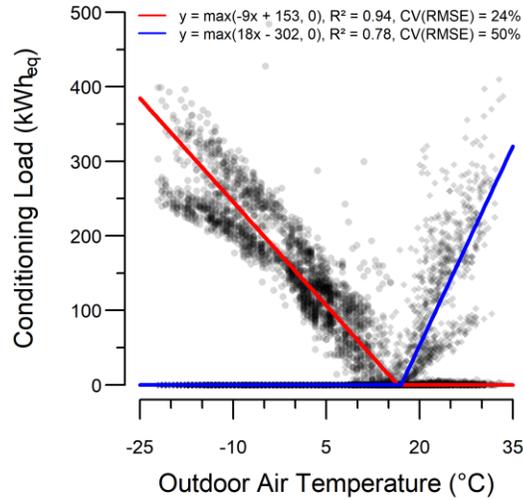


Figure 4.8: Changepoint model for baseline year.

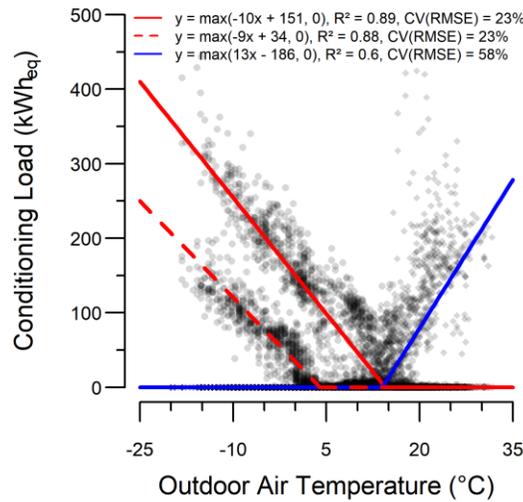


Figure 4.9: Changepoint model for implementation year.

4.3.3. Comfort impact

In addition to changes in energy use, changes in indoor air quality were explored. CO₂ concentrations from 125 zones were compared between the baseline year and implementation year, see Figure 4.10. A mean increase of 33 ppm was observed for occupied hours during the implementation year. The increase in hours where indoor CO₂ concentration exceeds atmospheric CO₂ concentrations by 700 ppm – the threshold set out in ASHRAE 62.1-2019 [110] – was from 0.06% to 0.23% of occupied hours from before

and after implementation, respectively. Further work needs to be done to determine how much of this increase is attributable to the outdoor air program versus drift in the CO₂ sensor accuracy over time.

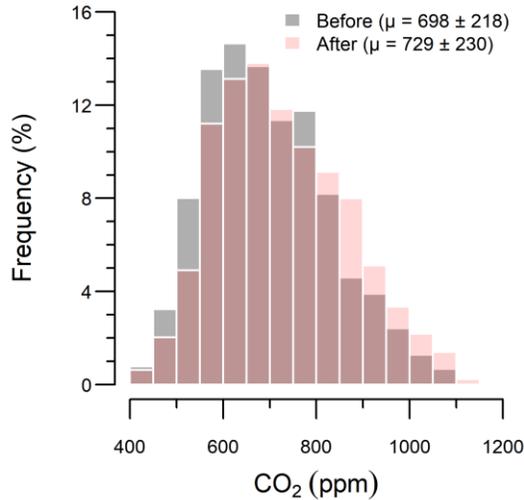


Figure 4.10: Changes in CO₂ concentration.

Changes in indoor air temperature for 138 zones were also examined, see Figure 4.11. While the mean zone air temperature did not change significantly before and after implementation, the time spent within $\pm 1^\circ\text{C}$ of the 22°C typical setpoint decreased from 73.8% to 70.0% (3.8% reduction) after implementation. However, this change in ability to meet setpoints may be partially attributable to the milder climate conditions during the implementation year, which led to a prolonged shoulder season (recall that switchover from heating to cooling occurred October 2nd and October 9th during the baseline and implementation year, respectively) and several seasonally warm days during the heating season. To explore this phenomenon, temperature data from 32 zones in another academic office building located on the same campus as the case study building were examined. It was similarly found that the mean zone air temperature did not change significantly,

however the number of occupied hours spent within $\pm 1^\circ\text{C}$ of the setpoint decreased from 65.9% to 63.5% (2.4% reduction) during the same study period.

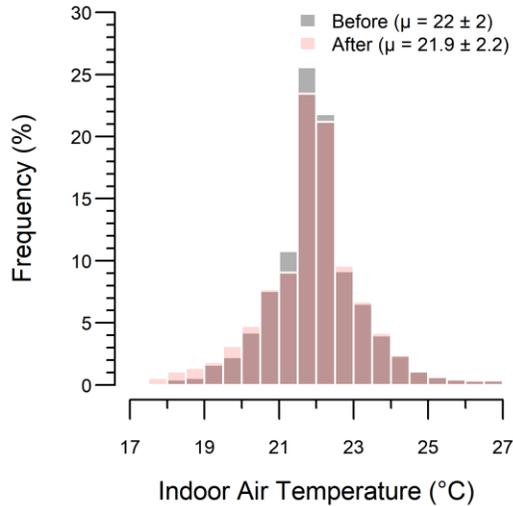


Figure 4.11: Changes in indoor air temperature.

Reducing the outdoor air fraction and increasing the return air fraction increased the mixed air temperature during the study period, reducing the heating coil output to meet temperatures within the supply air temperature setpoint range (13°C to 20°C). As a result, supply air temperatures during the implementation year were higher than the baseline year. The baseline year had a mean temperature of $16.1 \pm 4.1^\circ\text{C}$ with 39.1% of hours spent in the upper half of the setpoint range ($>16^\circ\text{C}$), compared to $16.5 \pm 4.7^\circ\text{C}$ with 50.9% of hours spent above 16°C during the implementation year. This significantly reduced the loads on the perimeter heaters and resulted in energy savings. A disadvantage of the implemented program was that higher return air (and hence higher supply air) temperatures would occur as the building's average room temperature increased, reducing the availability of cooling in problem zones and in some instances causing localized overheating, especially during very low 'L' occupancy days where the outdoor air temperature rose above 0°C . This problem could potentially be addressed by increasing the

airflow rate at the zone level in rooms where this issue was identified, or by introducing an override during days that match these criteria to ensure adequate outdoor airflow for cooling purposes.

4.4. Closing remarks

In this chapter, the rules-based approach for day-ahead prediction of occupancy-levels developed in the previous chapter was implemented in the air handling units of an academic office building in Ottawa, Canada. Daily electrical load profiles for each of the occupancy-levels along with rules from a classification tree that were previously trained offline were hardcoded into the BACnet controls programming language GCL+. Operational and IAQ data for concurrent 24-week periods from the previous year and implementation year were compared.

Key findings from this chapter are summarized as follows:

- It is possible to implement a rules-based occupancy-level forecast in commercial building controllers without additional sensing or controls hardware (in BASs comparable to the case study building).
- Three-parameter univariate changepoint models found an estimated savings of 38.4% and 10.3% for heating and cooling energy use, respectively, during the implementation year.
- The accuracy of the rules-based forecast as implemented was 58%. However, this program was conservative, classifying 97% of days at or above their actual occupancy yet still yielding considerable energy savings.
- The implemented program may be partially responsible for a 33 ppm increase in mean CO₂ concentrations and a 3.8% reduction in hours spent within $\pm 1^\circ\text{C}$ of

temperature setpoints (although the latter may be at least partially attributable to the milder climate and delayed switchover time during the implementation year).

While this chapter explored the benefits of implementing this control program, several questions and outstanding issues remain, including:

- *Exactly how much can ventilation rates be reduced to keep the benefits from higher ventilation on human comfort and IAQ?* While overly-excessive ventilation rates have little benefit to occupant comfort or satisfaction, well-ventilated buildings (e.g., above 10 L/s·person) have been associated with better human factor outcomes [134]. If ventilation can be reduced to levels that still provide IAQ benefits, significant savings may still be possible. For example, the minimum design ventilation rate in this case study was 13 L/s·person (6500 L/s for an assumed 50% maximum occupancy), which was sufficiently low to generate HVAC savings. Marginally increasing this ventilation rate may mitigate changes to the building's IAQ with a small cost to energy savings. The optimal balance between these two factors warrants further investigation.
- *How much of the savings predicted are attributable solely to the forecast implemented?* While the energy benefit of reducing the outdoor air fraction outright based on occupancy is apparent, the savings directly attributable to the forecasting method used should be further explored. As the variability of occupancy in a space increases, so does the energy savings potential [115]. While the case study building had highly variable occupancy levels, the occupancy patterns of an academic office building are relatively predictable and may not require such complex forecasting mechanisms where a schedule may be reliably used. Therefore, while the case study

building facilitated a useful methodological exercise in creating these rules-based forecasts, the benefits of such an approach should be thoroughly explored for other occupancy types and patterns.

This research highlights one possible use of occupancy-related data for data-driven operation of commercial and institutional buildings. While the approach adopted in this chapter is specific to the case study building, implementation of similar control schemes may be possible in other commercial and institutional buildings. Unique buildings' BASs age/computational capabilities, control infrastructure, or HVAC systems can potentially enhance or diminish the effectiveness of this occupancy-centric approach. In the same tune, there is much room to enhance the outlined approach in the case study building itself, although future optimization may be found to offer diminishing returns. This topic would be worthy of future research efforts as academic and commercial experience with building-level field implementations of occupancy-centric controls matures.

5. Conclusions

5.1. Summary

This research project aimed to develop a novel day-ahead occupancy-level prediction program that could be used with existing commercial BAS technology and data streams for the purpose of modulating air handling units' outdoor air dampers; this program was implemented in an academic office case study building in Ottawa, Canada. A summary of the conclusions presented in each chapter are outlined below.

5.1.1. Opportunistic occupant-count estimation

This chapter explored sensor fusion (i.e., the combination of several datatypes to enhance model accuracy beyond that achievable with a single datatype) of several datatypes available in commercial buildings to train models – namely multiple linear regression models and artificial neural networks – for occupant-count estimates. Eight-days' worth of ground truth occupancy data were collected over four floors of the case study building. Wi-Fi enabled device counts, PIR trigger events, CO₂ concentrations, plug-in equipment and lighting load data were recorded at concurrent five-minute timesteps from each floor for the study period. 2808 possible combinations of these five data streams across all four floors were examined to determine the utility of each data stream and model formalism. The main findings of this chapter were as follows:

- Artificial neural networks produce occupant-count estimates with slightly higher accuracy than multiple linear regression models. However, multiple linear regression models offer better performance when transferred from one space to

another, which largely negates any benefits from the added complexity of artificial neural networks.

- Increasing the number of data streams used in each model did not necessarily equate to higher accuracy, especially on models that have been trained using sensors that are highly impacted by the location and density of the sensor grid (i.e., PIR, CO₂).
- Wi-Fi enabled device count data offered the single largest contribution to model accuracy for both formalisms, whereas PIR sensors had the lowest utility.
- Multiple linear regression models that include Wi-Fi enabled device count data as one of the regressors produced occupant-count estimates with high accuracy (i.e., mean R² of 74.1% up to R² of 96.1%).
- Aggregated plug-in equipment and lighting load data had higher utility than considering these two loads separately, with modest accuracy when paired with Wi-Fi enabled device count data in a multiple linear regression model (i.e., mean R² of 72.1%) for occupant-count estimation.

Therefore, the findings suggest that linear regression is sufficient for occupant-count estimation for the purposed purpose. The Wi-Fi enabled device count and plug-in equipment/lighting load data streams showed promise for further practical application.

5.1.2. Rules-based occupancy prediction

Building on the findings of the previous chapter, and with consideration for the limited computational power and access to data within the BAS, a rules-based occupancy forecasting technique was developed. Seven-months of Wi-Fi enabled device count, plug-in equipment and lighting load data were collected at concurrent 15-minute timesteps in the case study building. Occupant-count estimates for the seven-month period were

extrapolated using the models explored in the previous chapter. Different clustering techniques were explored to determine the optimal technique for the proposed purpose. Clustering was used to find characteristic daily occupancy profiles over the study period. Corresponding representative electrical load profiles for these characteristic days were extracted (as these data are available in the BAS, whereas Wi-Fi data are currently not). Patterns in the occurrences of these characteristic days were explored using motif and discord identification. A classification tree was employed on these data to produce a set of rules-based Boolean logic for predicting occupancy-levels one day in advance. The main findings of this chapter were as follows:

- The maximum estimated occupancy in the case study building was 515 ± 54 persons, or approximately 46 to 57% of the 1000 occupant assumption used for traditional operations. The average peak occupancy during workdays was between 189 ± 20 and 226 ± 24 persons, or approximately 17 to 25% of the assumed occupancy.
- Agglomerative hierarchical clustering using correlation dissimilarity was a poorly suited clustering technique for the proposed application. Conversely, k-means clustering was selected as the preferred clustering technique.
- k-means clustering of electrical load data showed occupancy-levels on 84.5% of days could be correctly classified using plug-in equipment and lighting load in place of Wi-Fi enabled device count.
- A pruned classification tree trained using motifs and discords from these data was able to correctly predict the occupancy-level of the building based on the electrical load for 70.4% of days.

Therefore, this chapter showed a rules-based occupancy prediction could be developed and trained offline using k-means clustering, motif and discord identification, and a classification tree.

5.1.3. Air handling unit implementation

Occupancy profiles from the previous chapter were translated into a set of corresponding outdoor air damper position profiles. These profiles, the electrical load profiles needed for forecasting, and the rules-based occupancy predictions developed in the previous chapter were implemented in the BAS to control the case study buildings outdoor air dampers over a 24-week study period. Wi-Fi enabled device counts were recorded for this period to retroactively examine the implemented programs prediction accuracy. Zone-level CO₂ concentrations, PIR activations, and temperatures were recorded at concurrent 15-minute timesteps to assess the impact of indoor air quality. Operational data including chiller electricity use, steam mass flow rate and temperatures at the district heating loop, as well as AHU fan energy use were recorded at concurrent one-hour timesteps. These data were also extracted for the same timeframe during the previous year of operation to serve as a baseline. Univariate changepoint models were employed to evaluate the energy savings relative to changes in the weather between the implementation and baseline year. The main findings of this chapter were as follows:

- The implemented program reduced chiller electricity use from an estimated 85 MWh to 76 MWh (10.3%), and reduced heating loads from an estimated 297 MWh_{eq} to 183 MWh_{eq} (38.4%), with a negligible reduction in fan energy use.
- The implemented occupancy prediction algorithm had 58% prediction accuracy due in part to simplifications that had to be made for practical application; however,

the predictions were highly conservative with 97% of days either correctly or overclassified.

- Mean zone-level CO₂ concentrations increased by approximately 33 ppm during occupied hours for the implementation year, increasing the hours outside of ASHRAE 62.1-2019 requirements by 0.17% (from 0.06% to 0.23%).
- The number of hours spent within $\pm 1^{\circ}\text{C}$ of the zone temperature setpoints was reduced by 3.8% during the implementation year, although this may be partially attributable to climate.

Therefore, this chapter showed that even with high conservative occupancy predictions, significant energy savings could be achieved by modulating the outdoor air dampers in the case study building based on occupancy while in the minimum position mode. Tweaks to zone level parameters may help address any IAQ concerns raised during the initial implementation period.

5.2. Research contributions

5.2.1. Occupant-count estimation

The analysis presented in this chapter contributes to a growing body of literature on implicit occupancy sensing and the use of sensor fusion. This particular study focused explicitly on the use of commercially available and in-situ sensors, which is lacking in the literature [14, 20]. The data set from this study, including ground truth data, were published alongside a journal article in *Building and Environment* as “Opportunistic occupancy-count estimation using sensor fusion: A case study” [72] to allow other researchers to test and validate their occupant-count estimation functions. The results were presented at the Second Experts’ Meeting of IEA-EBC Annex 79 and the Fourth International Symposium on Occupant

Behaviour (OB-19) in San Antonio, Texas. In addition, the findings of this study have been shared with our collaborating industry partners.

5.2.2. Rules-based occupancy prediction

A novel forecasting technique formulated for the express purpose of implementing occupant-centric controls in commercial building controllers was presented in this chapter. The novel use of load profiling approaches for characterizing occupancy, and the use of clustering techniques for binning motifs and discords, contribute to the body of literature surrounding occupant-centric controls. These findings were first shared in *ASHRAE Transactions* as “Wi-Fi Based Occupancy Clustering and Motif Identification : A Case Study” [129] and presented at ASHRAE’s Winter Conference in Orlando, Florida. Further findings on this topic are under review for publication in *Energy and Buildings* as “Clustering and motif identification for occupancy-centric control of an air handling unit.” In addition, the findings of this study have been shared with our collaborating industry partners.

5.2.3. Air handling unit implementation

As noted in the literature review conducted by Park *et al.* [109], field implementations of system-level occupancy-centric controls are non-existent. This case study is therefore an early example that highlights the energy-savings potential and unique challenges associated with this scale of implementation. As such, the data and findings from this case study have been contributed to IEA-EBC Annex 79 as part of the case studies of field implementations of occupancy/occupant-centric controls being examined by Subtask 4 in their ongoing research efforts. These findings have been accepted for publication at IBPSA-Canada’s eSim 2020 Conference as “Occupancy-based predictive control of an

outdoor air intake damper: A case study” [135]. In addition, the findings of this study have been shared with our collaborating industry partners who have expressed their intention to integrate elements of the occupancy forecasting algorithms into their commercial energy management platform, as well as Facilities Management and Planning at Carleton University.

5.3. Recommendations for future work

Over the course of these works, several research topics arose that warrant further study, including:

- Wi-Fi shows great promise for occupant-count estimation purposes. However, there is a gap between facility managers and IT services when it comes to integrating this information into the BAS. As modern buildings have their BAS and Wi-Fi networks on the same local area network (LAN), developing affordable BAS-integrated Wi-Fi could potentially address some of the shortcomings of the proposed forecasting framework by making time series forecasts possible directly on building automation controllers, eliminating the need for rules-based forecasting. Significant progress must be made regarding the security and privacy concerns related to BAS-integrated Wi-Fi to move in this direction. In older non-LAN based building automation systems, or in systems with significantly less capable building automation systems, this may not be possible at all without significant investments in modern BAS infrastructure. The feasibility of similar occupant-centric control schemes should be explored in these buildings.
- Regarding the use of Wi-Fi enabled device count, there are also considerations that need to be made for the expected prevalence of Wi-Fi enabled personal mobile

devices across different occupancy-types, cultures, socio-economic backgrounds, and over time. These differences are expected to have a great impact on the usefulness of Wi-Fi for occupant-count estimation depending on these contextual factors. Recent work by Park *et al.* [120] that borrows from the field of ecology presented a capture and recapture (CRC) based method for sampling Wi-Fi and Bluetooth enabled device populations in buildings to help address this shortcoming. The benefits of a practical application of this theory in commercial buildings warrants further investigation.

- Plug-in equipment and lighting load data submetered at the floor level were used to develop representative electrical load profiles that excluded HVAC-related loads. The use of building-level (i.e., bulk metered) electrical load data in the proposed methodology should be explored, as this information is more commonly available across the building stock. Alternatively, the pairing of this technique with load disaggregation could potentially highlight the synergies between these two areas of study.
- The data used in this research were taken at five to fifteen-minute timesteps, which may be higher resolution than many traditional BAS systems are capable of logging. The effect of changing these parameters to a more representative timestep used in industry would be a worthy research question.
- A large change in occupancy, such as a change of use, would negatively impact the performance of the proposed technique. As the representative profiles are trained offline, the models would need to be recalibrated and reimplemented in the BAS periodically to address model drift, or when significant changes in occupancy-type

or large amounts of additional equipment are added. As we move towards the Internet of Things (IoT), the number of Wi-Fi enabled devices per person is likely to increase. New techniques to automatically establish the relationships between IoT devices and occupancy need to be explored, such as the method proposed by Park *et al.* [120].

- While overly-excessive ventilation rates have little benefit to occupant comfort or satisfaction, well-ventilated buildings (e.g., above 10 L/s·person) have been associated with better human factor outcomes [134]. If ventilation can be reduced to levels that still provide IAQ benefits, significant savings may still be possible. For example, the minimum design ventilation rate in this case study was 13 L/s·person (6500 L/s for an assumed 50% maximum occupancy), which was sufficiently low to generate HVAC savings. Marginally increasing this ventilation rate may mitigate changes to the building's IAQ with a small cost to energy savings. The optimal balance between these two factors warrants further investigation.
- The compatibility of the developed forecasting technique with common sequences of operation should be explored in any building before implementation. As discovered over the course of this study, operational faults in HVAC equipment can diminish or eliminate any savings potential these occupancy-centric controls offer (e.g., faults that cause the air handling unit to run excessively in the economizer mode). Therefore, fault detection and diagnostic techniques must be developed in hand with occupancy-centric controls if their full potential to be realized.

It should be noted that the findings of this research are explicitly related to the case study building. While the energy benefit of reducing the outdoor air fraction outright based on

occupancy is apparent, the savings directly attributable to the forecasting method used should be further explored. While the case study building had highly variable occupancy levels, the occupancy patterns of an academic office building are relatively predictable and may not require such complex forecasting mechanisms where a schedule may be reliably used. Therefore, while the case study building facilitated a useful methodological exercise in creating these rules-based forecasts, the benefits of such an approach should be thoroughly explored for other occupancy types and patterns.

Appendix A

Code and data relevant to this thesis can be found at:

<https://github.com/HobsonB/MASc-Thesis.git>

References

- [1] Natural Resources Canada, *Survey of Commercial and Institutional Energy Use - Buildings 2009 - Detailed Statistical Report 2012*. 2012.
- [2] ASHRAE, “ASHRAE Guideline 36: High-Performance Sequences of Operation for HVAC Systems,” vol. 8400, 2018.
- [3] T. Abuimara, W. O’Brien, H. B. Gunay, H. Burpee, and J. K. Day, “Designing for Occupants : A Review of the Integrated Design Practice,” in *ACEEE Summer Study on Energy Efficiency in Buildings*, 2018, vol. 66, no. August, pp. 2008–2013.
- [4] T. Abuimara *et al.*, “Modelling Occupants in Buildings: Stakeholders’ Workshop on Current Barriers, Challenges and Needs,” in *IBPSA-Canada eSim 2018*, 2018, no. July, pp. 312–320.
- [5] T. Abuimara, W. O’Brien, H. B. Gunay, and J. S. Carrizo, “How Assumptions About Occupants Can Misinform Building Design,” *ASHRAE J.*, vol. 62, no. 1, pp. 14–18, 2020.
- [6] L. Golden, “Flexible work Time: Correlates and consequences of work scheduling,” *American Behavioral Scientist*, vol. 44, no. 7. pp. 1157–1178, 2001.
- [7] K. W. Roth, D. Westphalen, P. Llana, and M. Feng, “The Energy Impact of Faults in US Commercial Buildings,” *Int. Refrig. Air Cond. Conf.*, pp. 600–609, 2004.
- [8] S. Katipamula and M. R. Brambley, “Methods for Fault Detection, Diagnostics , and Prognostics for Building Systems — A Review , Part I,” *HVAC&R Res.*, vol. 11, no. 1, pp. 3–25, 2005.
- [9] W. Shen, G. R. Newsham, and H. B. Gunay, “Leveraging existing occupancy-related data for optimal control of commercial office buildings: A review,” *Adv. Eng. Informatics*, vol. 33, pp. 230–242, 2017.
- [10] N. Cao, J. Ting, S. Sen, and A. Raychowdhury, “Smart sensing for HVAC Control: Collaborative intelligence in optical and IR cameras,” *IEEE Trans. Ind. Electron.*, vol. 65, no. 12, pp. 9785–9794, 2018.
- [11] M. Jin, N. Bekiaris-Liberis, K. Weekly, C. J. Spanos, and A. M. Bayen, “Occupancy Detection via Environmental Sensing,” *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no.

- 2, pp. 443–455, 2018.
- [12] O. Ardakanian, A. Bhattacharya, and D. Culler, “Non-Intrusive Techniques for Establishing Occupancy Related Energy Savings in Commercial Buildings,” in *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*, 2016.
 - [13] R. H. Dodier, G. P. Henze, D. K. Tiller, and X. Guo, “Building occupancy detection through sensor belief networks,” *Energy Build.*, vol. 38, no. 9, pp. 1033–1043, 2006.
 - [14] R. Melfi, B. Rosenblum, B. Nordman, and K. Christensen, “Measuring building occupancy using existing network infrastructure,” *2011 Int. Green Comput. Conf. Work. IGCC 2011*, 2011.
 - [15] S. Rosiek and F. J. Batlles, “Reducing a solar-assisted air-conditioning system’s energy consumption by applying real-time occupancy sensors and chilled water storage tanks throughout the summer: A case study,” *Energy Convers. Manag.*, vol. 76, pp. 1029–1042, 2013.
 - [16] F. Wahl, M. Milenkovic, and O. Amft, “A distributed PIR-based approach for estimating people count in office environments,” *Proc. - 15th IEEE Int. Conf. Comput. Sci. Eng. CSE 2012 10th IEEE/IFIP Int. Conf. Embed. Ubiquitous Comput. EUC 2012*, pp. 640–647, 2012.
 - [17] P. Liu, S.-K. Nguang, and A. Partridge, “Occupancy Inference Using Pyroelectric Infrared Sensors Through Hidden Markov Models,” *IEEE Sens. J.*, vol. 16, no. 4, pp. 1062–1069, 2016.
 - [18] C. Duarte, K. Van Den Wymelenberg, and C. Rieger, “Revealing occupancy patterns in an office building through the use of occupancy sensor data,” *Energy Build.*, vol. 67, pp. 587–595, 2013.
 - [19] Z. Nagy, F. Y. Yong, M. Frei, and A. Schlueter, “Occupant centered lighting control for comfort and energy efficient building operation,” *Energy Build.*, vol. 94, pp. 100–108, 2015.
 - [20] S. Gilani and W. O’Brien, “Review of current methods, opportunities, and challenges for in-situ monitoring to support occupant modelling in office spaces,” *J. Build. Perform. Simul.*, vol. 10, no. 5–6, pp. 444–470, 2017.

- [21] A. Wagner, W. O'Brien, and B. Dong, *Exploring Occupant Behavior in Buildings*. 2018.
- [22] ASHRAE, "ANSI/ASHRAE/IES 90.1-2019: Energy Standard for Buildings Except Low-Rise Residential Buildings," Atlanta, GA, 2019.
- [23] M. S. Zuraimi, A. Pantazaras, K. A. Chaturvedi, J. Yang, K. W. Tham, and S. E. Lee, "Predicting occupancy counts using physical and statistical Co2-based modeling methodologies," *Build. Environ.*, vol. 123, pp. 517–528, 2017.
- [24] J. Wang, N. C. F. Tse, T. Y. Poon, and J. Y. C. Chan, "A practical multi-sensor cooling demand estimation approach based on visual, indoor and outdoor information sensing," *Sensors (Switzerland)*, vol. 18, no. 11, 2018.
- [25] M. Aftab, C. Chen, C. K. Chau, and T. Rahwan, "Automatic HVAC control with real-time occupancy recognition and simulation-guided model predictive control in low-cost embedded system," *Energy Build.*, vol. 154, pp. 141–156, 2017.
- [26] J. P. Roselyn, R. A. Uthra, A. Raj, D. Devaraj, P. Bharadwaj, and S. V. D. Krishna Kaki, "Development and implementation of novel sensor fusion algorithm for occupancy detection and automation in energy efficient buildings," *Sustain. Cities Soc.*, vol. 44, pp. 85–98, 2019.
- [27] S. Petersen, T. H. Pedersen, K. U. Nielsen, and M. D. Knudsen, "Establishing an image-based ground truth for validation of sensor data-based room occupancy detection," *Energy Build.*, vol. 130, pp. 787–793, 2016.
- [28] G. R. Newsham *et al.*, "Testing the accuracy of low-cost data streams for determining single-person office occupancy and their use for energy reduction of building services," *Energy Build.*, vol. 135, pp. 137–147, 2017.
- [29] Y. Benezeth, H. Laurent, B. Emile, and C. Rosenberger, "Towards a sensor for detecting human presence and characterizing activity," *Energy Build.*, vol. 43, no. 2–3, pp. 305–314, 2011.
- [30] P. Carreira, A. A. Costa, V. Mansur, and A. Arsénio, "Can HVAC really learn from users? A simulation-based study on the effectiveness of voting for comfort and energy use optimization," *Sustain. Cities Soc.*, vol. 41, pp. 275–285, 2018.
- [31] N. Li, G. Calis, and B. Becerik-Gerber, "Measuring and monitoring occupancy with

- an RFID based system for demand-driven HVAC operations,” *Autom. Constr.*, vol. 24, pp. 89–99, 2012.
- [32] A. Akbar, M. Nati, F. Carrez, and K. Moessner, “Contextual occupancy detection for smart office by pattern recognition of electricity consumption data,” *IEEE Int. Conf. Commun.*, vol. 2015-Sept, no. June, pp. 561–566, 2015.
- [33] T. Labeodan, K. Aduda, W. Zeiler, and F. Hoving, “Experimental evaluation of the performance of chair sensors in an office space for occupancy detection and occupancy-driven control,” *Energy Build.*, vol. 111, pp. 195–206, 2016.
- [34] T. Ekwevugbe, N. Brown, and D. Fan, “A design model for building occupancy detection using sensor fusion,” in *IEEE International Conference on Digital Ecosystems and Technologies*, 2012, pp. 1–6.
- [35] Z. Chen, M. K. Masood, and Y. C. Soh, “A fusion framework for occupancy estimation in office buildings based on environmental sensor data,” *Energy Build.*, vol. 133, pp. 790–798, 2016.
- [36] Z. Yang, B. Becerik-Gerber, N. Li, and M. Orosz, “A systematic approach to occupancy modeling in ambient sensor-rich buildings,” *Simulation*, vol. 90, no. 8, pp. 960–977, 2014.
- [37] B. Dong *et al.*, “An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network,” *Energy Build.*, vol. 42, no. 7, pp. 1038–1046, 2010.
- [38] B. Dong and K. P. Lam, “Building energy and comfort management through occupant behaviour pattern detection based on a large-scale environmental sensor network,” *J. Build. Perform. Simul.*, vol. 4, no. 4, pp. 359–369, 2011.
- [39] T. Ekwevugbe, N. Brown, V. Pakka, and D. Fan, “Improved occupancy monitoring in non-domestic buildings,” *Sustain. Cities Soc.*, vol. 30, pp. 97–107, 2017.
- [40] W. Wang, J. Chen, T. Hong, and N. Zhu, “Occupancy prediction through Markov based feedback recurrent neural network (M-FRNN) algorithm with WiFi probe technology,” *Build. Environ.*, vol. 138, no. April, pp. 160–170, 2018.
- [41] S. H. Kim and H. J. Moon, “Case study of an advanced integrated comfort control algorithm with cooling, ventilation, and humidification systems based on occupancy

- status,” *Build. Environ.*, vol. 133, pp. 246–264, 2018.
- [42] T. H. Pedersen, K. U. Nielsen, and S. Petersen, “Method for room occupancy detection based on trajectory of indoor climate sensor data,” *Build. Environ.*, vol. 115, pp. 147–156, 2017.
- [43] S. H. Ryu and H. J. Moon, “Development of an occupancy prediction model using indoor environmental data based on machine learning techniques,” *Build. Environ.*, vol. 107, pp. 1–9, 2016.
- [44] K. Whitehouse *et al.*, “Towards occupancy-driven heating and cooling,” *IEEE Des. Test Comput.*, vol. 29, no. 4, pp. 17–25, 2012.
- [45] Q. Huang, “Occupancy-driven energy-efficient buildings using audio processing with background sound cancellation,” *Buildings*, vol. 8, no. 6, 2018.
- [46] W. Wang, J. Chen, Y. Lu, and H. H. Wei, “Energy conservation through flexible HVAC management in large spaces: An IPS-based demand-driven control (IDC) system,” *Autom. Constr.*, vol. 83, pp. 91–107, 2017.
- [47] W. Wang, J. Chen, G. Huang, and Y. Lu, “Energy efficient HVAC control for an IPS-enabled large space in commercial buildings through dynamic spatial occupancy distribution,” *Appl. Energy*, vol. 207, pp. 305–323, 2017.
- [48] W. Wang, T. Hong, N. Li, R. Q. Wang, and J. Chen, “Linking energy-cyber-physical systems with occupancy prediction and interpretation through WiFi probe-based ensemble classification,” *Appl. Energy*, vol. 236, no. August 2018, pp. 55–69, 2019.
- [49] W. Wang, J. Chen, and X. Song, “Modeling and predicting occupancy profile in office space with a Wi-Fi probe-based Dynamic Markov Time-Window Inference approach,” *Build. Environ.*, vol. 124, pp. 130–142, 2017.
- [50] W. Wang, J. Wang, J. Chen, G. Huang, and X. Guo, “Multi-zone outdoor air coordination through Wi-Fi probe-based occupancy sensing,” *Energy Build.*, vol. 159, pp. 495–507, 2018.
- [51] W. Wang, J. Chen, and T. Hong, “Modeling occupancy distribution in large spaces with multi-feature classification algorithm,” *Build. Environ.*, vol. 137, pp. 108–117, 2018.
- [52] O. Vermesan *et al.*, *Cluster of European Research Projects on the Internet of*

- Things: Vision and Challenges for Realising the Internet of Things*, no. March. Brussels: European Commission, 2010.
- [53] B. M. Simon, J. Y. Park, and Z. Nagy, “Inferring Space Occupancy by Bluetooth Device Tracking,” Austin, 2018.
- [54] R Core Team, “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria, 2017.
- [55] S. Fritsch, F. Guenther, M. Suling, and S. M. Mueller, “Training of neural networks. Package ‘neuralnet’.” The R project for statistical computing.” p. 13, 2016.
- [56] J. Heaton, *Introduction to the Math of Neural Networks*. St. Louis, MI: Heaton Research, Inc., 2012.
- [57] R Core Team, “The R Base Package. Package ‘base’.” The R project for statistical computing.” 2018.
- [58] W. O’Brien and H. B. Gunay, “The contextual factors contributing to occupants’ adaptive comfort behaviors in offices - A review and proposed modeling framework,” *Build. Environ.*, vol. 77, pp. 77–88, 2014.
- [59] IEA-EBC, “Total Energy Use in Buildings: Analysis and Evaluation Methods (Annex 53),” Hertfordshire, UK, 2016.
- [60] S. Baldi, C. D. Korkas, M. Lv, and E. B. Kosmatopoulos, “Automating occupant-building interaction via smart zoning of thermostatic loads: A switched self-tuning approach,” *Appl. Energy*, vol. 231, no. August, pp. 1246–1258, 2018.
- [61] F. Oldewurtel, D. Sturzenegger, and M. Morari, “Importance of occupancy information for building climate control,” *Appl. Energy*, vol. 101, pp. 521–532, 2013.
- [62] E. Barbour, C. C. Davila, S. Gupta, C. Reinhart, J. Kaur, and M. C. González, “Planning for sustainable cities by estimating building occupancy with mobile phones,” *Nat. Commun.*, vol. 10, no. 1, 2019.
- [63] X. Cui, B. Mohan, M. R. Islam, and K. J. Chua, “Modelling and Performance Evaluation of an Air Handling Unit for an Air Treatment System with Regulated Outdoor-air Fraction,” *Energy Procedia*, vol. 105, pp. 4718–4723, 2017.

- [64] K. I. Krakow, F. Zhao, and A. E. Muhsin, "Performance analysis of supply and return fans for HVAC systems under different operating strategies of economizer dampers," *ASHRAE Trans.*, vol. 106 (PA, no. 7, pp. 13–25, 2000.
- [65] W. Jung and F. Jazizadeh, "Human-in-the-loop HVAC operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions," *Appl. Energy*, vol. 239, no. January, pp. 1471–1508, 2019.
- [66] H. Saha, A. R. Florita, G. P. Henze, and S. Sarkar, "Occupancy sensing in buildings: A review of data analytics approaches," *Energy Build.*, vol. 188–189, pp. 278–285, 2019.
- [67] O. C. Daniel, V. Ramsurrun, and A. K. Seeam, "Smart Library Seat, Occupant and Occupancy Information System, using Pressure and RFID Sensors," *2nd Int. Conf. Next Gener. Comput. Appl. 2019, NextComp 2019 - Proc.*, pp. 1–5, 2019.
- [68] C. Lee and D. Lee, "Self-Error Detecting and Correcting Algorithm for Accurate Occupancy Tracking using a Wireless Sensor Network," *2019 4th Int. Conf. Smart Sustain. Technol. Split. 2019*, 2019.
- [69] J. Chen, H. Chen, and X. Luo, "Collecting building occupancy data of high resolution based on WiFi and BLE network," *Autom. Constr.*, vol. 102, no. February, pp. 183–194, 2019.
- [70] R. Dekimpe, P. Xu, M. Schramme, P. Gérard, D. Flandre, and D. Bol, "A battery-less BLE smart sensor for room occupancy tracking supplied by 2.45-GHz wireless power transfer," *Integration*, vol. 67, no. October 2018, pp. 8–18, 2019.
- [71] E. Longo, A. E. C. Redondi, and M. Cesana, "Accurate occupancy estimation with WiFi and bluetooth/BLE packet capture," *Comput. Networks*, vol. 163, 2019.
- [72] B. W. Hobson, D. Lowcay, H. B. Gunay, A. Ashouri, and G. R. Newsham, "Opportunistic occupancy-count estimation using sensor fusion: A case study," *Build. Environ.*, vol. 159, no. January, p. 106154, 2019.
- [73] P. Domingues, P. Carreira, R. Vieira, and W. Kastner, "Building automation systems: Concepts and technology review," *Comput. Stand. Interfaces*, vol. 45, pp. 1–12, 2016.
- [74] E. Finch, "Is IP everywhere the way ahead for building automation?," *Facilities*,

- vol. 19, no. 11, pp. 396–403, 2001.
- [75] National Round Table on the Environment and the Economy, “Geared for Change: Energy Efficiency in Canada’s Commercial Building Sector,” Government of Canada, Ottawa, 2009.
- [76] ISO, “ISO 16484-3:2005 Building automation and control systems (BACS) — Part 3: Functions,” Geneva, 2005.
- [77] CEN, “EN 15232 Energy performance of buildings - Impact of building automation, controls and building management,” Brussels, 2012.
- [78] H. N. Rafsanjani, C. R. Ahn, and K. M. Eskridge, “Understanding the recurring patterns of occupants’ energy-use behaviors at entry and departure events in office buildings,” *Build. Environ.*, vol. 136, no. March, pp. 77–87, 2018.
- [79] G. Calis, S. D. Atalay, M. Kuru, and N. Mutlu, “Forecasting Occupancy for Demand Driven HVAC Operations Using Time Series Analysis,” *J. Asian Archit. Build. Eng.*, vol. 16, no. 3, pp. 655–660, 2017.
- [80] K. M. Powell, A. Sriprasad, W. J. Cole, and T. F. Edgar, “Heating, cooling, and electrical load forecasting for a large-scale district energy system,” *Energy*, vol. 74, no. C, pp. 877–885, 2014.
- [81] C. Liu, S. C. H. Hoi, P. Zhao, and J. Sun, “Online ARIMA Algorithms for Time Series Prediction,” *Aaai*, pp. 1867–1873, 2014.
- [82] J. Morais, Y. Pires, C. Cardoso, and A. Klautau, “An Overview of Data Mining Techniques Applied to Power Systems,” *Data Min. Knowl. Discov. Real Life Appl.*, no. February, 2012.
- [83] G. Chicco, “Overview and performance assessment of the clustering methods for electrical load pattern grouping,” *Energy*, vol. 42, no. 1, pp. 68–80, 2012.
- [84] C. Miller, Z. Nagy, and A. Schlueter, “A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings,” *Renew. Sustain. Energy Rev.*, vol. 81, no. December 2016, pp. 1365–1377, 2018.
- [85] M. R. Anderberg, *Cluster Analysis for Applications: Probability and Mathematical Statistics*, 1st ed. New York: Academic Press, Inc., 1973.

- [86] W. H. E. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *J. Classif.*, vol. 1, no. 1, pp. 7–24, 1984.
- [87] D. Munoz, F. Bouchereau, C. Vargas, and R. Enriquez, *Position Location Techniques and Applications*, 1st ed. New York: Academic Press, Inc., 2009.
- [88] G. Jirong, Z. Jieming, and C. Xianwei, "An enhancement of K-means clustering algorithm," *2009 Int. Conf. Bus. Intell. Financ. Eng. BIFE 2009*, vol. 2, no. 2, pp. 237–240, 2009.
- [89] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.
- [90] N. Gaitani, C. Lehmann, M. Santamouris, G. Mihalakakou, and P. Patargias, "Using principal component and cluster analysis in the heating evaluation of the school building sector," *Appl. Energy*, vol. 87, no. 6, pp. 2079–2086, 2010.
- [91] E. Guelpa, S. Deputato, and V. Verda, "Thermal request optimization in district heating networks using a clustering approach," *Appl. Energy*, vol. 228, no. May, pp. 608–617, 2018.
- [92] T. Ahmad and H. Chen, "Short and medium-term forecasting of cooling and heating load demand in building environment with data-mining based approaches," *Energy Build.*, vol. 166, pp. 460–476, 2018.
- [93] C. Chang, N. Zhu, K. Yang, and F. Yang, "Data and analytics for heating energy consumption of residential buildings: The case of a severe cold climate region of China," *Energy Build.*, vol. 172, pp. 104–115, 2018.
- [94] P. Gianniou, X. Liu, A. Heller, P. S. Nielsen, and C. Rode, "Clustering-based analysis for residential district heating data," *Energy Convers. Manag.*, vol. 165, no. December 2017, pp. 840–850, 2018.
- [95] J. An, D. Yan, and T. Hong, "Clustering and statistical analyses of air-conditioning intensity and use patterns in residential buildings," *Energy Build.*, vol. 174, pp. 214–227, 2018.
- [96] Z. Ma, R. Yan, and N. Nord, "A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings," *Energy*, vol. 134, pp. 90–102, 2017.

- [97] J. F. Marquant, L. A. Bollinger, R. Evins, and J. Carmeliet, “A new combined clustering method to Analyse the potential of district heating networks at large-scale,” *Energy*, vol. 156, pp. 73–83, 2018.
- [98] L. Pampuri, N. Cereghetti, D. Strepparava, and P. Caputo, “Analysis of the electricity consumptions: A first step to develop a district cooling system,” *Sustain. Cities Soc.*, vol. 23, pp. 23–36, 2016.
- [99] W. K. O. Chong, H. Naganathan, S. T. Ariaratnam, J. Kim, and S.-Y. Moon, “Applications of Clustering and Isolation Forest Techniques in Real-Time Building Energy-Consumption Data: Application to LEED Certified Buildings,” *J. Energy Eng.*, vol. 143, no. 5, p. 04017052, 2017.
- [100] Z. Ma, R. Yan, K. Li, and N. Nord, “Building energy performance assessment using volatility change based symbolic transformation and hierarchical clustering,” *Energy Build.*, vol. 166, pp. 284–295, 2018.
- [101] T. Schütz, M. H. Schraven, M. Fuchs, P. Remmen, and D. Müller, “Comparison of clustering algorithms for the selection of typical demand days for energy system synthesis,” *Renew. Energy*, vol. 129, pp. 570–582, 2018.
- [102] J. Wang *et al.*, “Energy consumption prediction for water-source heat pump system using pattern recognition-based algorithms,” *Appl. Therm. Eng.*, vol. 136, no. March, pp. 755–766, 2018.
- [103] L. Diao, Y. Sun, Z. Chen, and J. Chen, “Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation,” *Energy Build.*, vol. 147, pp. 47–66, 2017.
- [104] T. Calders, F. Esposito, E. Hullermeier, and R. Meo, “Machine learning and knowledge discovery in databases,” in *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Part III*, 2014, p. 715.
- [105] C. Miller, Z. Nagy, and A. Schlueter, “Automated daily pattern filtering of measured building performance data,” *Autom. Constr.*, vol. 30, no. 49, pp. 1–17, 2015.
- [106] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, “A symbolic representation of time series, with implications for streaming algorithms,” p. 2, 2004.

- [107] A. Ashouri, G. R. Newsham, Y. Hu, W. Shen, and H. B. Gunay, “Model-based and Data-driven Anomaly Detection for Heating and Cooling Demands in Office Buildings,” Ottawa, 2018.
- [108] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks, 1984.
- [109] J. Y. Park *et al.*, “A critical review of field implementations of occupant-centric building controls,” *Build. Environ.*, vol. 165, no. May, p. 106351, 2019.
- [110] ASHRAE, “ANSI/ASHRAE Standard 62.1-2019: Ventilation for Acceptable Indoor Air Quality,” 2019.
- [111] W. Qiu, V. Carey, R. H. Zamar, R. Lazarus, and X. Wang, “Clustering method based on local shrinking. Package ‘clues’. The R project for statistical computing.” 2016.
- [112] T. Caliński and J. Harabasz, “A Dendrite Method For Cluster Analysis,” *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974.
- [113] C. M. Bishop, “Classification and regression training. Package ‘caret’. The R project for statistical computing,” *Handbook of Neural Computation*. 2004.
- [114] A. Ashouri, Y. Hu, G. R. Newsham, and W. Shen, “Energy Performance Based Anomaly Detection in Non-Residential Buildings Using Symbolic Aggregate Approximation,” *IEEE Int. Conf. Autom. Sci. Eng.*, vol. 2018-Augus, pp. 1400–1405, 2018.
- [115] M. M. Ouf, W. O’Brien, and H. B. Gunay, “On quantifying building performance adaptability to variable occupancy,” *Build. Environ.*, vol. 155, no. February, pp. 257–267, 2019.
- [116] I. U. Zeytinoglu, G. B. Cooke, and S. L. Mann, “Flexibility : Whose Choice Is It Anyway ?,” *Ind. Relat. (Berkeley)*, vol. 64, no. 4, pp. 555–574, 2019.
- [117] Z. Yang and B. Becerik-Gerber, “Modeling personalized occupancy profiles for representing long term patterns by using ambient context,” *Build. Environ.*, vol. 78, pp. 23–35, 2014.
- [118] Z. Yang and B. Becerik-Gerber, “The coupled effects of personalized occupancy profile based HVAC schedules and room reassignment on building energy use,” *Energy Build.*, vol. 78, pp. 113–122, 2014.

- [119] AMCA, *AMCA Publication 502-06 (R2009) - Damper Application Manual for Heating, Ventilating, and Air Conditioning*. Arlington Heights, IL, IL: AMCA, 2006.
- [120] J. Y. Park, E. Mbata, and Z. Nagy, “Good to see you again: Capture and recapture method on mobile devices to estimate occupancy profiles,” in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2019, pp. 145–148.
- [121] S. Lu, T. O. Karaguzel, and E. C. Hameen, “Simulations of dynamic ventilation operations with real-time occupancy estimation,” in *ACEEE Summer Study Proceedings*, 2018, no. August, pp. 1–10.
- [122] M. Mysen, S. Berntsen, P. Nafstad, and P. G. Schild, “Occupancy density and benefits of demand-controlled ventilation in Norwegian primary schools,” *Energy Build.*, vol. 37, no. 12, pp. 1234–1240, 2005.
- [123] Delta Controls Inc., *ORCAview Version 3.33 Technical Reference Manual*, 1.6. Vancouver, B.C., 2006.
- [124] S. S. Shrestha and G. M. Maxwell, “An experimental evaluation of HVAC-grade carbon dioxide sensors -Part I: Test and evaluation procedure,” *ASHRAE Trans.*, vol. 115 PART 2, pp. 471–483, 2009.
- [125] W. J. Fisk, D. Faulkner, and D. P. Sullivan, “A pilot study of the accuracy of CO₂ sensors in commercial buildings,” *IAQ Conf.*, no. October, pp. 1–11, 2008.
- [126] H. Zou, H. Jiang, J. Yang, L. Xie, and C. J. Spanos, “Non-intrusive occupancy sensing in commercial buildings,” *Energy Build.*, vol. 154, pp. 633–643, 2017.
- [127] B. S. Çiftler, S. Dikmese, I. Güvenç, K. Akkaya, and A. Kadri, “Occupancy Counting With Burst and Intermittent Signals in Smart Buildings,” *IEEE Internet Things J.*, vol. 5, no. 2, pp. 724–735, 2018.
- [128] J. Wang, N. C. F. Tse, and J. Y. C. Chan, “Wi-Fi based occupancy detection in a complex indoor space under discontinuous wireless communication: A robust filtering based on event-triggered updating,” *Build. Environ.*, vol. 151, no. November 2018, pp. 228–239, 2019.
- [129] B. W. Hobson, H. B. Gunay, A. Ashouri, and G. R. Newsham, “Wi-Fi Based

Occupancy Clustering and Motif Identification : A Case Study,” *ASHRAE Trans.*, vol. 126, no. 1, pp. 1–9, 2020.

- [130] ASHRAE, *2017 ASHRAE Handbook: Fundamentals (SI Edition)*. Atlanta, GA: ASHRAE, 2017.
- [131] L. G. Felker and T. L. Felker, *Dampers and Airflow Control*. Atlanta, GA: ASHRAE, 2009.
- [132] H. B. Gunay, W. Shen, G. R. Newsham, and A. Ashouri, “Detection and interpretation of anomalies in building energy use through inverse modeling,” *Sci. Technol. Built Environ.*, vol. 25, no. 4, pp. 488–503, 2019.
- [133] ASHRAE, “ASHRAE Guideline 14: Measurement of Energy, Demand, and Water Savings,” vol. 8400, pp. 1–165, 2014.
- [134] K. E. Charles and J. A. Veitch, “Outdoor Ventilation Rates in Offices and Occupant Satisfaction (IRC-RR-160),” Ottawa, Canada, 2002.
- [135] B. W. Hobson, H. B. Gunay, A. Ashouri, and G. R. Newsham, “Occupancy-based predictive control of an outdoor air intake damper: A case study,” in *Proceedings of IBPSA-Canada’s eSim 2020 Conference*, 2020, pp. 1–8.