Investigating the impact of households' occupancy patterns and activity routines on daily load profiles: a data-driven approach

Saba Akbari

A Thesis

In

the Department

of

Building, Civil, and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

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

Concordia University

Montreal, Quebec, Canada

January 2021

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Saba Akbari

Entitled: Investigating the impact of households' occupancy patterns and activity routines on daily load profiles: a data-driven approach

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

Master of Applied Science (Civil Engineering)

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

Signed by the final examining committee:

_____ Chair
Dr. Mazdak Nik-Bakht

_____ Examiner
Dr. Ursula Eicker

_____ External to Program
Dr. Amin Hammad

_____ Thesis Supervisor
Dr. Fariborz Haghighat

Approved by _____
Dr. Ashutosh Baghchi, Chair
Department of Building, Civil and Environment Engineering

January 13, 2021 _____
Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering & Computer Science

Abstract

Investigating the impact of households' occupancy patterns and activity routines on daily load profiles: a data-driven approach

Saba Akbari, M.A.Sc

Concordia University 2021

Examining individual households' load profiles and discovering contextual and temporal factors of energy usage (e.g., occupancy, time of use, and occupant activity) gain lots of popularity in recent studies. Given the proliferation of Home and Building Energy Management Systems (HEMS and BEMS) and the availability of high-resolution data of households' energy usage, it is possible to gain a deeper understanding of temporal factors of load profiles and take advantage of the services offered by these systems. The incorporation of smart meters in the grid has several economic and environmental benefits. These technologies (1) provide the opportunity for appliance scheduling, which can reduce electricity costs on the customer-side, and (2) optimize the integration of intermittent renewable energy sources to the electricity grid. Despite the importance of temporal determinants, previous works mainly focused on determinants of annual end-use load. Additionally, studies on residential energy mainly address district and city scales, while small-scale analyses are highly overlooked. Based on the identified limitations, in this study, two time-series analysis methods (k-shape clustering and change point detection) are implemented on historical, sensor-collected data of three residential units in order to discover the frequent occupancy schedule patterns of each household and identify the high- and low- consumption periods within each occupancy pattern. Then LASSO regression is utilized to find the comparative contribution of various activity factors on households' energy usage (e.g., kitchen-, living room-, bathroom-, or bedroom-related activities indicated by plug loads recorded in specific rooms of the apartments) during the identified energy consumption periods. The results suggest that occupancy patterns are able to explain temporal variations in daily load profiles, and the shape of daily load

profiles can be characterized by the occupancy schedule pattern of a day. Furthermore, the analysis of this study can make households aware of the most influencing activities during high-consumption periods. And as a result, households can reduce their energy bills by shifting the energy-consuming activities from high-consumption periods to off-peak periods.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my supervisor, Professor Fariborz Haghighat, for letting me be part of his incredible research group and funding my master's studies. His guidance, thoughtful comments, and recommendations have guided me through this research.

The people with the greatest indirect contribution to this thesis are my parents, Raheleh Esfahani and Mahmoud Akbari, whose emotional support, patience, and devotion is the greatest gift one could ever have. I sincerely appreciate their unconditional love and feel blessed to have their support in my life.

I am also thankful to meet my colleagues at Concordia University (Maryam, Shahin, Milad, Sajjad, Soroush, Niousha, Mohammad, Moein, Karthik, Ying, Jun, Bowen, and Behrang), who soon became my treasured friends. My journey at this University could have been significantly different without their help and the positive, joyful environment of our office and the Energy and Environment Group.

Last but not least, I would like to thank Dr. Mohamed El Mankibi at École Nationale des Travaux Publics de l'État (ENTPE) for providing the data of this study.

Contents

List of figures

List of tables

List of acronyms

AC	Air-Conditioning
AI	Artificial Intelligence
BAS	Building Automation System
BATH	Bathroom
BED	Bedroom
BEMS	Building Energy Management System
CPD	Change Point Detection
CVI	Cluster Validation Index
DB star	Davis-Bouldin star
DM	Data Mining
DTW	Dynamic Time Warping
ED	Euclidean Distance
EV	Electric Vehicle
HEMS	Home Energy Management System
HVAC	Heating, Ventilating, and Air-Conditioning
IQR	Interquartile Range
KIT	Kitchen
LIGHT	Lighting energy consumption
LIV	Living room
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MSE	Mean Squared Error
OLS	Ordinary Least Squares
OTH	Other
RSS	Residual Sum of Squares
SBD	Shape-Based Distance
SD	Standard Deviation
SIC	Schwarz Information Criterion

TSS	Total Sum of Squares
TUS	Time-Use Survey
VIF	Variance-Inflation factors

List of symbols

P	Plug load variable
μ	Mean
σ	Standard deviation

K-shape parameters

z_i	Z-score values
R_0	Autocorrelation
CC	Cross-correlation
NCC	Normalized cross-correlation
k	Number of clusters
μ_k	Centroid time-series

Change point detection parameters

Q	Maximum number of change points
H_0	Null hypothesis
H_1	Alternative hypothesis
q	Number of change points
k	Location of change points
n	Length of time-series

Regression parameters

β	Estimated regression coefficient
λ	Regression penalty parameter
x_i'	Min-max normalized value
R^2	R-squared

Chapter One

1. Introduction

1.1. Background and motivations

The residential sector accounts for almost one-third of U.S. electricity use (Figure 1-1), and around 39% of this consumption is for household appliances (28%) and lighting (11%) (Figure 1-2) (Schwartz et al., 2017).

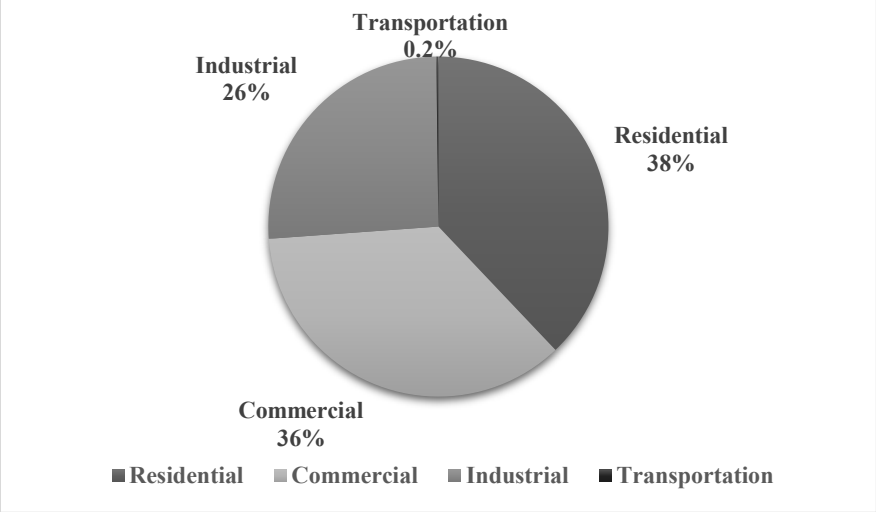


Figure 1-1. Electricity consumption by sector in U.S., 2014 (Schwartz et al., 2017)

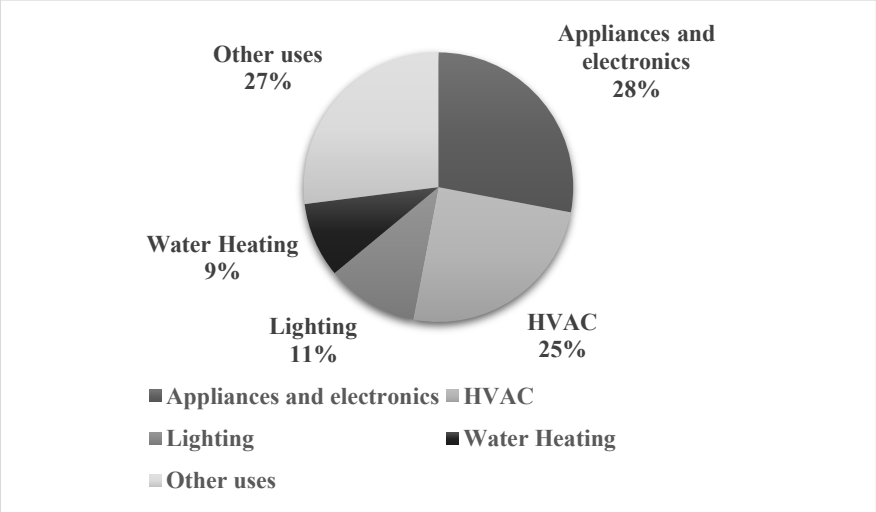


Figure 1-2. Residential electricity consumers in U.S., 2014 (Schwartz et al., 2017)

According to the Annual Energy Outlook report published by U.S. Energy Information Administration, the residential electricity consumption will grow by 0.6% per year through 2050 because of the increasing demand for electric appliances and devices (Figure 1-3) (Sugawara & Nikaido, 2019). Given the importance of residential electricity demand, it is essential to have a good understanding of its influencing factors to enhance demand response in this sector.

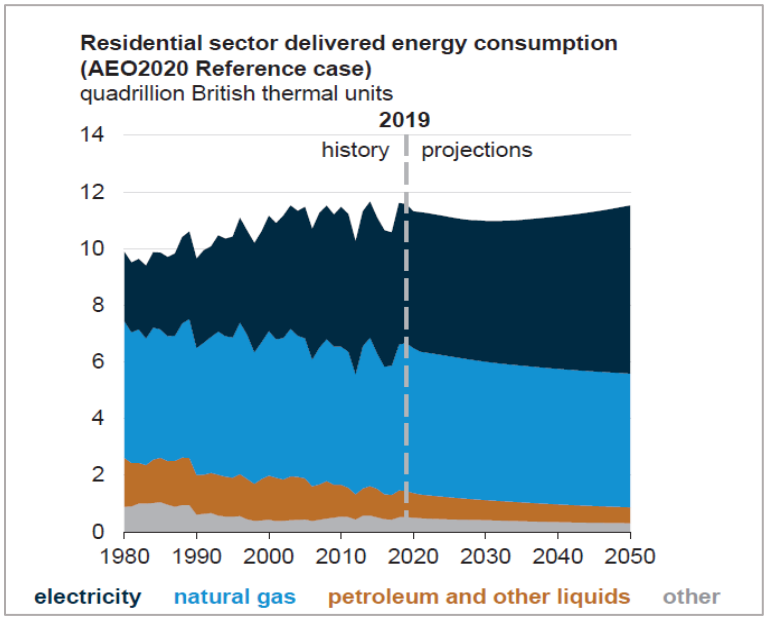


Figure 1-3. Annual prediction of residential energy demand through 2050 by energy type; adopted from (Sugawara & Nikaido, 2019)

One of the most important challenges regarding residential energy consumption is peak demand. The economic and environmental issues introduced by imbalances between supply and demand during peak periods are becoming increasingly significant. As mentioned by (Thieblemont et al., 2018), the mismatches between demand and supply, especially during the peak hours, force energy suppliers to (1) buy energy from their neighbors at an expensive rate or (2) turn to fossil fuels instead of renewable sources to respond to the peak demand. As a result of the mentioned issues, the smart grid solutions have gained popularity in recent years since they can bring economic and environmental benefits to the demand-supply system (Grunewald & Diakonova, 2018) (Kirschen, 2003). On the small scales (i.e., household or building), (Zhou et al., 2016) emphasized the role of Home Energy Management Systems (HEMS) in alleviating the two-way interaction between consumers (demand-side) and energy providers (supplier-side). (Beaudin & Zareipour, 2015)

reported that HEMS can reduce the cost of electricity by 23.1%, and the residential peak demand can be reduced by 29.6% using these systems. (Zhou et al., 2016) summarized functionalities of smart HEMS as: 1. "monitoring" real-time energy consumption of households, 2. "collecting" and storing data regarding the amount of appliance consumption, amount of energy generated from available resources in the distributed grid, and energy storage capacity, 3. "controlling" and 4. "management" of energy usage in smart homes through consideration of renewable energy systems, home appliance operations, energy storage systems, and plug-in electric vehicles (EV), and eventually 5. "alarming" in case of fault and abnormality detection in the systems. Based on these functionalities, some of the advantages of smart HEMSs are as follows:

- HEMSs can contribute to the integration of multiple renewable sources of energy through the provision of a sophisticated storage and controlling platform that schedule charging/ discharging operations based on the energy demand and availability of different renewable sources at each time (Vijayapriya & Kothari, 2011) (Zhou et al., 2016);
- These smart systems can be programmed to follow an appliance scheduling scheme based on real-time electricity pricing during peak and off-peak periods to reduce costs for consumers and reach energy efficiency (Ozturk et al., 2013) (Zhou et al., 2016). Appliance scheduling can reduce the peak-to-average ratio of one household energy usage by 19.7%, and the same statistics for aggregated usage of 10 households is 34.6% (Zhao et al., 2013).
- Furthermore, (Nilsson et al., 2018) noted that HEMSs have the potential to improve energy feedback effectiveness and increase households' awareness regarding their energy consumption. The authors further stipulated that through employing a variety of metrics, formats, and data aggregation levels, HEMSs can provide transparent and interpretable energy feedbacks. The concise feedback with high-level interpretablity can bridge the gap between households' everyday activities and the amount of energy consumption and are easier for households to use them and reduce energy consumption.

Advanced metering technologies connected to these systems can collect high-resolution data from households and their energy-related behaviors over a long period, making it possible to analyze households' energy usage on a daily or hourly basis. Perceiving energy consumption as load profiles rather than daily, monthly or annual end-uses can enhance energy estimations on higher resolutions and provide opportunities for shifting load from peak to off-peak periods (Torriti,

2020). Investigating temporal drivers of energy usage can help us realize when, how, and why the energy is consumed, so understanding these temporal factors leads to improved energy management in the residential sector.

1.2. Objectives

This study aims to develop a methodology framework that reveals the impact of occupants' routines on the shape of load profiles in residential apartments. The proposed framework of this study is examined on data of three residential apartments collected by HEMS (Home Energy Management System) over one year. This study aims to achieve the following objectives:

- To develop a methodology framework to extract the usual routines of individual households regarding their presence and energy-related activities
- To investigate the effect of temporal factors of energy consumption (i.e., occupancy and occupants' activity) on the shape of load profiles

The purpose of this investigation is to find out when and why the shape of load profiles regularly changes. Occupancy and occupants' activity data will be analyzed to find the answers to these questions. Finding the factors of energy consumption on an hourly basis is beneficial for energy interventions since occupants can get aware of their activities during the peak demand periods; plus, suggestions regarding appliance scheduling and peak shifting can be presented to households in order to reduce their peak demand and avoid high electricity cost rates during the peak hours. Fulfilling the mentioned objectives can be useful for regulating the automation of buildings' systems to the frequently practiced routines of households' members.

1.3. Organization of the thesis

In this study, in **Error! Reference source not found.**, the studies exploring the determinants of r esidential energy are overviewed, and different types of data collection methods and their impact on residential energy analyses are critically reviewed. Based on the previous works, at the end of chapter **Error! Reference source not found.**, the existing research gaps are mentioned, and the a pproach of this study to addressing them is explained. Chapter 0 reports the characteristics of the

data sets used in this study. Afterward, the proposed methodology framework of the current research and its potentials to achieve the defined objectives are introduced in chapter **Error! R eference source not found.**. Lastly, the results of the analyses are thoroughly discussed in chapter 0.

Chapter Two

2. Literature review

2.1. Residential energy: determinants, data, and scale in previous analyses

So far, the available data regarding residential energy consumption mainly contain information about socio-demographic, dwelling characteristics, and appliance ownership factors. Based on the review of (Jones et al., 2015), most of the studies exploring the mentioned factors are mainly conducted on large scales with a large number of dwellings (e.g., usually more than hundreds). Plus, each household's annual or monthly end-use loads are mainly considered for impact assessment of the mentioned factors. (Satre-Meloy, 2019) found the comparative contribution of several structural factors (e.g., dwelling characteristics and ownership of certain appliances) and occupant-related drivers (e.g., occupants' energy literacy and attitudes towards consumption, socio-economic factors, etc.) on the annual electricity consumption of almost 1000 households. The findings suggest that dwelling size, number of households' members, ownership of air-conditioner (AC) and electric vehicles (EV) are strongly associated with high consumption. In contrast, some habitual behaviors like unplugging unneeded appliances and turning off AC during the times of absence are associated with low energy consumption in residential dwellings. With the same approach to find the comparative contribution of appliance ownership and appliance usage factors, dwelling characteristics, socio-demographic parameters, and occupants' attitudes on annual energy consumption of 845 households, (Huebner et al., 2016) discovered that the regression model employing only appliance-related variables is able to explain the residential energy consumption sufficiently, and inclusion of socio-demographic and dwelling factors to the same model can only enhance the energy predictions slightly; this is while occupant attitude factors proved to be neutral predictors. The findings of the above-mentioned studies suggest that a combination of socio-demographic, dwelling characteristics, and appliance ownership factors can sufficiently explain the size of end-use load (Satre-Meloy, 2019). However, these factors cannot explain fluctuations in the shape of load profiles and timing of consumption since none of the mentioned factors are sufficient representatives of the daily routines of different socio-demographic groups (Torriti, 2020). (Yu, Fung, et al., 2011) summarized the factors influencing

building energy consumption in two main categories of related and unrelated to occupant behavior. In the mentioned study, the authors suggested that occupant behavior's impact on buildings' energy can be grasped if the factors unrelated to occupant behavior (e.g., climatic conditions, building characteristics, building services features, number of households' members, etc.) are similar among buildings. Inspired by the same idea, (Ashouri et al., 2019) took advantage of k-means clustering to group buildings which have similar physical characteristics (i.e., occupant number, climate, and building characteristics, etc.) to rank buildings regarding their end-use consumption and provide energy feedbacks to the occupants of buildings with similar physical characteristics. The results suggested that the consumption gap between end-use loads of buildings with similar physical characteristics can be attributed to occupant behavior. However, in this study, temporal occupant-related factors such as occupied and unoccupied periods and hourly consumption of different activities have not been considered, and only daily end-use loads are compared and analyzed for the energy feedback purposes. As a result, there is the possibility of comparing households who spend most of their time inside their apartments and consuming energy against those who are usually absent and inevitably consume less. Two of the well-recognized factors of energy consumption that are highly variable at different time-periods of day are occupancy and occupant activity (Torriti, 2020). Timing of occupied periods (Yao & Steemers, 2005) (Richardson et al., 2008) (Buttitta et al., 2017), along with the timing and type of occupant activities (Satre-Meloy et al., 2020) (Widén & Wäckelgård, 2010) (Gajowniczek & Zabkowski, 2017), have an evident impact on daily load profiles. Recent research emphasizes the need to investigate energy consumption factors on an hourly or daily basis rather than monthly or annually (Satre-Meloy, 2019). Nowadays, owing to the proliferation of monitoring and availability of high-resolution occupancy and energy consumption data, it is possible to explore factors influencing the shape of load profiles on an hourly or even higher-resolution basis.

In this section, some of the common data collection methods and the models obtained from them in the previous works of residential energy are reviewed. Based on the advantages and drawbacks of data collection methods, their potentials for different occupant behavior and energy analyses are also discussed. The most common monitoring techniques of occupant behavior are sensor monitoring and surveys (Gilani & O'Brien, 2017). Self-reports like time-use surveys are commonly used to address occupancy and occupant activities in the residential sector (Torriti, 2020). In some of these studies, the metered consumption data accompany self-reports and surveys

to assess the impact of occupants on energy consumption (Satre-Meloy et al., 2020) (Huebner et al., 2016) (Viegas et al., 2016) (Vassileva et al., 2012). Surveys and self-reports are easier and more economical to collect than sensor-monitoring, which can be relatively expensive at first installation (Zhang et al., 2018) (Gilani & O'Brien, 2017); the mentioned qualities make surveys more suitable for studies on larger scales. However, the main issue with reports is that actual and reported events are not always in accordance with each other (Hong et al., 2017) (Zhang et al., 2018) (McKenna et al., 2018). To examine the consistency between reported and actual consumption, (Durand-Daubin et al., 2013) investigated several indicators such as duration, time of use, and intensity of energy consumption obtained from questionnaires, diaries, and measured energy consumption data of 60 buildings in France. Despite the overall consistency between intensity and operation time of appliances obtained via the mentioned data collection methods, some level of inconsistency is reported for some appliances and indicators (Durand-Daubin et al., 2013). Additionally, self-reports data are difficult to be maintained for long-term data collection practices (Gilani & O'Brien, 2017), and they are usually collected for a limited number of days. For instance, national Time Use Surveys (TUS), one of the most common available data in residential energy studies, are collected for one single weekend day and one single working day (McKenna et al., 2018). Hence, the models developed from these surveys are based on the assumption of two repetitive behavioral patterns for all working days and all weekend days (Buttitta et al., 2017) (Buttitta et al., 2019). As stated by (McKenna et al., 2018), this issue with TUS leads to ignorance of the inherent flexibility of occupant behavior. TUS data are reports filled by occupants at, usually, 10-minute intervals. The data contains information about the location of occupants and their activities at each time-interval. National time-use surveys are one of the most popular data collection methods when it comes to residential energy. Surveys like TUS can be collected for a large number of households (e.g., on district or city levels). Many studies addressing residential occupancy and occupant activities have used TUS to find groups of occupants (or households) with similar behavioral characteristics (i.e., customer segmentation) (Diao et al., 2017) (Buttitta et al., 2017) (Buttitta et al., 2019) (Torriti, 2020). Although customer segmentation and developing models for a small number of customer groups can be useful for demand-response programs on large scales, (O'Brien et al., 2017) argued that aggregating data of several occupants on small scales (16 occupants in this study) ignores the inter-diversity in behavior of occupants. According to the authors, attention must be paid to the selection of sample size so that the obtained

model can sufficiently capture diversity among occupants' behavior on large scales. However, for small-scale occupant behavior analyses, the authors did not recommend aggregating multiple occupants' behavior since the diversity in occupants' behavior (e.g., arrival and departure hours) is more tangible on small scales. (Li et al., 2019) reported that aggregation can introduce a smoothing effect on load profiles since aggregated data of several occupants with diverse peak load timing will result in smoothed load profiles during peak periods. Using WikiEnergy data, (Gajowniczek & Zabkowski, 2017) also demonstrated how the aggregated load profile of 46 households represents a smoother peak load in the morning compared to a random single household load profile. This aggregation and smoothing effect can lead to smaller prediction errors in large-scale analyses that cover data from a large number of consumers. (Gajowniczek & Zabkowski, 2017) reported that the mean absolute percentage error (MAPE) in the prediction of individual household's consumptions (20 - 100%) is extremely higher than that of aggregated consumption of multiple households (1 - 2%). It can be concluded that the impact of occupant behavior tends to introduce some level of uncertainty to energy predictions on small-scale levels, while this uncertainty is less effective on larger scales (Li et al., 2019). This fact emphasizes the need to address individual households' traits in small scales to increase the prediction accuracy of each household's consumption. (Pereira & Ramos, 2019) discussed the importance of adapting energy management systems (e.g. HEMS, BEMS, BAS, etc.) to the specificities of each household in order to automate the operation of some building systems such as windows and roller shutters. This automation can increase the visual and thermal comfort and improve the indoor air quality. Additionally, a good knowledge of specificities of each occupant's routines can be useful in optimizing the zone-level designs (e.g., terminal HVAC units) (O'Brien et al., 2017). So far, the modelling efforts primarily tried to fit data to a limited number of customer groups, while the diversity in routines of individuals is largely ignored (O'Brien et al., 2017). Sensor monitoring allows for collecting data for a prolonged period and can provide more suitable granularity of real data (Zhang et al., 2018). In general, sensor-collected data are more suitable for discovering the diverse energy consumption routines of individual households.

2.2. Data analysis in occupant behavior modeling

Data science is a fast-growing science that offers a variety of powerful, quantitative, analytical tools and techniques capable of handling large volumes of data that can be applied in many fields of study, including building energy. Some of these tools are data warehousing, artificial intelligence (AI), machine learning (ML), data mining (DM), and data visualization, which can be very helpful in decision-making processes. The mentioned techniques are compelling in handling large volumes of data collected by BEMS and HEMS since this data is of high resolution (e.g., records data at one-minute or even less than one-minute intervals) and can be collected for an extended period, so the size of datasets is relatively large. Considering building energy and occupant behavior studies, DM techniques (e.g., clustering, association rule mining, and classification methods, etc.) and statistical models such as regression models are some of the frequently practiced tools in exploring associations among different variables and discovering hidden, recurring patterns in behavior of occupants (Li et al., 2019) (Zhang et al., 2018) (Fan et al., 2018). Next, some of the most common data analysis methods practiced for pattern discovery and profiling are reviewed.

2.2.1. Pattern discovery and profiling

Among all the available methods, clustering has broadly been applied to find typical behavioral patterns of occupants. For example, (D'Oca & Hong, 2014) used k-means clustering and discovered four distinctive patterns of window opening/closing behavior among the occupants of 16 naturally-ventilated office units. The obtained patterns are used to categorize office users based on their attitudes regarding interaction with windows. By taking a clustering-then-classification approach, (Yu, Haghighat, et al., 2011) first identified groups of buildings characterized by similar consumption behaviors using k-means. Then they applied decision tree to extract rules that explain the consumption characteristics of each group. Nonetheless, the clustering method's main popularity in occupant behavior and energy studies is associated with the fact that this method can be applied to time-series data to extract occupancy and energy load profiles. The profiles are characterized by similar fluctuation patterns of a certain value (e.g., level of occupancy, amount of consumption, etc.) over a certain time-window (usually a 24-hour window is considered). Load profiling has several merits, such as providing a basis to explore factors that have relations to the obtained profiles (Satre-Meloy et al., 2020) (McLoughlin et al., 2015), providing tailored energy

recommendation, load shifting, and enhancing demand-response managements (Kwac et al., 2014) (Kwac et al., 2018). Although conventional similarity distance measures like Euclidean distance (ED) are used very frequently to find clusters of load profiles (Lavin & Klabjan, 2015) (Kwac et al., 2018) (McLoughlin et al., 2015) (Liu et al., 2012) (Chicco et al., 2006), some recent studies (Satre-Meloy et al., 2020) (Teeraratkul et al., 2018) suggested that methods employing conventional distance measures cannot properly capture the temporal variations in time-series of load profiles. It is due to the fact that Euclidean distance can only compare the distance between corresponding time-points of two given time-series, so do not appropriately capture the temporal variations in load profile time-series. However, instead of using conventional distance measures to find dissimilarities between time-series of raw energy consumption, some studies attempted to make changes in the time-series of raw energy consumption to extract load profiles using conventional distance measures. For instance, (Xiao & Fan, 2014) segmented daily time-series into three modes of morning, afternoon, and night period, then used minimum, maximum, mean, and standard deviation of energy consumption within each mode to find the days with similar energy consumption statistics and extract the load profiles. In another study, to discover the peak-time consumption profile of 269 households, (Satre-Meloy et al., 2020) utilized daily time-series of cumulative consumption instead of daily raw consumption time-series and compared the results of several clustering algorithms with different conventional distance measures (e.g., Euclidean, Manhattan distance measure). The results show that the use of cumulative load time-series allows for taking advantage of the simplicity of Euclidean distance as a distance measure to extract load profiles. On the other hand, there are similarity measures with more flexibility towards temporal variations of time-series and are tailored for time-series data. (Teeraratkul et al., 2018) used dynamic time warping (DTW) distance to find distinctive load profiles and extracted a smaller number of clusters comparing to the previously practiced clustering methods with conventional distance measures. The obtained clusters with DTW showed more cohesiveness with lower variability among the time-series clustered in the same groups. (Yang et al., 2017) compared clustering results obtained from DTW and Shape-based distance (SBD) measures on consumption time-series of 10 institutional buildings. They found out SBD outperforms DTW, and the obtained profiles of K-shape increase the overall accuracy of energy predictions. In addition to consumption time-series, profiling has also been used for time-series of occupancy. Occupancy is commonly recognized as the pre-requisite for occupants' energy-related behaviors and activities (Li et al.,

2019). (Wei et al., 2019) proved the importance of occupancy level as an input for the energy prediction model to achieve higher prediction accuracy. (J. Zhao et al., 2014) evaluated the impact of office occupancy profiles on the HVAC energy use by integrating these profiles in energy simulations. They concluded that the influence of occupancy profiles on HVAC consumption varies for different climate zones. Plus, for some zones, consideration of hourly occupancy rate can reduce the energy consumption significantly. (D'Oca & Hong, 2015) found that the occupancy state of office buildings (occupied or vacant) can be accurately predicted by decision tree (C4.5) that uses time-related variables (e.g., season, day of week, etc.) and window change behaviors as input. In this study, the profiling is done by k-means clustering, and four identical occupancy schedule patterns are identified. Furthermore, (Liang et al., 2016) showed that occupancy profiles could also be explained and predicted using time-related variables (e.g., season, day of week) in office buildings. It can be seen in the literature that sensor-collected data has generally been used to find occupancy profiles in office buildings and the figures usually indicate the number of present occupants (i.e., occupancy level) at each time interval. However, in the residential sector, occupancy is mainly addressed using national TUS. Markov-Chain techniques are applied to TUS data of residential users to generate occupancy models that determine occupancy state (e.g., absent, present and active, present and non-active) at each time interval. These statistical models are regularly trained based on different categories such as weekday/weekend, number of household members (Richardson et al., 2008), or buildings' type (e.g., detached houses, apartments, etc.) (Widén & Wäckelgård, 2010). However, (Buttitta et al., 2017) identified the limitation of such models in the incapability of categorizing users based on their presence routines and generating representative occupancy profiles for each category of users. This limitation exists because the generated occupancy models using probabilistic approaches like Markov-Chain for building stock users can only output the probability of change in occupancy states of a large number of users who might have utterly different at-home presence routines. To tackle this issue, (Buttitta et al., 2017) applied k-modes clustering on TUS data of building stock users and discovered several occupancy profiles representing groups of users with similar at-home presence schedules. As stated by the authors, these profiles are more suitable to be integrated in scalable energy-use models of building stocks. (Buttitta et al., 2019) proved that consideration of distinctive occupancy schedules for different groups of customers can increase the accuracy of annual heating energy demand estimations. (Diao et al., 2017) used k-modes clustering on sequences of 9 identified activities

obtained from American TUS (i.e., away from home, grooming, dishwashing, laundry, sleeping, cooking, cleaning, leisure, other) to find groups of occupants with similar activity patterns. The discovered patterns have been further used for energy estimation purposes. However, as mentioned earlier, one of the shortcomings of TUS data is the limited duration of data collection, and these surveys are collected for one single weekday and one single working day (Buttitta et al., 2017), so they are incapable of capturing intra-diversity of individuals' behavior and are not appropriate to address the diversity of occupant behavior in small-scale analyses.

2.3. Identified research gaps

Based on the review, there is a trend in the literature regarding the discovery of temporal factors that impact residential energy consumption patterns, especially at the household level. The reason behind this growing trend is to find associations between time-related factors, user activity, and, eventually, the shape of load profiles. The insights obtained from this analysis can provide the opportunity for improved energy forecasts (Gajowniczek & Zabkowski, 2017) and enhanced energy management in the residential sector. Furthermore, in modeling approaches, especially in the residential sector, it is mainly attempted to fit data to a limited number of customer groups, and diverse routines of individuals have not been captured. In the residential sector, TUS data are essentially used for collecting data on occupancy and occupant activities. These surveys are only collected for a limited number of days, so models obtained from them cannot capture the diversity in the presence and activity routines of households.

Based on the review of previous works on residential energy, the following research gaps are identified:

1. While fitting models to data of a small number of customer groups is frequently practiced in the previous works, modeling specific routines of individual households have not been satisfactorily exercised. As a result, the inherent flexibility of individual households regarding their presence and energy consumption routines is ignored.
2. It is also discussed that although the factors that impact the size of end-use load are well-addressed, little is known about temporal factors of residential energy consumption.

Therefore, in this study, the attempt is to develop a data-driven framework to explore the impact of temporal and contextual determinants like occupancy and occupants' activity on individual households' load profiles. The goal is to answer the following questions:

- How diverse are a single household's routines regarding their presence and energy consumption throughout a long period?
- Which hours are the time points when load profiles are bound to change (i.e., increase or decrease) significantly?
- What types of activities are the key drivers of energy consumption during high- and low-consumption periods?

The next chapter explains the case study apartments and the available data from the apartments and their potentials to answer the mentioned questions.

Chapter Three

3. Data

The datasets used in this study belong to three residential units of a building complex called Hikari. Hikari is an energy-efficient complex equipped with high-tech building/home energy management systems (BEMS/HEMS), which monitors a variety of data, including occupant-, energy-, and environmental-related. The three residential units selected for the current study have similar floor plans (see Figure 3-1) and are located on different levels of a building in the Hikari complex (see Table 3-1). The datasets of these units contain one-minute records of occupancy, lighting, and plug load in the year 2016:

- Occupancy variables are motion detection data recorded for each apartment's room and have a binary format (i.e., contains values of zero and one representing unoccupied and occupied status, respectively).
- Each lighting variable represents the amount of energy consumption for lights in a specific room of an apartment, so the lighting variables type is numerical, and the unit is W.h.
- Plug variables contain energy consumption values recorded from each outlet in an apartment; therefore, they represent the households' appliances energy consumption. Each plug variable belongs to a certain room. However, as opposed to motion variables, the plugs' location is unknown. In section 5.1, a possible solution is brought, which helps to have an assumption about the location of each plug sensor. Since specific appliance loads are not available in this study, each plug variable is perceived as occupants' activities and interactions with appliances situated in a specific room to make these variables more understandable. The term "activity" refers to the interactions of occupants with appliances that lead to consumption in a specific zone. It is evident that each plug variable can represent the consumption value of more than one device.

It should be noted that the initial data sets hold one-minute records of cumulative load for lighting and plug variables. Therefore, to obtain the actual lighting and plug energy consumption values on a one-minute basis, the cumulative consumption values of consecutive records are simply subtracted (i.e., the subtraction of cumulative consumption value in 5:01 from the value in 5:00 is assigned to 5:00).

Table 3-1. General information on residential units

Apartment ID	Floor No.	Floor Area (square meter)	No. of residents
112	1	109.3	1 person
152	5	110	2 people
162	6	110.2	2 people

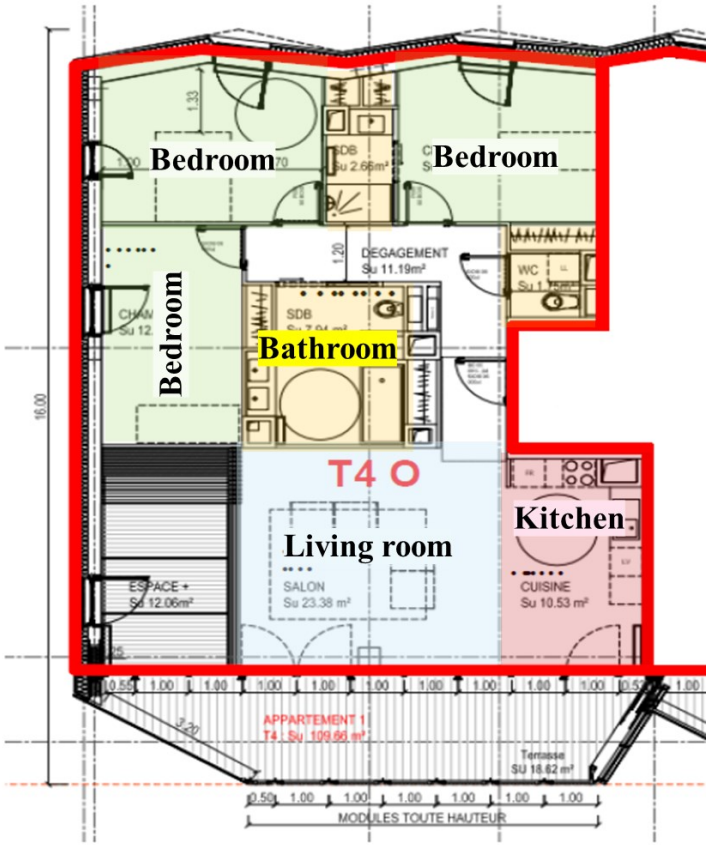


Figure 3-1. Floor plan of the residential units (apartment 112, 152, and 162)

Chapter Four

4. Methodology

4.1. Methodology framework

The data-driven framework, illustrated in Figure 4-1, is proposed to achieve the objectives of the current study. As mentioned in chapter 0, data sets of three residential apartments are available for this study. The proposed framework is applied to the data of each apartment separately to see whether it can be generalized to households with different characteristics. After data preprocessing (first step of framework), to find the occupancy patterns, a clustering method, called K-shape, is used. The function of this method is more suitable for time-series compared to other conventional clustering methods such as K-means. Daily time-series of occupancy (obtained from motion detection data) are used as input of K-shape, and the outputs are clusters containing days with similar occupancy schedule profiles. In the third step, change point detection (CPD) is applied to daily time-series of energy consumption for days grouped in the same occupancy cluster. The obtained results of change point detection are hours when energy consumption is highly probable to either decrease or increase significantly. These hours can specify the periods of day characterized by a particular energy consumption behavior (i.e., high or low) within each occupancy cluster. In the fourth stage of the analysis, a regression model is trained for each consumption period in order to find the influencing activity factors within each period. Thus, the plug variables (i.e., plug number 1, 2, …, n) coupled with the total lighting load are utilized as predictors, while the total energy consumption is considered as the target variable of the regression models. Each plug variable belongs to a certain room of each apartment, so the energy consumed from each plug can be seen as an activity that belongs to the appliance(s) located in that room. Therefore, the regression model outputs some coefficients that indicate the most influencing activity factors on a household's energy consumption for a given period. Regularization penalties are applied to the regression models to better differentiate the variables regarding their contribution to total energy consumption. The penalized regressions shrink the coefficients of all predictors to zero or near-zero values. This shrinkage is more severe for irrelevant predictors rather than influencing ones. In this study, the LASSO penalty is selected since this penalty term can achieve highly interpretable and sparse results by shrinking insignificant variables' coefficients to zero.

Therefore, the coefficient set acquired from LASSO regression is expected to be sparser as some predictors' coefficients will be set to zero.

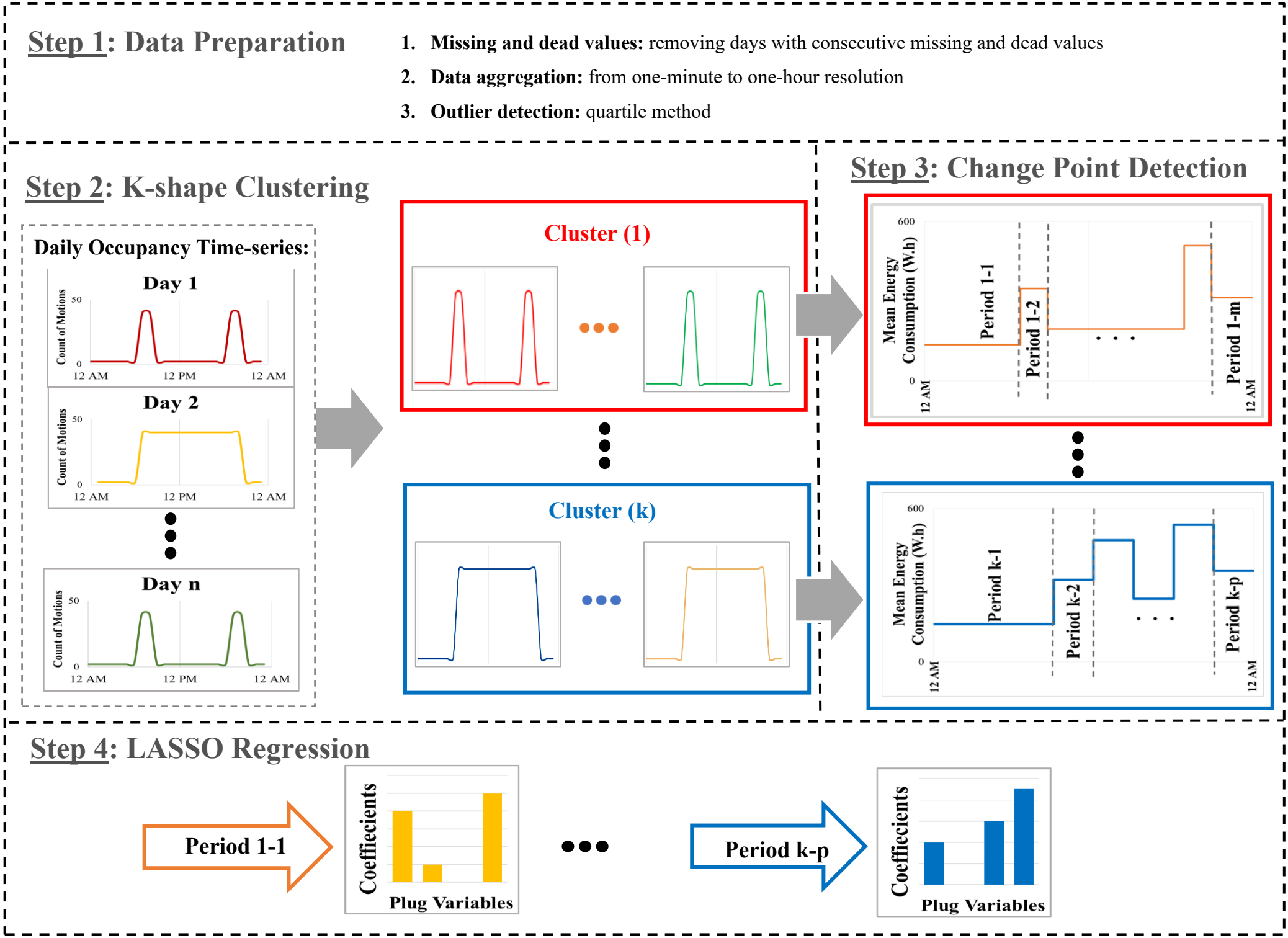


Figure 4-1. Methodology framework

Microsoft Excel and R programming language are utilized to implement the framework. For the data preparation step, Microsoft Excel is used to handle missing and dead values, and "lubridate", "dplyr", and "reshape2" packages in R programming language are utilized to read and manipulate time variables (e.g., hour, day of week, month, season, etc.), aggregate data, and separate time-series of length 24. "ggplot2" is used for plotting. For time-series clustering and cluster validation, "dtwclust" is used. Change point detection is carried out using the "changepoint" package. Regression analysis is implemented using both "glmnet" and "caret" packages in R.

4.2. Data preparation

4.2.1. Missing and dead values

Two of the common issues with sensor-collected data are missing values and dead values. Missing values are data points for which the value is not recorded. As mentioned earlier, for lighting and plug variables, cumulative consumption values are transformed to minute-wise consumption values. Before this transformation, missing values found in one-minute cumulative consumption variables (i.e., lighting and plug variables) were filled with their respective previous value. Similarly, regarding one-minute binary values of motion detection, the value of the previous record is used to fill a missing record. Furthermore, missing values can occur at consecutive records in the datasets. In cases where missing values appeared at more than 30 consecutive records (i.e., 30 minutes), the entire data of the respective day in which the continuous missing values have occurred should be removed entirely. Another type of issue in raw sensor-collected data is dead values. Dead values are defined as continuous data points (i.e., records) for which the sensor has recorded the same value for a long period (Xiao & Fan, 2014). The same measure applied for continuous missing values is adopted for dead values as well. Therefore, days containing more than 30-minutes of dead values are removed from the data sets. It should be mentioned that in our datasets, the cases of consecutive missing records and dead values usually occur in several days, which is much greater than a 30-minute window. However, a 30-minute window is suggested for removing the entire day instead of filling the continuous missing values since the replaced values can lead to misleading extreme values in the data aggregation step.

4.2.2. Data aggregation

Another data preparation stage is data aggregation. In this study, the one-minute values are aggregated to one-hour values in order to make variables more interpretable and tangible. As stated earlier, motion detection data is available for each room of the apartments and shows whether a motion is detected in a room within a one-minute interval. Given the occupancy in the entire apartment matters for the purpose of this study, to hourly aggregate the occupancy variables, it is necessary to sum all the motions detected in the entire apartment within each hour. After aggregation, the hourly values of occupancy account for all the motions detected in the entire apartment at each hour. These hourly values represent the occupancy level (count of motions recorded) within an hour. The number of motions detected in the period between two consecutive hours is assigned to the beginning hour of that period (e.g., the number of motions detected from 1:00 to 1:59 a.m. are summed up, and the obtained value is assigned to 1:00 a.m.). The same aggregation is used for plug and lighting variables, so the sum of values within an hour is calculated and assigned to the beginning hour of each one-hour interval. After this aggregation, it is possible to obtain sequences (time-series) of length 24, which depict changes in hourly count of motions and hourly energy consumption throughout a day. With a simple transformation in the arrangement of data, it is possible to create datasets whose rows represent daily time-series of motion count or energy consumption. In the next steps of the methodology, daily time-series of motion count (i.e., occupancy) are used as input of time-series clustering to find groups of days with occupancy time-series of similar shapes. Plus, daily load profiles are utilized in change point detection (CPD) to find the hours at which a sudden change in the load profile is expected.

4.2.3. Outlier detection

The quartile method is used to detect the outliers. However, some of the plug variables in the data sets are very sparse. As mentioned earlier, these variables represent the amount of energy consumption recoded from the apartments' outlets. It is evident that occupants might occasionally plug appliances into the outlets, or some plugged devices might not have a significant standby energy consumption, which causes sparsity in the respective plug variables. Due to the sparsity in the plug variables, the quartile method is only applied to the non-zero values in order to find the abnormal values. Hence, using equations presented in Appendix G, the plug variables' outliers are

set to zeros. The quartile method is also implemented on hourly motion count values and hourly lighting energy consumption values.

4.3. Time-series clustering analysis

So far in the literature, the partitional clustering algorithms, especially K-means with Euclidean distance, have proved to be very efficient in occupancy and energy profiling. This is while due to its pair-wise distance calculation, K-means has some shortcomings when it comes to measuring the distance between time-series. (Paparrizos & Gravano, 2015) mentioned that in addition to the choice of clustering method (e.g., hierarchical, partitional, spectral, etc.), the choice of distance measure is necessary to achieve higher accuracy and efficiency in time-series clustering algorithms. In the mentioned study, the authors introduced a novel time-series clustering method, called K-shape, which is a domain-independent, accurate, and scalable clustering method. However, the major popularity of the K-shape lies in the fact that this method is able to properly address the distortions and temporal variations in time-series when measuring the similarity among time-series.

4.3.1. Distortions in time-series

In time-series clustering, it is essential to mitigate the impact of distortions and temporal variations in time-series data so that the comparisons are meaningful. In other words, the clustering algorithm should be invariant towards the distortions. Some of the distortions are as follows:

1. **Translation distortion** is related to the difference in the amplitude of two time-series with similar patterns. Figure 4-2 (a) shows that while the illustrated time-series have similar patterns, one of them has a bigger amplitude.
2. **Noise distortion** happens when there are noises in time-series' values, while time-series' overall patterns are similar. Noise invariance is a characteristic of clustering algorithms to handle noises and measure similarity in time-series regardless of noises (see Figure 4-2 (b)).
3. **Shift distortion** is associated with the difference in phases of two time-series with similar overall patterns; time-series can be of the same pattern, but one can be slightly shifted in time (see Figure 4-2 (c)).

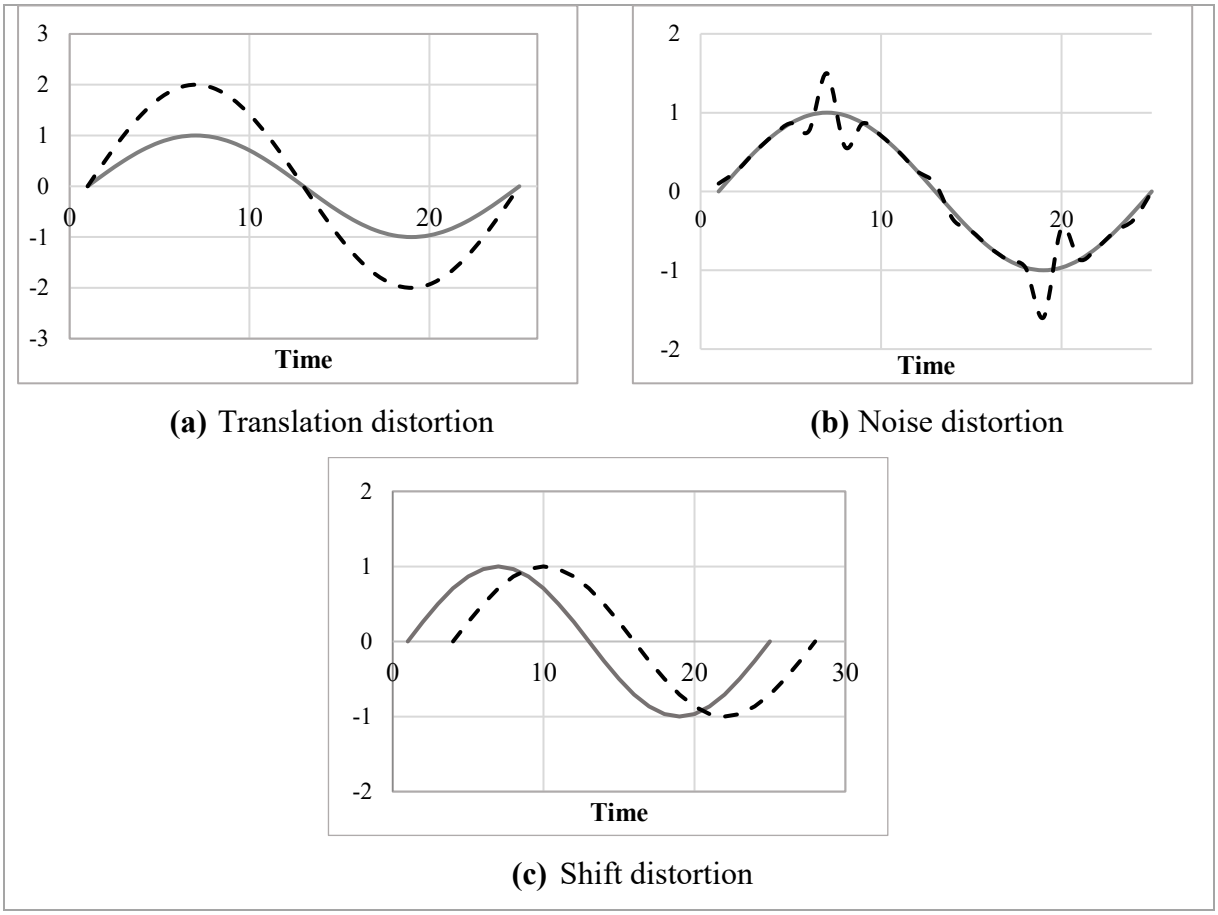


Figure 4-2. Different types of distortions in time-series data

Some of the intrinsic distortions in time-series data can be eliminated using the normalization of time-series. Z-normalizing is the most common normalization method used for time-series clustering, which also helps achieve translation-invariance (Paparrizos & Gravano, 2015). Hence, it is vital to z-normalize input time-series before comparing the similarity between them. Z-normalization is applied on values of each time-series separately, so for a time-series of length n the z-normalization equation is as follows:

$$z_i = \frac{x_i - \mu}{\sigma}, \qquad for\ \ i = 1, 2, \dots, n. \qquad (Equation 1)$$

In case of daily time-series used in this study, i ranges from 1 to 24. Nonetheless, z-normalizing time-series solely cannot provide the area for meaningful comparison of time-series. The choice of distance measure and centroid calculation should also be accounted for to make robust comparisons. K-shape method, which is shift-, translation-, and complexity-invariance, can

compare time-series properly and handle the temporal variation by satisfying the mentioned invariances (Paparrizos & Gravano, 2015).

4.3.2. K-shape clustering procedure

Similar to the well-known K-means, K-shape uses the iterative assignment and refinement procedure to group time-series with similar shape (pattern) in distinctive clusters. Accordingly, in the K-shape algorithm, after the initial random selection of k clusters, the cluster centroids will be computed. Once the initial centroids are obtained, each time-series will be assigned to the closest centroid. The assignment step depends on the shape-based distance measure computation (Equation 3). Then, in the refinement step, the new centroids' position will be determined based on the new cluster members using Equation 4. The assignment and refinement steps will be repeated until either (1) no changes occur in cluster membership of time-series or (2) the algorithm reaches its iteration limit, which can be predetermined by the user. Both assignment and refinement procedures in K-shape clustering relies on the normalized cross-correlation (NCC) computation (Equation 2). The nature of cross-correlation is similar to the convolution of two functions. Cross-correlation ($cc_w\ (\vec{x}, \vec{y})$) calculates the inner-product of two sequences in the way that one sequence remains static and the other one slides over it; meanwhile, the inner-product of each slide is calculated. The slide with the highest inner product gives the alignment of two time-series, where the similarity between the two time-series is at its peak. Therefore, if two time-series with length m are considered ($x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$), w is defined as a set like $\{1, 2, \ldots, 2m-1\}$ which shows a slide. The goal of NCC alignment is to find the one w that represents the position of the slide with the highest cross-correlation value (see Equation 2).

$$NCC = \frac{CC_w(\vec{x}, \vec{y})}{\sqrt{R_0\ (\vec{x}, \vec{x})\ .\ R_0\ (\vec{y}, \vec{y})}} \qquad (Equation 2)$$

where $R_0(\vec{x}, \vec{x})$ and $R_0(\vec{y}, \vec{y})$ are the autocorrelations of sequences x and y, or in other words, cross-correlation of a single sequence with itself. To measure the similarity between time-series, K-shape uses shape-based distance (SBD), which also works based on cross-correlation computation. In SBD computation, the maximum value of NCC between the two time-series is considered:

$$SBD(\vec{x},\vec{y}) = 1 - max_w \left(\frac{CC_w(\vec{x},\vec{y})}{\sqrt{R_0(\vec{x},\vec{x}) \,.\, R_0(\vec{y},\vec{y})}}\right) \qquad (Equation 3)$$

As mentioned earlier, centroid computation also depends on the NCC calculation. In general, the objective of centroid computation in partitional clustering methods is to find the position where the sum of squared distances between all time-series and the centroid position is minimized. In other words, the general approach of partitional clustering is to find the position of the centroids where the dissimilarities among all sequences (time-series) of a cluster and their respective centroids are minimized. However, the nature of cross-correlation is to measure similarity rather than dissimilarity (Paparrizos & Gravano, 2015) since in cross-correlation calculous, one time-series is sliding over the other one, and the inner product of each slide is calculated. Therefore, the maximum cross-correlation value indicates the highest similarity between two time-series. Hence, the configuration with the maximized cross-correlation shows the highest similarity. The optimization problem for centroid computation in K-shape is defined as Equation 4; the aim is to compute the centroid sequence (time-series) for which the sum of squared NCCs from the centroid to all other sequences assigned to that centroid is maximized. Since k-shape relies on NCC to extract centroid, the centroid computation process is called shape-extraction (Paparrizos & Gravano, 2015), and it is calculated as follows:

$$\overrightarrow{\mu_k^*} = \underset{\overrightarrow{\mu_k}}{argmax} \sum_{\overrightarrow{x_i} \in P_k} NCC(\overrightarrow{x_i},\overrightarrow{\mu_k})^2 = \underset{\overrightarrow{\mu_k}}{argmax} \sum_{\overrightarrow{x_i} \in P_k} \left(\frac{max_w\, CC_w(\overrightarrow{x_i},\overrightarrow{\mu_k})}{\sqrt{R_0(\overrightarrow{x_i},\overrightarrow{x_i}) \,.\, R_0(\overrightarrow{\mu_k},\overrightarrow{\mu_k})}}\right)^2 \quad (Equation 4)$$

where $P = \{p1, \ldots, pk\}$ represents the k disjoint clusters, so k is the number of clusters, and $\overrightarrow{\mu_k^*}$ represents the centroid time-series (centroid sequence).

Figure 4-3 demonstrates the modifications applied to time-series after NCC-based alignment. This alignment reflects the situation where the inner-product of the two time-series is maximized.

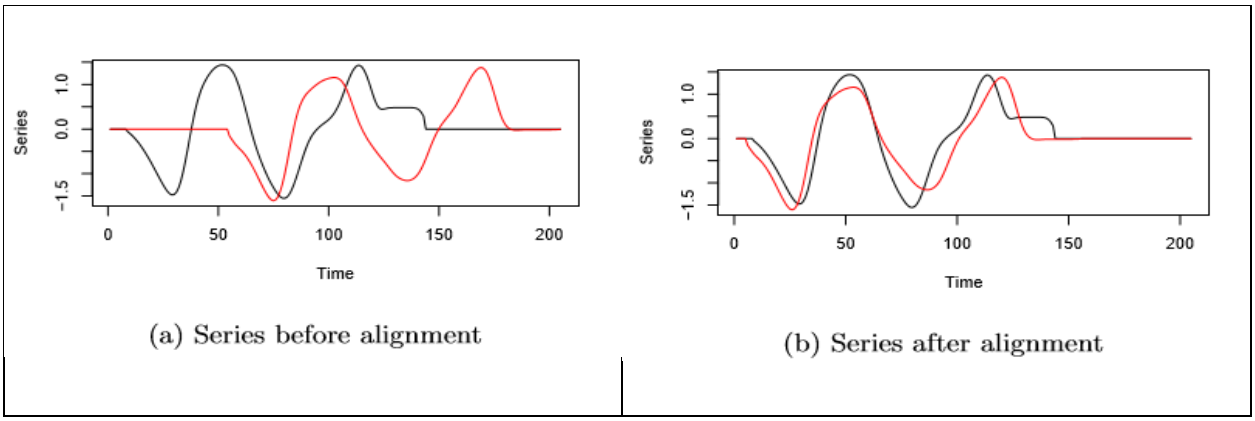


Figure 4-3. NCC-based alignment, adopted from (Sardá-Espinosa, 2019)

4.3.3. Cluster validation

To determine the most optimum number of clusters that achieves high intra-cluster and low inter-cluster similarity, the "cvi" function in the "dtwclust" package is utilized. This function provides a variety of cluster validation indices, namely internal and external indices. External clustering validation indices evaluate the clustering results based on externally provided labels or "ground truth". Readers are referred to (Aghabozorgi et al., 2015) for more information on ground truth and external indices. Since the "ground truth" is not available in this study and the purpose is to find the one clustering arrangement which is best fitted to the provided data, internal indices are used to find the most optimum number of clusters (k). Generally, internal indices compare the goodness of fit between the obtained clustering results from different k values. Internal indices work on the basis of inter-and intra-cluster distances, so the most optimal clustering result has the low intra-cluster distances and high inter-cluster distances. To discover the optimum number of clusters, the majority vote of several indices is taken when these indices are measured for $2 \leq k \leq 10$. The validation indices considered for this purpose are Silhouette (Rousseeuw, 1987), Calinski-Harabasz (Caliński & Harabasz, 1974), Davis-Bouldin (Davies & Bouldin, 1979), Dunn (Dunn, 1973); please check Appendix F for equations of the mentioned indices.

In this study, K-shape is applied on daily time-series of occupancy (as mentioned earlier, these time-series indicate the count of motions detected at each hour of a day). The outputs of the K-shape are clusters grouping days with similar occupancy profiles, and each cluster contains days during which the presence patterns of the household are very similar. In the next step, another

time-series analysis method is implemented to find the hours when the significant changes in energy consumption are probable during the days with similar occupancy patterns.

4.4. Change point detection (CPD)

Changepoint detection, also known as change point inference, enables detecting segments of a signal (i.e., time-series) where the statistical properties like mean or variance are significantly different from their adjacent segments. In general terms, this method can detect segments of a signal for which the probability distribution functions are different from each other. There are a variety of change point detection algorithms for various applications and analyses. For instance, after implementing an energy efficiency measure, (Touzani et al., 2019) sought changes in both mean and variance of daily energy consumption to find the non-routine consumption events. On the other hand, (Pereira & Ramos, 2018) only considered changes in the mean of environmental parameters (e.g., CO_2, RH, etc.) to detect the timing of occupants' action (e.g., cooking, showering, heating, etc.). Figure 4-4 (a) and (b) show the change points detected in mean and variance of time-series values, respectively (the solid orange lines depict changepoints). In general, change point detection algorithms are designed to test time-series regarding the existence of changes in their patterns and, in case of change detection, reveal the location of changes (Chen & Gupta, 2012).

4.4.1. Statistical hypotheses of change point detection

According to (Chen & Gupta, 2012), if $x_1, x_2, \dots, x_n$ is a time-series of independent random variables, and $F_1, F_2, \dots, F_n$ are the respective probability distribution functions, the change point detection problem is to test the null hypothesis:

$$H_0 : F_1 = F_2 = \cdots = F_n \qquad (Equation 5)$$

against the alternative hypothesis:

$$H_1 : F_1 = \cdots = F_{k_1} \neq F_{k_1+1} = \cdots = F_{k_2} \neq F_{k_2+1} = \cdots = F_{k_q} \neq F_{k_q+1} = \cdots = F_n \quad (Equation 6)$$

In Equation 6, q is the number of change points, and $k_1, k_2, \dots, k_q$ are the location of them. It is also possible that the distributions $F_1, F_2, \dots, F_n$ belong to a common parametric family $F(\theta)$, where $\theta \in R^p$; therefore, the previous hypotheses can be translated to:

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_n = \theta \qquad (Equation 7)$$

$$H_1 : \theta_1 = \cdots = \theta_{k_1} \neq \theta_{k_1+1} = \cdots = \theta_{k_2} \neq \theta_{k_2+1} = \cdots = \theta_{k_q} \neq \theta_{k_q+1} = \cdots = \theta_n \ (Equation 8)$$

For instance, if the change point problem is to find changes in the mean value of a normally distributed time-series like x_i, where $i = 1, 2, \dots, n$, the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n = \mu \qquad (Equation 9)$$

Should be tested against the alternative:

$$H_1 : \mu_1 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n \qquad (Equation 10)$$

It should be mentioned that the hypothesis (Equation 10) only deals with detecting a single change point, and k is the unknown location of that change point. The process of testing the existence of one change point against zero change point is applied in one of the most popular changepoint methods, called "binary segmentation". Binary segmentation can be summarized in three stages (Chen & Gupta, 2012). In the first stage, the entire time-series (sequence) will be tested for the existence of a single change point, and if no change is detected and the null hypothesis (Equation 9) is accepted, the process will be stopped at this stage. But if the null hypothesis is rejected (or H_1 (Equation 10) is accepted), it means that one change point is detected, and its location will be revealed. The algorithm will then move on to the second stage, considering that the initial time-series is already divided into two segments (two time-series). In the second stage, each segment will be tested with the same set of hypotheses as the first stage (Equations 9 and 10), to find a single change within each segmented time-series. This process will continue until no change is detected in any subsegments and no null hypothesis is rejected. At the end of the binary

segmentation process, the estimated number (q) and locations $(k_1, k_2, \dots, k_q)$ of all change points in the initial time-series will be obtained. The popularity of binary segmentation method is due to the simultaneous detection of the number of change points and their location, so it is computationally fast. However, this fast computation introduces some level of accuracy loss as well (Touzani et al., 2019).

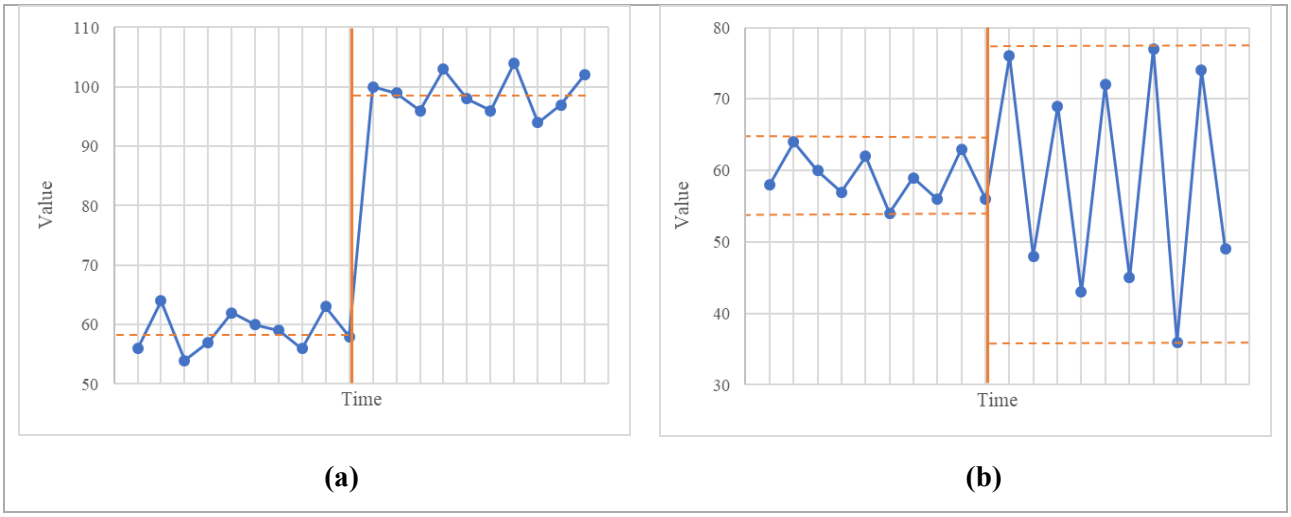


Figure 4-4. Change point detected in (a) mean and (b) variance

The question that remained unanswered so far is about the criterion that determines whether the null hypothesis is rejected or accepted. This criterion also determines the changepoints location at each step of the binary segmentation process. Information criteria are some of the most common approaches applied to address the mentioned issues. In this study, Schwarz Information Criterion (SIC), introduced by (Schwarz, 1978), is applied to the change point detection process.

To clarify the role of SIC in changepoint detection, (Chen & Gupta, 2012) brought an example of change point detection in the variance of time-series values:

> Consider a time-series like $x_1, x_2, \dots, x_n$ consisting of independent, identically distributed random variables. The purpose is to detect changes in the variance (Figure *4-4* (b)) within the time-series x_i, where $i = 1, \dots, n$. Since changes in the mean are not desired in this example, this value is considered as a constant (denoted by μ) for all the subsegments of the time-series x_i. In this case, $SIC(n)$ denotes the model for the null hypothesis, where no

change in the variance can be observed within the time-series x_i. Therefore, $SIC(n)$ can be defined as follows:

$$SIC(n) = n\,log\,2\pi + n\ log\,\hat{\sigma}^2 + n + log\,n \qquad (Equation 11)$$

- $\hat{\sigma}^2 = {(\sum_{i=1}^{n}(x_i - \mu)^2)}/{n}$ is the maximum likelihood estimator of σ^2 under null hypothesis H_0.

It should be noted that there is one possible value for $SIC(n)$, which is determined by Equation 11. However, $(n-3)$ SICs can be calculated under the hypothesis H_1 denoted by $SIC(k)$, where k ranges from 2 to $n-2$. $SIC(k)$ is defined as Equation 12:

$$SIC(k) = n\,log\,2\pi + k\ log\,\widehat{\sigma_1}^2 + (n-k)\,log\,\widehat{\sigma_n}^2 + 2\,log\,n \qquad (Equation 12)$$

- $\widehat{\sigma_1}^2 = {(\sum_{i=1}^{k}(x_i - \mu)^2)}/{k}$ and $\widehat{\sigma_n}^2 = {(\sum_{i=k+1}^{n}(x_i - \mu)^2)}/{(n-k)}$ are the maximum likelihood estimators of σ_1^2 and σ_n^2, under the hypothesis H_1, respectively.

Accordingly, calculation of maximum likelihood estimator is only available for points of time-series located between 2nd and $(n-2)^{th}$ positions. Therefore, based on the information criterion principle, the null hypothesis (H_0) is accepted if:

$$SIC(n) < \min_{2 \le k \le n-2} SIC(k) \qquad (Equation 13),$$

and is rejected (i.e. H_1 is accepted) if:

$$SIC(n) > SIC(k), \qquad 2 \le k \le n-2 \qquad (Equation 14).$$

Evidently, the changepoint position is determined based on the one k that minimizes $SIC(k)$ (Equation 12).

For more details about the function of information criteria and calculous of maximum likelihood, the readers are referred to (Chen & Gupta, 2012).

In this study, the binary segmentation method coupled with the SIC penalty is utilized to detect changes in the daily time-series of energy consumption. Changes in the mean energy consumption are considered so that the segments of time-series with high and low consumptions are separated

from one another. Figure 4-5 demonstrates an example of change point detection in the mean value of energy consumption. Based on the four change points detected at 6 am, 8 am, 6 pm, and 8 pm in the daily time-series, five periods with different mean energy consumption can be found. It should be mentioned that the energy consumption values of hours detected as change points have closer values to their previous hours. So, the hours at which a change has occurred will be in the same period as their respective previous hours. Consequently, the identified periods are as follows: period 1: $[12\ am - 6\ am]$, period 2: $[7\ am - 8\ am]$, period 3: $[9\ am - 6\ pm]$, period 4: $[7\ pm - 8\ pm]$, period 5: $[9\ pm - 11\ pm]$. It should be mentioned that each period expands from the first minute of the starting hour to the last minute of the ending hour. So, period $[12\ am - 6\ am]$ is equivalent to $[12{:}00 - 6{:}59]$.

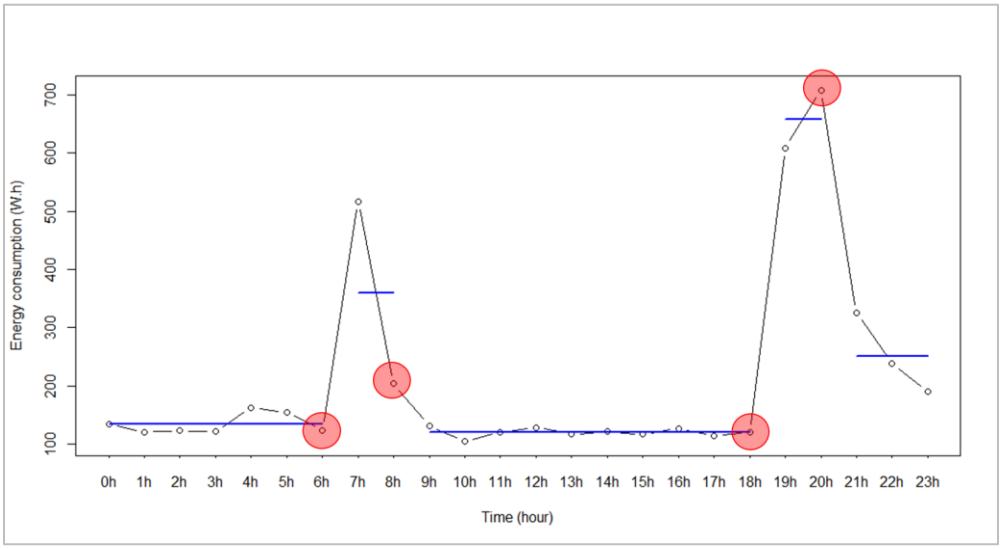


Figure 4-5. Change points detected in mean electricity consumption. The hours indicated by red circles are the points of time-series where changes are detected, and the blue lines depict the mean values throughout the periods specified by change points

4.4.2. Detection of usual routines of households using CPD

In this study, the purpose of CPD is to find the general consumption routines of occupants and separate the regular high- consumption periods from the regular low-consumption ones. Therefore, implementing change point detection on a single load profile is not intended since it cannot deliver the frequent change hours across several days. With the aim of discovering occupants' general consumption routine, in this study, change point detection is applied in the same way as used by (Li, Panchabikesan, et al., 2019). Accordingly, the maximum number of change points (Q) are

primarily determined by trying several values for Q and observing the results. When the limit for the maximum number of change points (Q) increases, the algorithm might detect insignificant changes in the mean value that are not desirable. Since the purpose of CPD analysis in this study is to distinguish regular high consumption hours from low consumption periods, it is not desirable to find the points of time-series (hours) with insignificant changes in mean energy consumption.

In the previous step of the methodology framework, the days during which households follow a similar occupancy schedule are grouped using K-shape clustering. In this step, to separate the regular high-consumption periods from regular low-consumption periods within each occupancy cluster, CPD is applied to the daily load time-series of each cluster. A set of change points (hours) is obtained for each day. Eventually, the number of times a change point is detected at each hour is counted to discover the most probable hours for change point occurrence in each occupancy cluster. As a result, within each occupancy cluster, the relative frequency of change point occurrence can be calculated for each hour. For example, if the relative frequency of 7 am in occupancy cluster_n is 0.5, it means that in 50 percent of the days clustered in occupancy cluster_n, a change is observed in the daily load profiles at 7 am. Based on the calculated frequencies, a threshold must be selected to determine the hours when the possibility of change point occurrence is relatively high. Obviously, as the Q increases, the limit that determines the most frequent change points (i.e., hours with a high possibility of change point occurrence) should be increased as well. In this study, based on the observed results, $Q = 5$, and in some cases, $Q = 6$ are suitable as the maximum number of change points. Therefore, these Qs can sufficiently capture the hours where mean energy consumption is probable to change significantly. Based on the results obtained from selected Qs, the proper limit for change point frequency in each occupancy cluster is determined as 0.4. Therefore, the hour detected as a change point in at least 40 percent of days within an occupancy cluster is considered the frequent change point in that cluster. Using the frequent change points (i.e., most probable change hours) in each occupancy cluster makes it possible to find the periods during which the mean consumption value is significantly higher or lower than the adjacent periods. These periods will give us an insight into the households' routine in case of energy consumption. In the next step, the factors that explain the consumption behavior of occupants within high- and low-consumption periods are determined.

4.5. Statistical analysis with LASSO regression

At this stage, a regression model will be trained for each period to determine what types of activities (indicated with plugs located in different rooms of the apartments) have noticeable impacts on the total load within each period. The statistical analysis of this step helps identify the comparative contribution of occupants' activities to the energy consumption within each period. Variable selection is the key role of regression models in energy studies. Based on the characteristics of data, the variable selection task can be faced with some challenges. One of these challenges is due to the high dimensionality of the dataset. The high dimensionality happens when the number of predictors is higher than the number of training examples (i.e., the number of rows or data points to train the regression model). This issue raises the possibility of overfitting, which can impair the interpretability of the trained model (Satre-Meloy et al., 2020). Overfitting is a product of complex models that perfectly fit the training set while failing to generalize to new data (testing set). In other words, although the overfitted model outperforms on the training set, when new data is introduced, the performance of the model decreases noticeably. Another challenge is associated with the sparsity of predictors (i.e., variables which contain many zeros). Multicollinearity is one of the possible products of sparsity in the predictors' matrix, and it occurs when a predictor can be linearly explained by other predictors. Multicollinearity can introduce instability to the model so that small changes in data or the model can lead to noticeable changes in model parameters (Satre-Meloy et al., 2020). Although multicollinearity might not affect the model's accuracy, it influences model coefficients. So, the delivered variable importance can be misleading if multicollinearity exists among predictors. Regularized regression can be the solution to overcome the challenges mentioned above. In this study, plug variables and total lighting load are used to predict total energy consumption using regression analysis. Some of the plugs might have been occasionally used by occupants or are associated with appliances that do not constitute the base load consumption of the household (i.e., has lots of zeros in their consumption pattern), so sparsity in predictors matrix can be an issue in the variable selection task. On the contrary, some plug variables are associated with base load consumption of households (e.g., fridge, freezer, and appliances with significant standby energy consumption like TV); therefore, these variables always have a value and are not sparse. Considering the different characteristics of the predictors, regularization is applied to linear regression, to be able to find the comparative contribution of all predictors with different characteristics. The penalized regressions are designed to avoid

overfitting in the way that they introduce some level of bias[1] to the best-fitted model to the training set in order to achieve lower variance[2] in the performance of the model when it is applied on the testing set. Another term for regularization is penalized regression, as these models add a penalty term to the ordinary least square (OLS) optimization problem (Equation 15). By doing so, the predictors' coefficients will be shrunk to zero or near-zero values, and this shrinkage is more sever for irrelevant predictors rather than influencing ones. Due to this shrinkage, the obtained model from penalized regression will be more interpretable than the output of simple regression models. This quality of penalized regressions makes them a suitable choice in case of variable selection purposes. The OLS regression problem is to estimate coefficient values (β_j) that minimize the residual sum of squares (RSS) (see Equation 15):

$$RSS(\beta) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \qquad (Equation 15)$$

Where p is the number of predictor variables, x_{ij} denotes the j^{th} predictor value for the i^{th} observation, and y_i is the target value for the i^{th} observation. As mentioned earlier, penalized regressions change the OLS optimization problem (Equation 15) by adding a penalty to the equation. Here two of the most popular regularization methods, naming RIDGE (Hoerl & Kennard, 1970) and LASSO (Robert Tibshirani, 1996), and their shortcomings and advantages are discussed. Equation 16 is the optimization equation of RIDGE regression:

$$\underset{\beta}{argmin}\left(RSS(\beta) + \lambda\sum_{j=1}^{p}\beta_j^2\right) \qquad (Equation 16)$$

[1] In machine learning concept, when a model fits perfectly to the training set so that it can properly capture the distribution of training data, it can be said that the model has low bias.

[2] In machine learning, variance is associated with the difference in fits between training and testing set. Meaning that, when the performance of a model on training data is close to its performance on unseen (i.e., testing) data, it can be said that the model has low variance.

where $\lambda \geq 0$ is a constant value that controls coefficient shrinkages (Satre-Meloy, 2019). Increasing the penalty term of RIDGE regression (i.e., $\lambda \sum_{j=1}^{p} \beta_j^2$) leads to shrinkage of model coefficients; however, not all the way to zero. On the other hand, the penalty used in LASSO regression (i.e., $\lambda \sum_{j=1}^{p} |\beta_j|$) can set variable coefficients to zero and make the final coefficients more distinguishable regarding their importance in the prediction task (see Equation 17); the optimization problem of LASSO regression is to minimize the Equation 17. Therefore, when the aim is to obtain sparsity in results and yield more interpretable models, the LASSO penalty is a more desirable choice than the RIDGE penalty (James et al., 2013) (Satre-Meloy, 2019).

$$\underset{\beta}{argmin}\left(RSS(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right) \qquad (Equation 17)$$

As mentioned earlier, increasing λ is equivalent to the increase of penalty, which results in higher bias to the model. Regarding penalized regressions, it is necessary to check how increasing the penalty impacts the trade-off between bias and variance. In other words, it is crucial to check how much bias can be tolerated by the model without sacrificing the variance. Therefore, finding the optimum λ that achieves high accuracy when applied to the new data (i.e., achieves low variance) is necessary. In other words, the purpose here is to increase the bias (by increasing the penalty) as long as the variance improves and is not impaired. Cross-validation is frequently used to determine λ in the way that it calculates mean squared error (MSE) for a range of λ values to find the one λ that achieves the least prediction error. For each λ value, the k-fold cross-validation uses a certain proportion of data to train a regression model and then test the obtained model on the remaining data. If the number of folds (k) is equal to 10, this process will be repeated ten times, so 90 percent of data will be used for training and 10 percent for testing, and eventually, the average of the 10 MSEs will be calculated. The k-folds cross-validation process is repeated for a range of λ values so that the λ that gives the lowest average of MSEs is chosen as the most optimum.

Out of the two most popular regularization methods (i.e., RIDGE, LASSO), LASSO is selected because one of the purposes of this study is to discover the most influencing factors of energy consumption, and the interpretability of the model is essential for the analyses. LASSO regression

can yield more sparse models as it set variable coefficients to zero and makes the final coefficients more distinguishable regarding their importance in the prediction task. Therefore, it can achieve more interpretable models by shrinking the insignificant variables' coefficient to zero. The aim of using penalized regression in this study is to avoid overfitting, enable identifying the comparative contribution of several plug variables with different energy consumption patterns, and perform variable selection that leads to understanding the influencing factors of energy usage in residential apartments.

4.5.1. Data preparation for LASSO regression

It should be mentioned that min-max normalization is applied to variables before training the regression models. This transformation allows for a more improved evaluation of variable importance through the estimated coefficients of the model (Ren et al., 2019). Accordingly, the values of all variables are scaled between 0 and 1. The min-max normalization formula is shown in Equation 18:

$$x_i{}' = \frac{x_i - min(x)}{max(x) - min(x)} \qquad (Equation 18)$$

For a vector like x_i, where $i = 1, \dots, n$, $x_i{}'$ is the normalized value of the ith element from vector x_i. $min(x)$ and $max(x)$ are the minimum and maximum values in the vector x_i.

4.5.2. Evaluating regression models

Regarding the evaluation of regression models, 80 percent of the data is used for training the models, and the remaining 20 percent is considered for testing the trained model. The evaluation measure applied to the testing set is R^2 (R-squared). Therefore, first, the model is trained using the training set, and the obtained model is implemented on the unseen, testing data, and the R^2 demonstrates how well does the trained models can explain the variations in the target variables of the testing set. The higher the R^2, the better the goodness of fit. R^2 is determined using Equation 19 and ranges from 0 to 1 inclusive.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(\overline{y} - y_i)^2} \qquad (Equation 19)$$

where RSS is the residual sum of squares, and TSS is the total sum of squares. y represents the target value, so y_i is the actual target value of the ith observation, $\hat{y}_i$ is the predicted value for the ith observation, and $\overline{y}$ is the average of all observations' actual values.

Chapter Five

5. Results and Discussions

5.1. Data preparation

After removing days containing continuous missing or dead values and excluding days with zero occupancy, the final datasets of apartments 112, 152, and 162 include 271, 271, and 216 days, respectively. As mentioned earlier, to have an assumption about the location of plug variables, correlation analysis is employed to find the relations between the count of motions in different rooms and plug loads. So, it is assumed that if the consumption of a plug demonstrates correlations with the count of motions recorded in a specific room, that plug sensor is probably located in the same room as the motion sensor. This assumption is also used in (Li, Panchabikesan, et al., 2019). After discovering the associations between plug variables and motion variables in different rooms of each apartment, every plug variable can be seen as a zone-related activity. Therefore, plug variables can give us an insight into occupants' activities in different rooms of the apartments. For example, based on the correlation analysis among motion and plug variables in apartment 152 (Figure *1-1*Figure 5-1), the plug number 16 (P16) has the highest Pearson's coefficient with living room motions. It can be interpreted that when occupants are active in the living room, they use appliance(s) plugged into the plug number 16. Although plug 16 also has a high correlation with motions detected in the kitchen, only the highest correlation value found for each plug variable is considered to label the plug. Accordingly, the labels of each plug variable are shown in Table *5-1*. Please check 0 for the results of correlation analysis in apartments 112 and 162. It should be mentioned that for some of the apartments, correlations only appear when the aggregation level is increased to an hourly level. Thus, aggregation is essential in order to find trends and patterns in the data. For the same reason, hourly aggregation is frequently practiced in most of the previous studies, and in this study, the same level of aggregation is utilized.

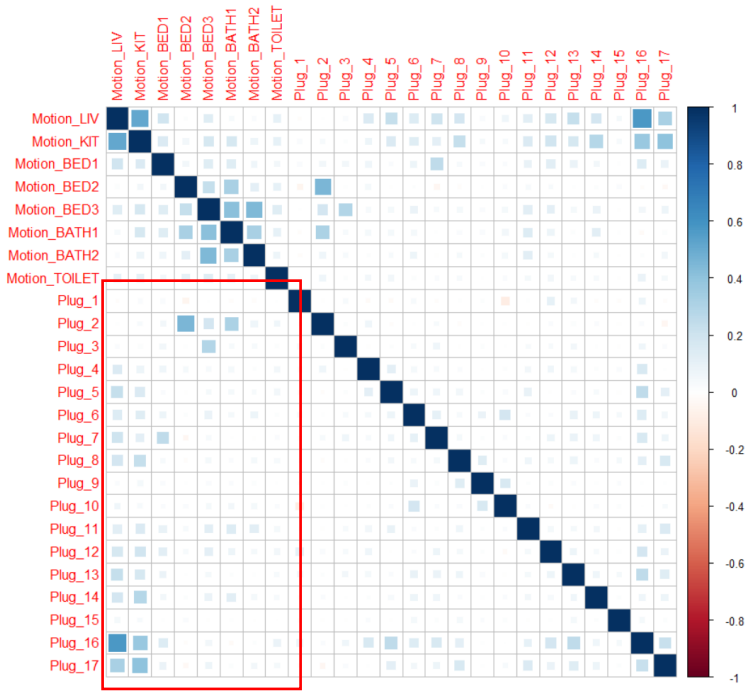


Figure 5-1. Pearson correlation coefficients between plug variables and motion detection variables of apartment 152

Table 5-1. Zonal labels of plug variables in apartment 152

Zonal labels of the apartment 152	**Plug Power Variables**
Bedroom-related	P2, P3, P7
Livingroom-related	P4, P5, P13, P16
Kitchen-related	P8, P14, P17
Bathroom-related	-
Others	P1, P6, P9, P10, P11, P12, P15

After plotting hourly aggregated values of plug variables, three types of consumption patterns could be observed:

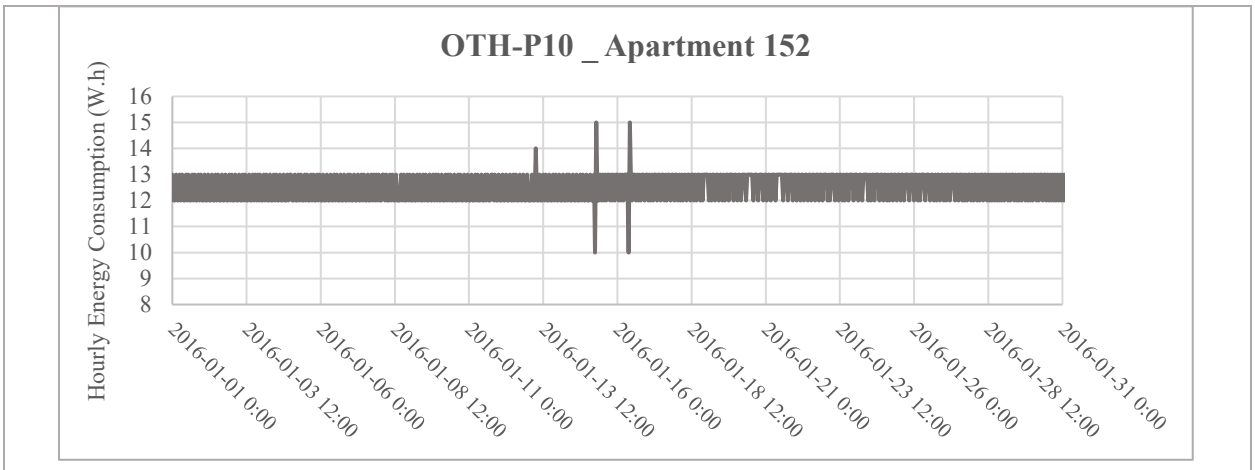
1. The first category indicates the consumption pattern of appliance(s) that are constantly consuming energy at each hour with negligible fluctuations in their consumption value (Figure 5-2 (a)). The consumption of these plugs forms a proportion of each household's base load[3], so this consumption exists during sleeping, or unoccupied periods as well as active, occupied hours. As shown in Figure 5-2 (a), the plugs having the first category of

[3] The base load is associated with standby load of electric appliances (e.g. TV, set-up boxes, etc.), or consumption of the appliances that are always working (e.g. fridge, freezer, etc.).

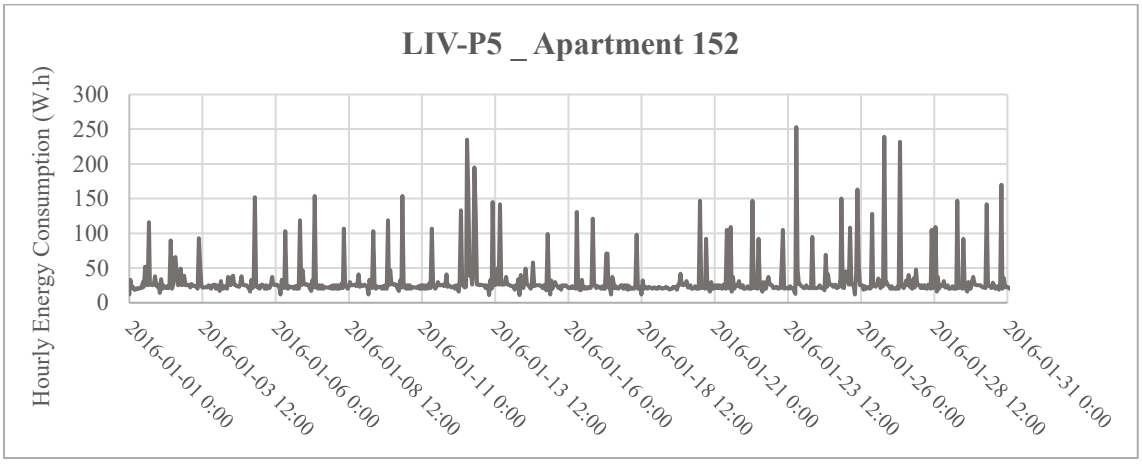
consumption pattern are not affected by occupants' presence and activities since the pattern of consumption has not been changed significantly during the one-month period shown in Figure 5-2 (a). Examples of the appliances belonging to this category are internet modem, cordless phone etc. (Standby Power Summary Table, 2018). Accordingly, this category of plug variables has low standard deviations (SD) (i.e., below $SD = 2$), and the mean and three percentiles (25^{th}, 50^{th}, 75^{th} percentiles) of these variables are very close to each other (see Table 5-2 and Appendix B).

2. Figure 5-2 (b) demonstrates the consumption pattern of electric appliance(s), which constitute a proportion of the household's base load and have high fluctuations in their consumption pattern. The appliances which belong to this category are desktop computer, TV, etc. (Standby Power Summary Table, 2018). This category of plug variables is characterized by approximately high mean and SD values, and the difference between the three percentiles (25^{th}, 50^{th}, 75^{th} percentiles) is more noticeable compared to the other two categories of plugs (see Table 5-2 and Appendix B).
3. The third category of plug variables represents the consumption pattern of electric appliance(s) that are only plugged-in occasionally or has negligible standby power, such as hairdryer, electric kettle, dishwasher etc. (Standby Power Summary Table, 2018). These plug variables are very sparse, meaning that they contain zero values most of the time (Figure 5-2 (c)). Correspondingly, the mean of these variables is relatively small, while the SD is usually high. Another common characteristic of these variables is that their 25^{th}, 50^{th} (median), and 75^{th} percentiles are negligible (i.e., usually zero). This fact emphasizes the sparsity in 3^{rd} category of plug variables (see Table 5-2 and Appendix B).

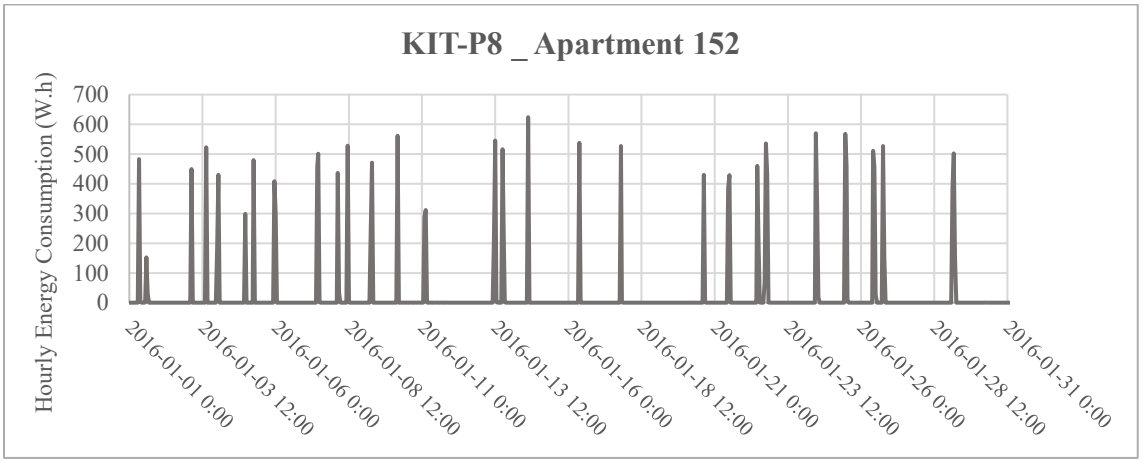
It should be noted that each plug variable can represent the consumption pattern of more than one appliance that is plugged into an outlet, and the names of appliances brought as examples are only mentioned to make the patterns shown in Figure 5-2 more tangible for the readers. It can be seen that sparsity exists among the plug variables (Figure 5-2 (c)). As mentioned in section 4.5, sparsity should be handled in the regression analysis step (step 4 of the methodology).



(a)



(b)



(c)

Figure 5-2. Categories of plug variables based on hourly consumption pattern; (a) constant consumption with negligible fluctuations (b) continuous consumption with noticeable fluctuations (c) sparse consumption

Table 5-2. Summary statistics of variables in apartment 152

Plug Variables in Apt. 152	Mean (W.h)	SD	25th percentile	50th percentile	75th percentile	Category (based on consumption pattern)
OTH_P1	35.71	23.09	20	32	50	2nd
BED_P2	5.59	16.73	0	0	1	3rd
BED_P3	0.02	0.53	0	0	0	3rd
LIV_P4	19.63	87.46	0	0	0	3rd
LIV_P5	40.19	52.62	21	25	29	2nd
OTH_P6	18.12	1.76	17	17	20	1st
BED_P7	1.75	5.67	0	0	0	3rd
KIT_P8	8.80	71.96	0	0	0	3rd
OTH_P9	5.87	0.38	6	6	6	1st
OTH_P10	12.40	0.58	12	12	13	1st
OTH_P11	2.16	11.92	0	0	0	3rd
OTH_P12	0.26	1.00	0	0	0	3rd
LIV_P13	32.81	14.32	23	30	37	2nd
KIT_P14	3.03	12.62	0	0	0	3rd
OTH_P15	0.02	0.41	0	0	0	3rd
LIV_P16	34.12	41.35	6	12	72	2nd
KIT_P17	21.11	95.95	0	0	0	3rd
LIGHTS	19.39	43.86	0	1	14	-

5.2. Time-series clustering results

In this section, the cluster validation results are first presented to discover how many clusters can extract all the distinctive occupancy patterns in each apartment. Secondly, the results of K-shape clustering are brought and discussed for each apartment.

5.2.1. Cluster validation results

For this study, four cluster validation indices (CVIs) are evaluated for $k = 2$ to $k = 10$ clusters. To determine the optimum number of clusters, Silhouette, Calinski–Harabasz, and Dunn Index should be maximized, while Davis–Bouldin should be minimized (see Table 8-4). The values indicated with the bold font in Table 5-3. Apartment 112, cluster validation indices Table 5-4, and Table 5-5 represent the optimum value of each index. The majority vote of the four indices is selected to determine the best number of clusters capturing all the distinctive occupancy patterns. For example, in apartment 112, three CVIs (Silhouette, Dunn Index, Davis–Bouldin) show that optimum results can be achieved with three clusters while only one CVI (Calinski–Harabasz) votes

for two as the most optimum number of clusters; in this case, $k = 3$ is considered as the optimum number of clusters for apartment 112. Following the same logic, two clusters enable discovering all the distinctive occupancy schedule patterns in apartments 152 and 162 (see Table 5-3. Apartment 112, cluster validation indices, Table 5-4, and Table 5-5).

Table 5-3. Apartment 112, cluster validation indices

Number of Clusters (k)	2	3	4	5	6	7	8	9	10
Silhouette	0.116	**0.128**	0.125	0.108	0.101	0.113	0.102	0.102	0.091
Calinski-Harabasz	**192.960**	88.426	71.914	59.211	61.133	68.143	55.186	36.898	34.111
Davis-Bouldin	2.034	**1.680**	1.717	2.131	1.791	1.788	1.861	1.891	1.806
Dunn	0.102	**0.124**	0.095	0.075	0.064	0.087	0.061	0.088	0.084

Table 5-4. Apartment 152, cluster validation indices

Number of Clusters (k)	2	3	4	5	6	7	8	9	10
Silhouette	**0.251**	0.188	0.190	0.160	0.163	0.122	0.179	0.173	0.132
Calinski-Harabasz	181.174	**200.959**	91.241	98.650	68.619	54.881	57.604	42.607	36.539
Davis-Bouldin	**1.061**	1.525	1.862	2.108	2.213	2.306	1.777	2.012	1.801
Dunn	0.046	0.039	0.045	0.025	0.026	0.051	**0.080**	0.056	0.052

Table 5-5. Apartment 162, cluster validation indices

Number of Clusters (k)	2	3	4	5	6	7	8	9	10
Silhouette	**0.213**	0.211	0.192	0.147	0.124	0.159	0.165	0.143	0.151
Calinski-Harabasz	**234.634**	143.728	79.479	75.909	46.290	57.719	42.846	40.527	40.310
Davis-Bouldin	**1.132**	1.322	1.654	1.420	2.042	1.493	1.523	1.535	1.468
Dunn	**0.128**	0.100	0.084	0.062	0.038	0.040	0.041	0.021	0.017

5.2.2. K-shape clustering results

Based on the optimum number of clusters, K-shape is implemented on daily time-series of occupancy. In clustering, it is commonly regarded that discovering clusters containing a small number of data points (in our case, daily time-series of occupancy) can be an indicator of overfitted clustering results or the existence of outliers. Table 5-6 shows that daily time-series in the dataset

of each apartment are almost equally distributed among the occupancy clusters obtained for that apartment.

Table 5-6. Distribution of daily occupancy time-series among the occupancy clusters of each apartment

Apartment ID	Cluster-ID	Number of days
112	Cluster1_Day-time Absence	119
	Cluster2_Mostly Present	82
	Cluster3_Mostly Absent	70
152	Cluster1_Day-time Absence	152
	Cluster2_Mostly Present	119
162	Cluster1_Mostly Absent	112
	Cluster2_Mostly Present	104

Figure 5-3 (a), (b), and (c) show the obtained occupancy patterns for apartments 112, 152, and 162, respectively. Each gray string depicts a daily time-series of motion counts, and the time-series values are z-normalized hourly count of motions. The red string demonstrates the centroid of each cluster. To further evaluate the clustering results and grasp a clearer insight into the obtained patterns, raw values of motion count (i.e., values before z-normalization) are plotted using heatmaps shown in Figure 5-4, Figure 5-5, and Figure 5-6. These figures show the intensity of occupants' motions throughout the days grouped in each cluster. Each row of the heatmaps represents the occupancy profile of a specific day. Therefore, each row contains 24 tiles representing the count of occupants' motions at each hour. Based on the obtained profiles shown in Figure 5-3, the patterns obtained for each apartment are explained to get an insight about each household's presence routines:

Occupancy clusters in apartment 152: Considering occupancy cluster_1 in apartment 152, the number of detected motions usually starts to increase at 7 am (Figure 5-3 (b)), so it can be interpreted that the usual waking-up hour in cluster_1 of apartment 152 is 7 am. This is while the waking-up hour in cluster_2 varies from 7 to 8 am (Figure 5-3 (b)). The schedule pattern of days grouped as cluster_1 (Figure 5-3 (b)) shows that occupants tend to leave the apartment before 9 am on these days (the count of motions decreases after 8 am) and return home around 7 pm. On the other hand, it can be seen in Figure 5-3 (b) and Figure 5-5 (b) in cluster_2 of

apartment 152, occupants are usually present, and the count of motions only decreases from 2 to 6 pm.

Occupancy clusters in apartment 112: the waking-up hour is around 7 to 8 am in all the three obtained clusters (Figure 5-3 (a)). The pattern of cluster_1 in Figure 5-3 (a) shows the household's schedule on days when they are absent from 12 to 7 pm, while in the days grouped in cluster_2, the apartment is usually occupied. On the other hand, cluster_3 is characterized by the high level of motion count during the morning (i.e., 7 to 10 am). In this cluster, occupancy is also recorded during the afternoon and evening hours, but the number of motions is lower than in the morning hours. Figure 5-7 shows that the hourly averaged count of motions from 11 am to 11 pm is smaller than the same values in the period from 8 to 9 am in cluste_3 ("day-time absence").

Occupancy clusters in apartment 162: Occupancy profiles obtained for apartment 162 shows that the usual waking-up hour is around 5 to 6 am (see Figure 5-3 (c)). The daily occupancy patterns of days gathered in cluster_1 of apartment 162 are highly variable. These days are either characterized by presence during the morning or the evening (Figure 5-6 (a)). This type of inconsistency in the period of occupancy presence observed in cluster_1 is due to the fact that K-shape clustering is shift-invariant (see Figure 4-2 in section **Error! Reference source not found.**). Hence, time-series with similar patterns and different phases can be clustered together. The shift-invariant quality of the K-shape method is also in favor of our analysis since it can handle insignificant shifts in phases of time-series with similar overall shapes. However, this algorithm ignores noticeable shift distortions in time-series as well. Therefore, occupancy time-series with different presence periods can be clustered together. Cluster_2, on the other hand, represents days with active occupancy throughout the day, except for sleeping hours (Figure 5-6 (b)).

The results suggest that for a single apartment, K-shape enables discovering occupancy schedule patterns which have utterly distinguishable characteristics from one another. However, some similarities can be observed among the occupancy patterns of different apartments. With the exclusion of sleeping hours, there is an occupancy cluster, in all apartments, that groups days during which the count of motion is usually high; in other words, occupants are mostly present throughout those days (see Figure *5-4* (b), Figure 5-5 (b), and Figure 5-6 (b)). According to the

observed pattern, the name “mostly present” is chosen for these three clusters. It can also be seen that the count of motion ranges between 0 and 150 in apartments 152 and 162, while this value ranges from 0 to 100 in apartment 112 (Figure 5-4, Figure 5-5, Figure 5-6). The reason for this difference lies in the fact that the number of occupants in 152 and 162 are twice as many as the number of occupants in 112. Another similarity in occupancy schedule patterns of different apartments can be found between cluster_1 of apartment 112 and cluster_1 of apartment 152. The mentioned occupancy clusters demonstrate households' presence schedule when they are absent for some hours in the middle of the day. In cluster_1 of apartment 112, the absence period (when the hourly count of motions is zero or near zero) starts from 12 pm and ends at 7 pm (see Figure 5-4 (a)), while this period is longer for cluster_1 of 152 as it starts from 9 am and ends at 6 pm (see Figure 5-5 (a)). According to the timing of absence, the name chosen for these occupancy clusters is “day-time absence”. Additionally, centroids of cluster_3 in apartment 112 (Figure 5-3 (a)) and cluster_1 in apartment 162 (Figure *5-3* (c)) demonstrate the same pattern, and both show a peak in the count of motions in the morning. However, as discussed earlier, in cluster_3 of apartment 112, occupant presence also happens during afternoons and evenings, but the motion count values during these hours are much lower than in the mornings (Figure 5-4 (c)). While in 162, the occupancy presence happened either in the mornings or evenings (Figure 5-6 (a)). Due to the similarity in centroids’ patterns of cluster_3 in apartment 112 and cluster_1 in apartment 162, and few hours of occupancy presence in these occupancy clusters, these clusters are named “mostly absent”.

Occupancy clusters are formed based on the similar daily time-series shapen by hourly occupancy level (count of motions detected within an hour). Although occupancy level cannot represent presence and absence, heatmaps (Figure 5-4, Figure 5-5, and Figure 5-6) can show hours with a relatively low number of movements (below ten motions detected) during which the occupants are either absent or sleep. It is discussed that occupancy clusters are explicit regarding the number of definite occupied hours and the occupancy level fluctuations. In the current study, the purpose of occupancy pattern extraction is only to find the impact of these occupancy patterns on the shape of load profiles, which is discussed in the following section. So, the uniqueness of occupancy clusters from one another and their distinctiveness regarding the shape of daily load profiles fulfill the purpose of the current study, and discussing the probability of presence is not the case here.

Interested readers are referred to (Panchabikesan et al., 2021), which further extract the presence probability profiles from the occupancy patterns obtained from K-shape clustering.

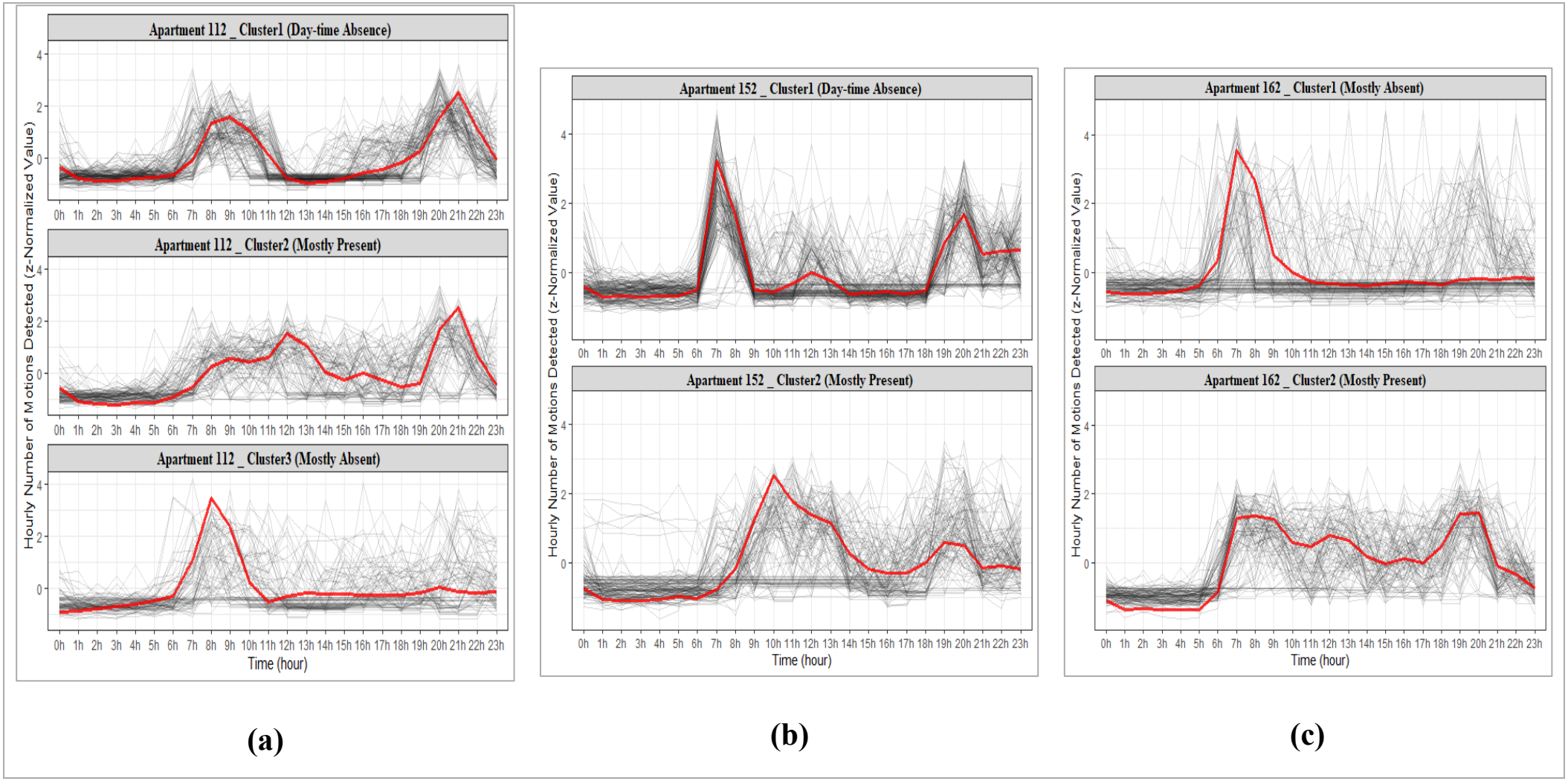


Figure 5-3. Occupancy schedule patterns in apartments (a) 112, (b) 152, and (c) 162; Red lines depict the centroids computed for each cluster; grey lines are the z-normalized daily occupancy time-series.

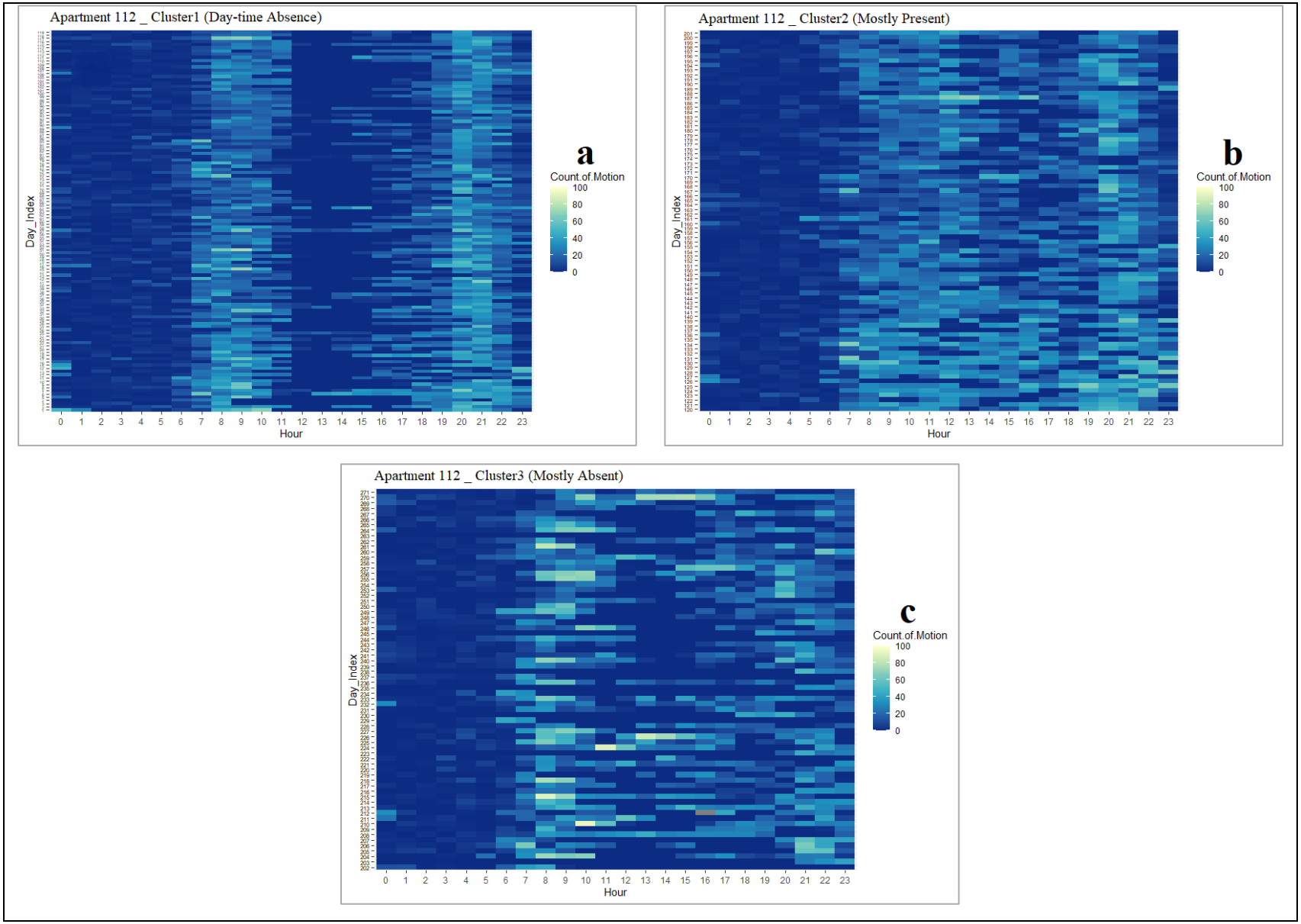


Figure 5-4. Heatmaps of occupancy clusters in apartment 112; each row of heatmaps show a daily occupancy time-series, so each time-series is assigned to a day index and contains the hourly count of motions before z-normalization

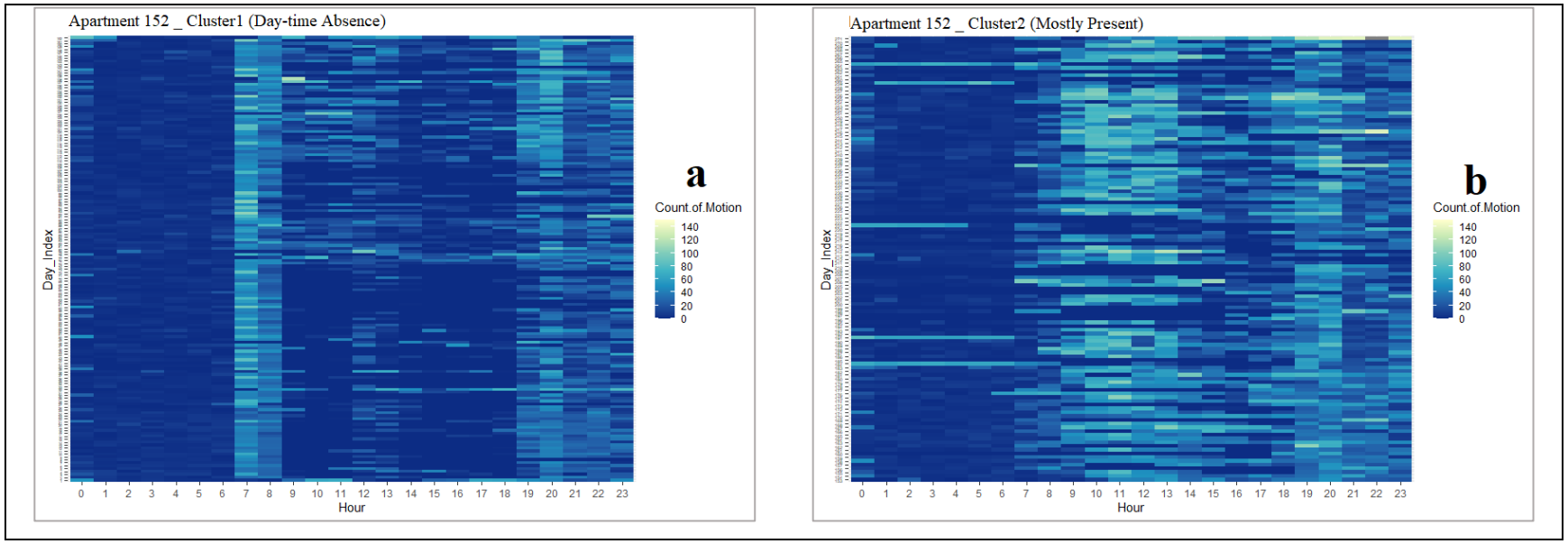


Figure 5-5. Heatmaps of occupancy clusters in apartment 152; each row of heatmaps show a daily occupancy time-series, so each time-series is assigned to a day index and contains the hourly count of motions before z-normalization

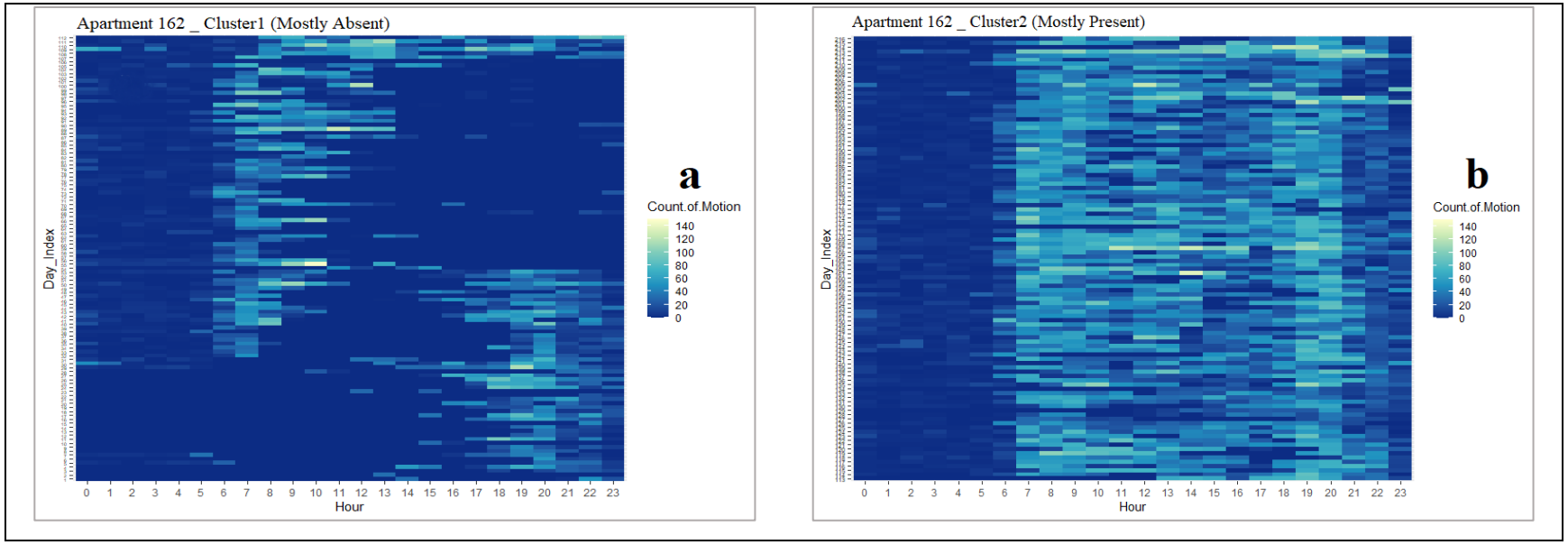


Figure 5-6. Heatmaps of occupancy clusters in apartment 162; each row of heatmaps show a daily occupancy time-series, so each time-series is assigned to a day index and contains the hourly count of motions before z-normalization

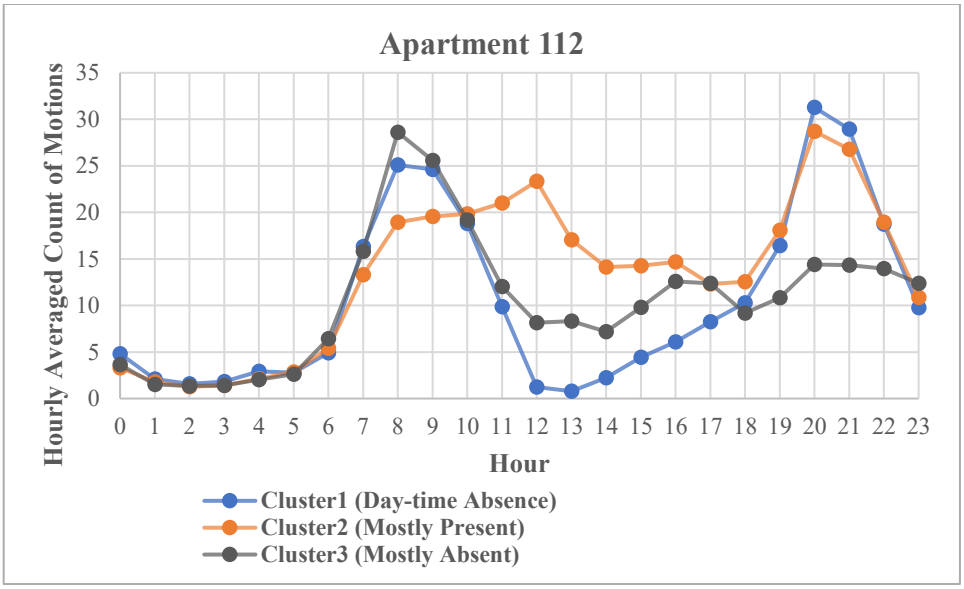


Figure 5-7. Average of hourly count of motions for each occupancy cluster in apartment 112; the hourly averaged values are real values before z-normalization

As mentioned in section 2.1, the limitation of occupancy pattern discovery practices in previous works is associated with short-term occupancy data collection in the residential sector. Since surveys like TUS are mainly collected for a limited number of days, the obtained occupancy models from these surveys cannot properly capture the variation in the presence schedule of individual households. The bar chart in Figure *5-8* (a) demonstrates that apartment 152 is mostly occupied on Mondays and weekends since the "mostly present" occupancy cluster has usually occurred on these days. And it is also obvious that occupancy cluster_1, "day-time absence", usually occurs on all working days except for Mondays. This result contrasts the assumption of prior occupant models, which used TUS to find occupancy patterns. The mentioned studies made an assumption regarding the repetition of a single occupancy pattern for all the working days and another occupancy pattern for working days (Buttitta et al., 2017). Analyzing one-year historical data demonstrates that working days and weekend days can have similar occupancy patterns. However, for the two other apartments, occupancy clusters do not demonstrate transparent relations to specific weekdays like apartment 152. For example, occupancy clusters in apartment 162 are almost equally distributed among all weekdays (see Figure 5-8 (b)). However, the occupancy clusters of apartment 162 show relations to seasons (see Figure 5-9 (b)); cluster_1 in 162 can be observed in spring and summer, while cluster_2 mainly happens during fall and winter

days. The obtained number of clusters for each apartment indicates that presence routines of individual households can be grasped using two to three clusters, and these clusters can be explained using time-variables like weekdays and seasons.

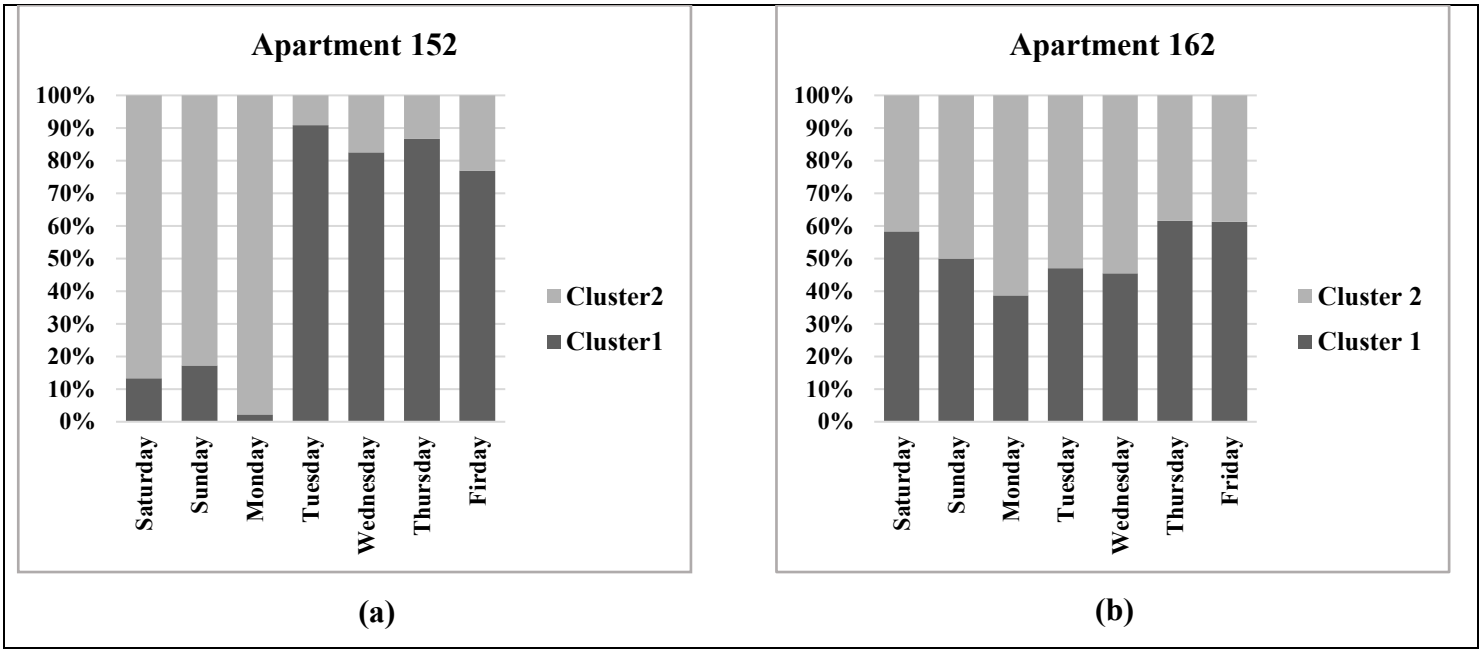


Figure 5-8. Distribution of occupancy patterns among weekdays (a) in apartment 152 and (b) in apartment 162

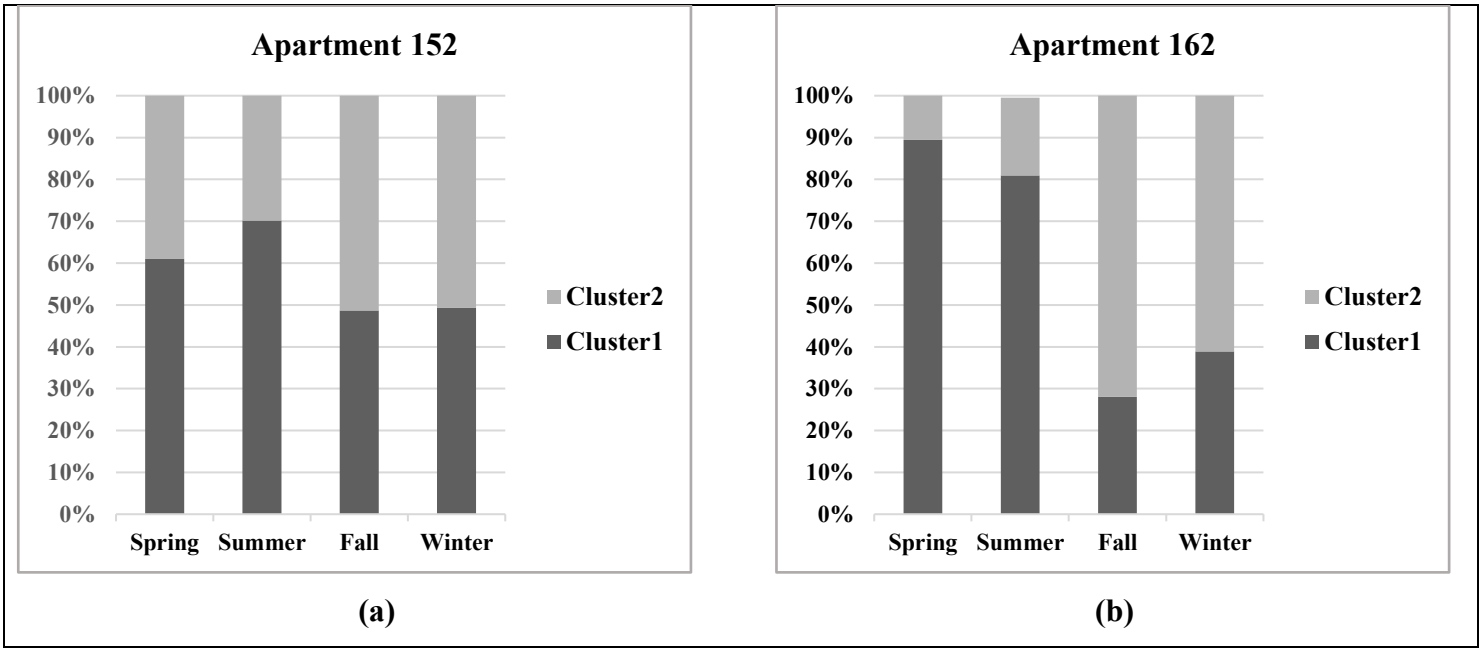


Figure 5-9. Distribution of occupancy patterns among seasons (a) in apartment 152 and (b) in apartment 162

5.3. Relationship between load profiles and occupancy clusters

It has been seen that K-shape clustering can group days with distinctive occupancy patterns in each apartment and capture the flexibility in each household's presence routines. Now, it is important to see the influence of occupancy patterns on the shape of load profiles and realize whether distinctive occupancy clusters are unique in case of load profiles. Using the 25th, 50th, and 75th percentiles of hourly consumptions in each occupancy cluster, the most probable peak periods in a typical day of each occupancy cluster are depicted in Figure 5-11, Figure 5-12, and **Error! Reference source not found.**. Based on the obtained percentile profiles, 7 am, 8 am, 7 pm, and 8 pm are the common peak hours in clusters of all apartments.

Additionally, it can be seen that during days when apartments are mostly occupied (i.e., "mostly present" cluster in apartments 112, 152, and 162), a peak consumption is bound to happen at noon (12 pm) (see Figure 5-11 (b), Figure 5-12 (b), and **Error! Reference source not found.** (b)); this peak does not exist in other occupancy clusters, so the occurrence of noon peak differentiates the consumption profiles of "mostly present" clusters from other occupancy clusters. This noon period cannot be easily identified without considering the occupancy clusters, meaning that when the hourly consumption percentiles of all days (without grouping days by occupancy clusters) are depicted, the noon peak consumption might not even appear at the 25th percentile. To better understand the significance of occupancy clustering in the determination of energy usage during the noon, Pareto figures of hourly energy consumption during 12 pm is depicted to see the distribution of energy usage at 12 pm of days grouped in each occupancy cluster (see Figure 5-13, Figure 5-14, Figure 5-15). The hourly average energy consumption throughout the year is calculated for each apartment to have a basis for defining high energy consumption values at 12 pm; the obtained annual average energy consumption is 201 W.h in apartment 112, 260 W.h in apartment 152, and 223 W.h in apartment 162. If the value of energy usage at 12 pm is higher than the annual average, that value is considered high-energy consumption. The Pareto lines in all three "mostly present" clusters demonstrate that around 60 to 70 percent of the days grouped in "mostly present" clusters experience energy usage values of higher than annual average energy consumption. On the other hand, the percentage of days with energy usage of less than annual average energy consumption at 12 pm in each apartment is as follows:

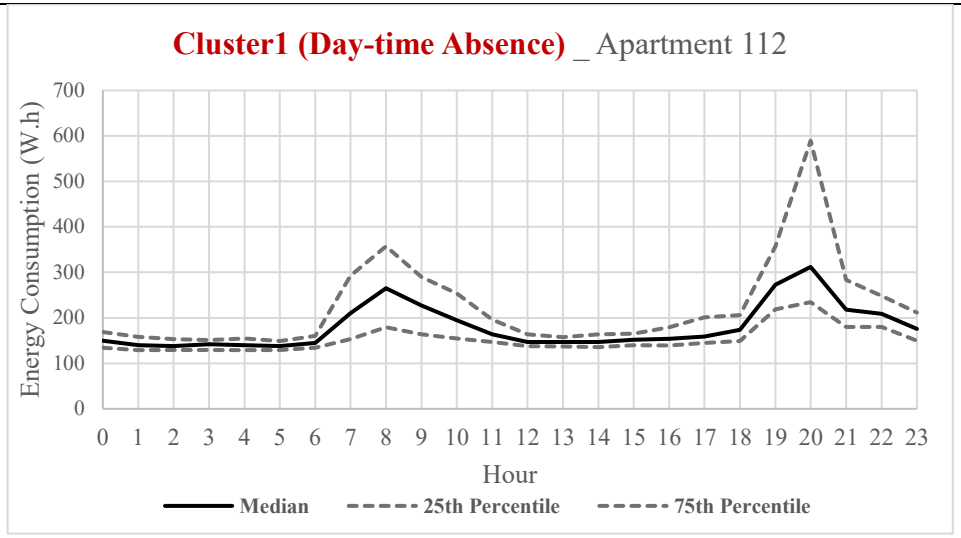
- 90% in "day-time absence" cluster and 80% in "mostly absent" cluster of apartment 112,

- 75% in the “day-time absence” cluster of apartment 152,
- 90% in the “mostly absent” cluster of apartment 162.

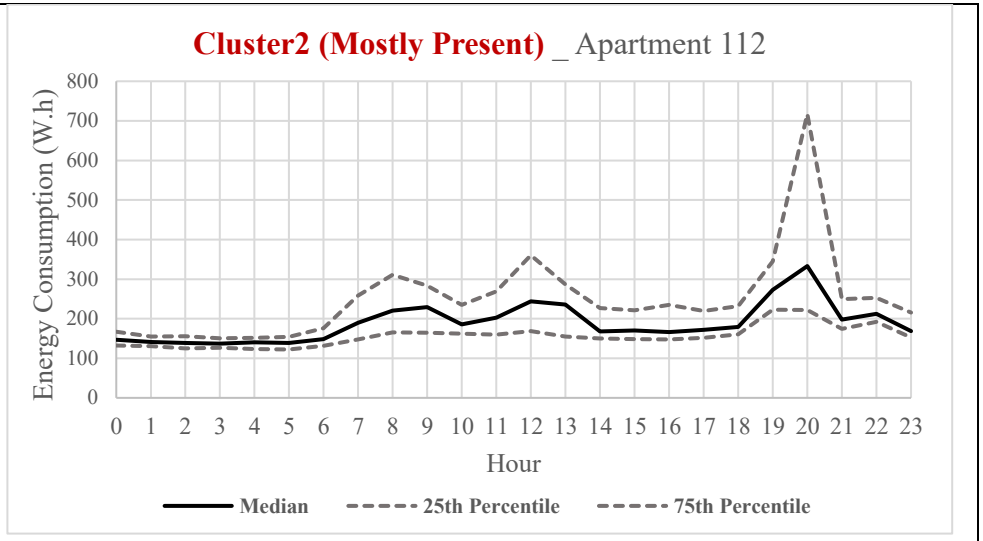
It can be concluded that the energy consumption during the noon (12 pm) is expected to be higher than the annual average usage when the households follow the “mostly present” schedule. In general, the determination of the occupancy profile of a day can give useful information about the timing of high consumption. Knowing the occupancy schedule of a day can contribute to the energy estimations on an hourly basis.

Furthermore, in apartment 112, the percentile profiles of “day-time absence” and “mostly absent” clusters have similar behaviors during the mornings, and the difference appears in the evening period. For days grouped in the “mostly absent” cluster, the 50th and 75th percentile during the evening only reaches a little higher than 200 and 300 W.h, respectively (see Figure 5-11 (c)). While the same percentiles during the evening can be as high as 300 and 600 W.h in the “day-time absence” cluster (Figure 5-11 (a)).

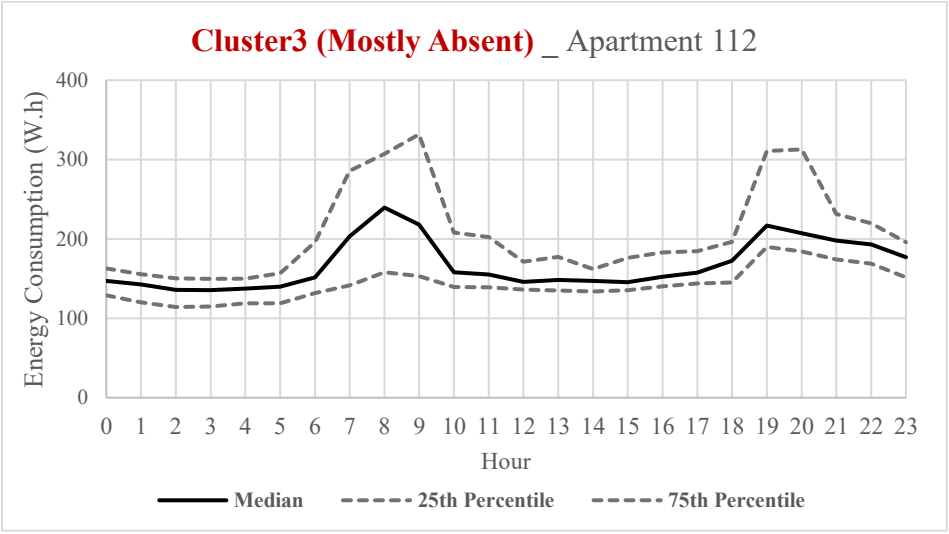
Based on the differences that exist in the load profiles of each occupancy cluster, it can be concluded that occupancy is an essential determinant of the shape of load profiles. In general, it is true to say different occupancy clusters in each apartment are characterized by different consumption profiles, and it is possible to identify the timing of high- and low-consumption periods using the occupancy pattern of a day.



(a)



(b)



(c)

Figure 5-10. Hourly percentiles of electricity consumption in apartment 112 in (a) “day-time absence” and (b) “mostly present” (c) “mostly absent” cluster

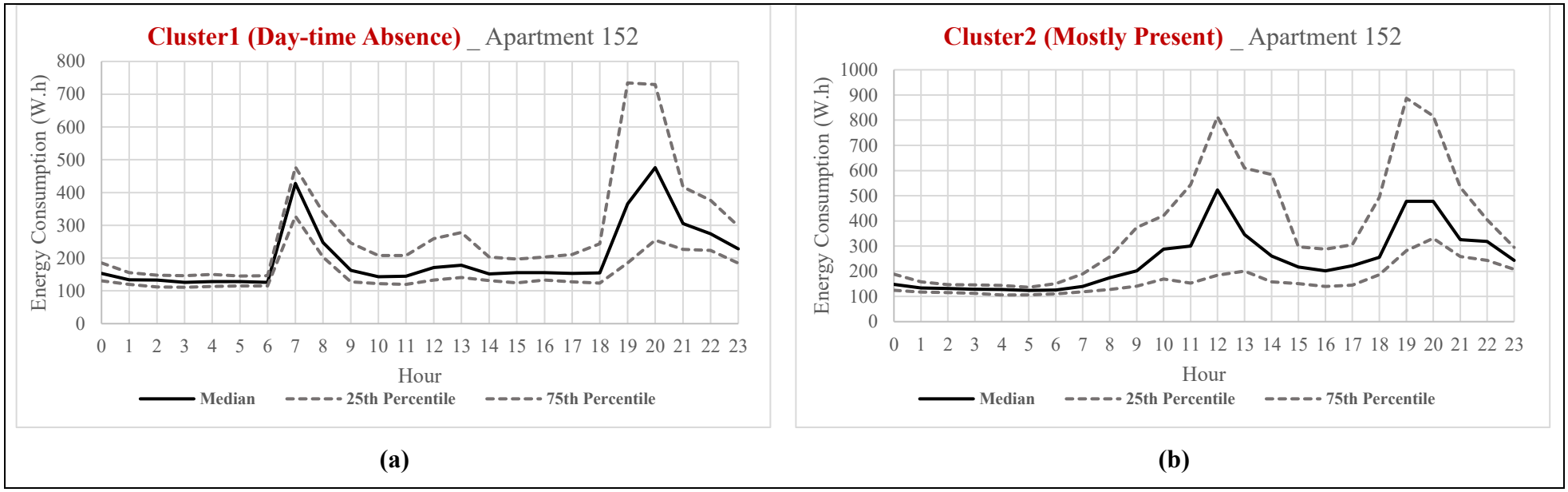


Figure 5-11. Hourly percentiles of electricity consumption in apartment 152 in (a) "day-time absence" and (b) "mostly present" cluster

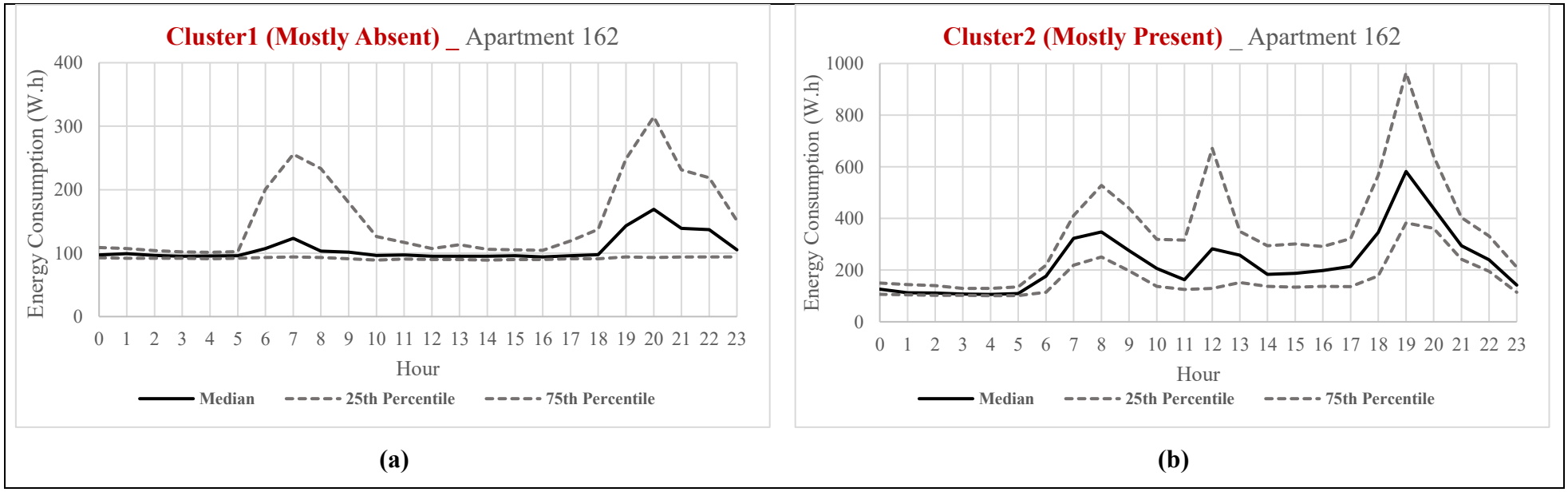


Figure 5-12. Hourly percentiles of electricity consumption in apartment 162 in (a) "mostly absent" and (b) "mostly present" cluster

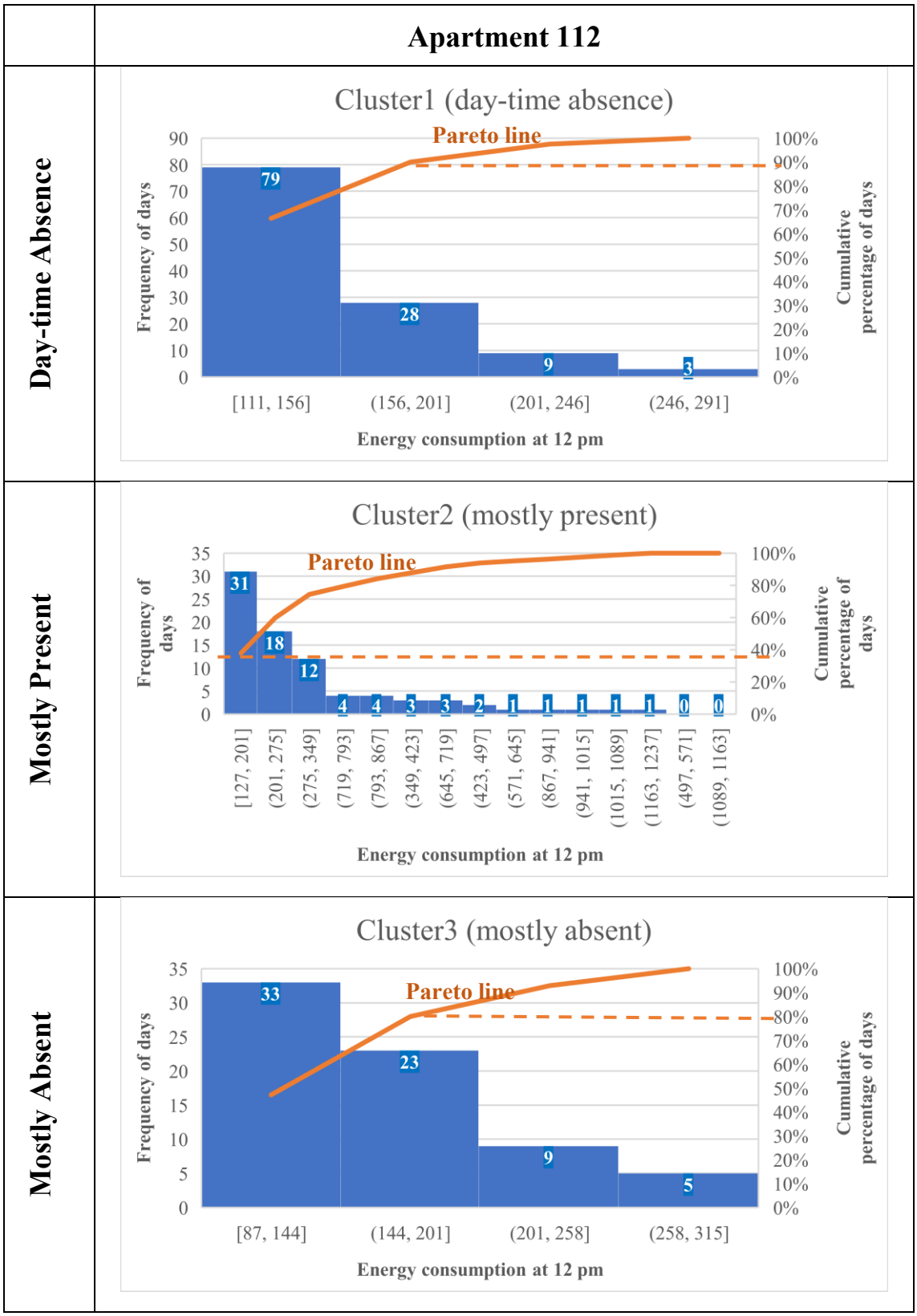


Figure 5-13. Distribution of energy consumption values recorded at 12 pm in days of each occupancy cluster of apartment 112. In cluster2 ("mostly present"), for almost 60% of days, an energy consumption of higher than 201 W.h (the average hourly energy consumption over the year in apartment 112) is recorded at 12 pm.

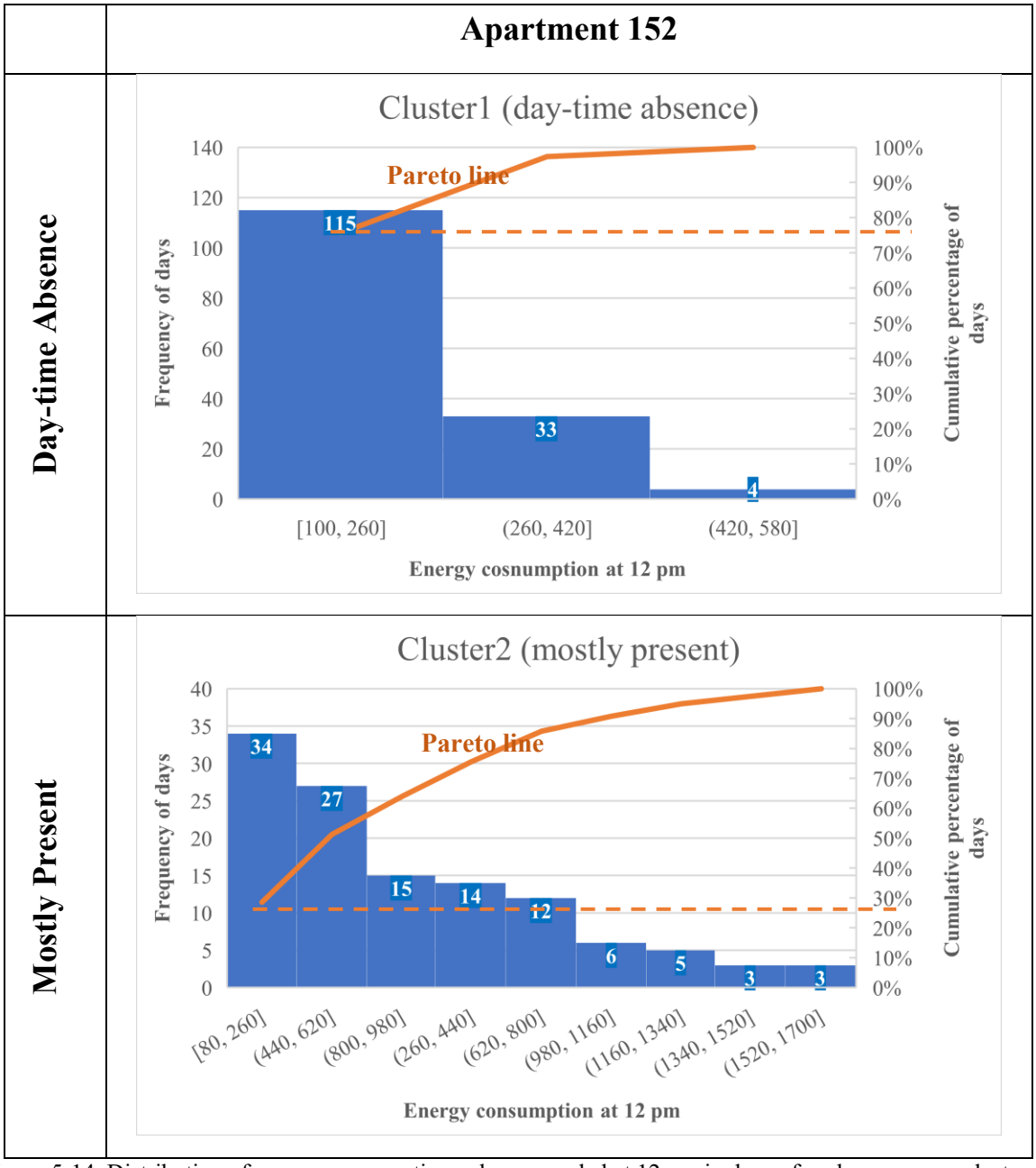


Figure 5-14. Distribution of energy consumption values recorded at 12 pm in days of each occupancy cluster of apartment 152. In cluster2 ("mostly present"), for almost 70% of days, an energy consumption of higher than 260 W.h (the average hourly energy consumption over the year in apartment 152) is recorded at 12 pm.

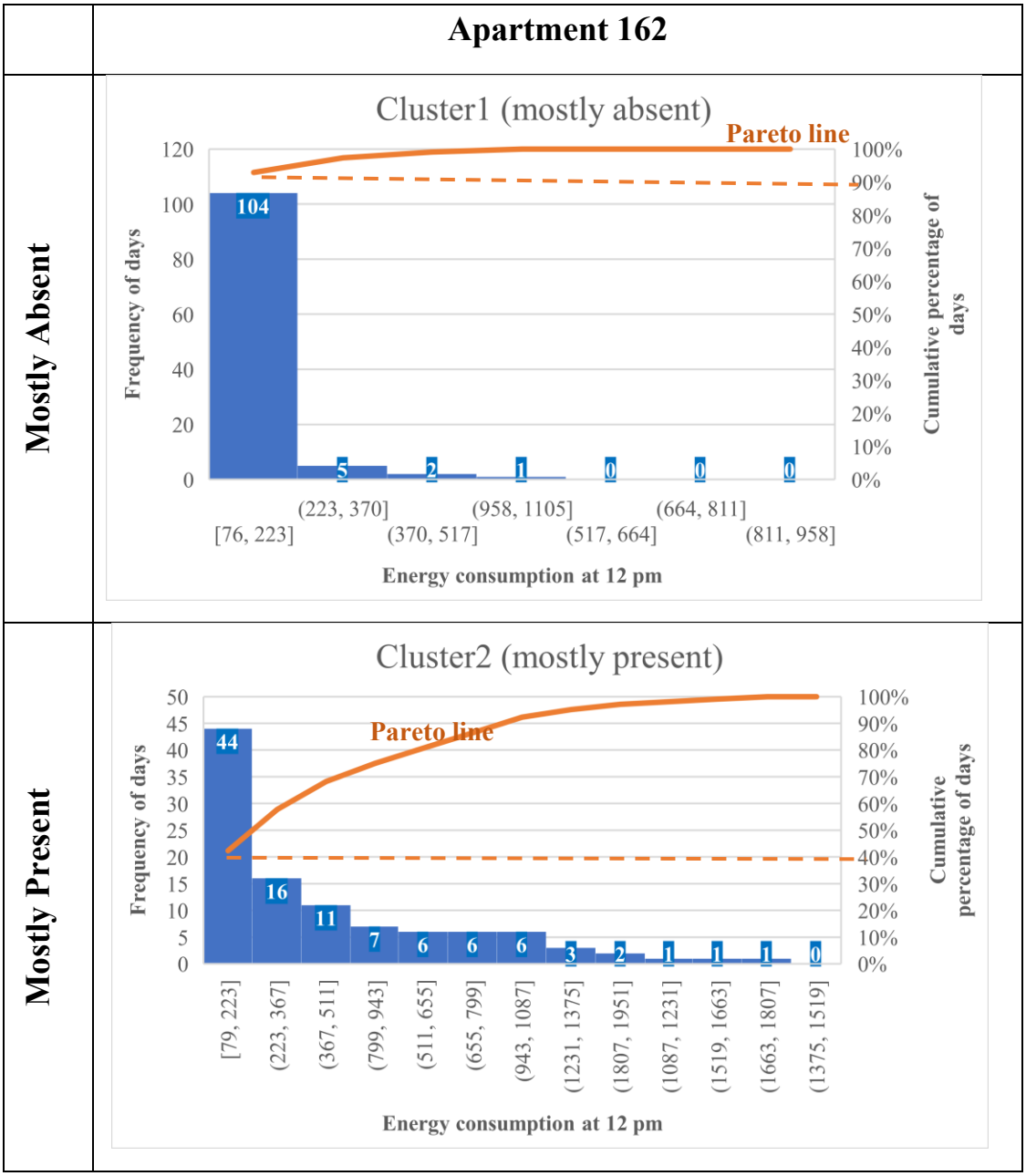


Figure 5-15. Distribution of energy consumption values recorded at 12 pm in days of each occupancy cluster of apartment 162. In cluster2 ("mostly present"), for almost 60% of days, an energy consumption of higher than 223 W.h (the average hourly energy consumption over the year in apartment 162) is recorded at 12 pm.

5.4. Change point detection results

As mentioned earlier, the change point is applied to find the hours when the consumption pattern is bound to experience a significant change (i.e., a rise or a drop). The information provided by the

change point results would give us an insight into each household's energy consumption routine. Since occupancy clusters are characterized by distinctive energy consumption profiles, the change point is applied to the daily load profiles of each occupancy cluster separately. Hence, to determine the most probable change hours (change points) in an occupancy cluster, change point detection (CPD) is applied to daily load profiles of all the days grouped in that cluster separately. The change hours obtained from each daily load profile are recorded (an example of a change point detection result for a single daily load profile is shown in Figure 4-5). When the records of change hours are obtained from all daily load profiles, the number of change occurrences at each hour should be counted in order to determine the most probable change hours in each occupancy cluster. As mentioned earlier several values for the maximum number of change points (Q) is tested, and the most suitable value is selected based on the obtained results. In this study, $Q = 5$ is selected for all occupancy clusters except for "mostly present" clusters; $Q = 6$ is shown to be more appropriate for "mostly present" clusters in all three apartments. Based on the chosen Q, the proper limit to determine the frequent changing hours is 0.4. Hence, the hours detected as change points in at least 40 percent of the days of an occupancy cluster are the most probable changing hours within that cluster. In this section, the results acquired for apartment 152 are only discussed, and the results of other apartments are available in Appendix D. The results show that change point is detected at 6 am, 8 am, 6 pm, and 8 pm in 86, 70, 59, and 57 percent of days in cluster_1 ("day-time absence" cluster) of apartment 152, respectively (see Figure 5-16). The relative frequencies of frequent change hours in the "day-time absence" cluster are higher than the "mostly present" cluster (see Figure 5-16), which indicates that occupants are regularly following a certain energy consumption routine in days of the "day-time absence" cluster. In contrast, in the "mostly present" cluster, discerning the frequent change hours from the rest of the hours is not as clear as they are in the "day-time absence" cluster. Therefore, it is true to say that the variation in energy consumption routines of occupants in the "day-time absence" cluster is much lower than the "mostly present" cluster. Despite the variations in daily load profiles of the "mostly present" cluster, there are hours at which changes in the energy consumption pattern are more probable. 8 am, 11 am, 2 pm, 5 pm, and 8 pm have a relative frequency of higher than 0.4 for change point occurrence (see Figure 5-16), so they are considered highly probable change hours for days grouped in "mostly present" cluster. In other words, during days when occupants of apartment 152 are mostly present throughout the day, it is expected for their energy consumption pattern to change

at the mentioned hours. The frequent change points in each occupancy cluster indicate the hours at which the mean consumption is usually increasing or decreasing. Based on the most frequent change hours, it is possible to separate high- and low-consumption segments (periods). Figure 5-17 shows that the mean consumptions of adjacent periods are well-separated from one another. One of the common high-consumption periods in all clusters of each apartment occurs at 7 and 8 pm; so, it is true to say that this period is not characterized by occupancy clusters. On the other hand, it is revealed that a high-consumption period in the noon is only discovered in "mostly present" clusters, and this high-consumption period does not exist in other occupancy clusters. Therefore, the high or low state of energy consumption during the noon can be determined through the occupancy schedule patterns. This result emphasizes the importance of occupancy schedules on the shape of daily load profiles. Furthermore, it can be seen that no change has been detected at hours 12 am, 10 pm, and 11 pm in any of the occupancy clusters. The reason is rooted in the calculation of maximum likelihood estimates, which cannot be obtained for the 1^{st}, $(n)^{th,}$ and $(n-1)^{th}$ points of a time-series (see section 4.4 4.4.1). Therefore, it is recommended to consider the starting point of the daily time-series with the hour that has the least chance of change occurrence (e.g., 4 am). Accordingly, the same starting hour should be considered for all steps of analysis, including removing the days in the data cleaning step and separating time-series of occupancy for K-shape clustering. For example, if 4 am is considered as the starting point, the occupancy time-series used for time-series clustering must start from 4 am and end at 3:59 am. In this study, 12 am is considered as the time-series' starting point. The results indicate that the obtained periods are well-separated regarding their mean consumption value in all clusters of apartments 112, 152, and 162. Therefore, the results obtained by considering midnight (12 am) as the starting point of the time-series is accepted. The change point results can also be justified using Figure *5-11*, Figure 5-12, and **Error! Reference source not found.**, which show the peak consumption hours of 7 and 8 pm, while a drop is expected from 9 pm in occupancy clusters of all apartments.

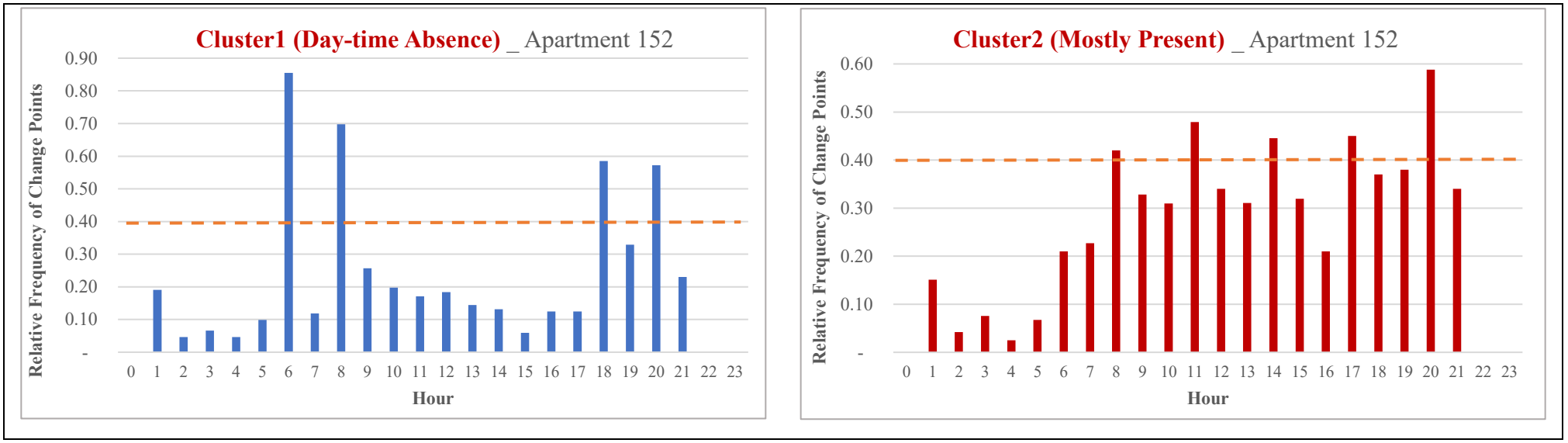


Figure 5-16. relative frequency of change occurrence at each hour in cluster1 (Day-time Absence) and cluster2 (Mostly Present) of apartment 152

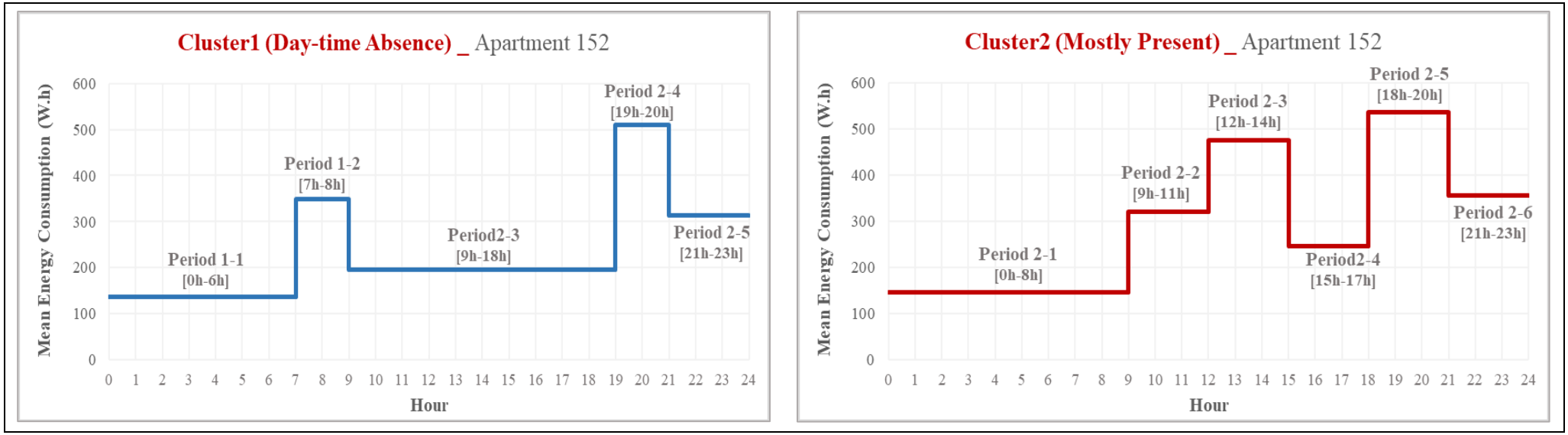


Figure 5-17. Mean electricity consumption within each specified period of cluster1 (Day-time Absence) and cluster2 (Mostly Present) in apartment 152

5.5. LASSO regression results

Up to now, the periods characterized by high- and low-consumption are determined for each occupancy cluster. The purpose of the analysis at this stage is to identify the most influencing activity types within each period. As mentioned earlier, several plug variables are available, and for each plug variable, the location in the apartments is assumed using correlation analysis in section 5.1. In this step, regression analysis is used to realize which of these plug variables are the most influencing ones at each period. Since the plugs are linked to zone-related activities, the regression analysis results can help us understand which types of zonal activities contribute to the variations in the total energy consumption at each time-period. Plus, it is possible to provide useful feedback to occupants about their consumption behaviors throughout the day. Therefore, in this step, a regression model is trained for each time-period, using lighting and plug variables as predictors and total load as the target value. As mentioned in section 5.54.5, regression models should be able to handle sparsity and multicollinearity that might exist among predictor variables. In section 5.1, sparsity could be found among plug variables (Figure 5-2 (c)). In addition to sparsity, multicollinearity can affect the estimated coefficients of regression models, and regularization can handle this issue as well. Variance-inflation factors (VIFs) are utilized to identify multicollinearity in the predictor matrices. The VIFs determine whether it is possible to explain one particular predictor using the linear combination of other predictors. For instance, to obtain the VIF for predictor A, a multiple linear regression model will be trained for this predictor using the rest of the predictors so that predictor A is considered the target variable. The R-squared (R^2) of the trained model shows how well predictor A can be explained linearly by the combination of other predictors. According to the Equation 20, when R^2 gets closer to its maximum value (i.e. $R^2 = 1$), the VIF will approach infinity. Therefore, high VIF values imply strong multicollinearity among predictors, while VIFs closer to 1 rejects the existence of multicollinearity.

$$VIF = \frac{1}{1 - R^2} \qquad (Equation\ 20)$$

- $0 \leq R^2 \leq 1$
- $1 \leq VIF \leq \infty$

In previous works, $VIFs = 3.3$ is considered the threshold based on which the multicollinearity among predictors is either rejected or accepted (Satre-Meloy, 2019) (Roberts & Thatcher, 2009), so VIFs higher than 3.3 indicate multicollinearity among variables. In this study, the same threshold is considered to check the existence of collinearity. In datasets of the case study apartments, multicollinearity exists among plug variables (see Table 8-3 in Appendix E); OTH_P4, OTH_P7, OTH_P8 in apartment 112, OTH_P6, OTH_P9, OTH_P10 in apartment 152, and OTH_P4, OTH_P7, OTH_P8 in apartment 162 have shown strong multicollinearity since their VIFs are extremely high. After checking the energy consumption pattern of these plug variables from Table *5-2*, Table 8-1, and Table 8-2, it is revealed that the consumption pattern of the plugs with high multicollinearity belongs to the first category of plug variables (see Figure 5-2 (a) in chapter 5.1). As discussed earlier, the consumption pattern of these plugs shows that hourly consumption of these plugs only experiences minor fluctuations and the changes in their values of consumption are insignificant. Some examples of appliances with consumption patterns of first category are cordless phone, internet modem, etc., which does not show strong relations to the presence of occupants and their activities. Since, the plug variables in this category are not able to explain the variations in total load, it is expected that the coefficient of these variables to be estimated at zero after implementing the LASSO penalty.

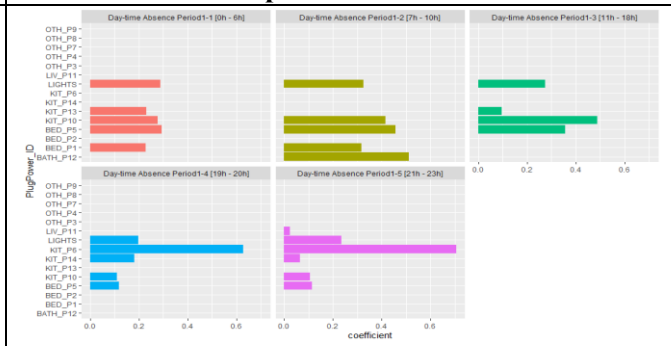
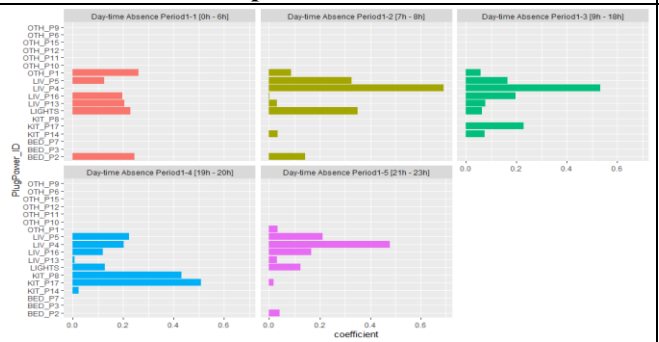
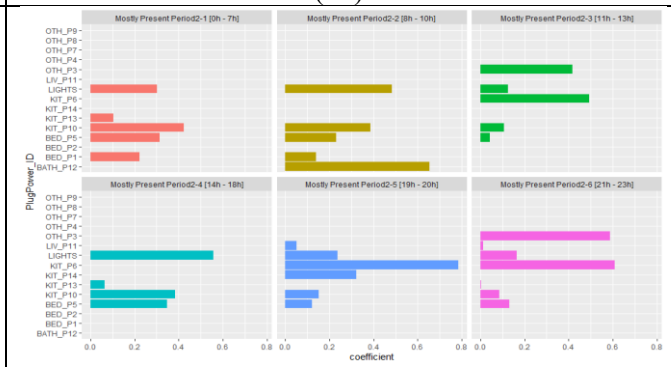
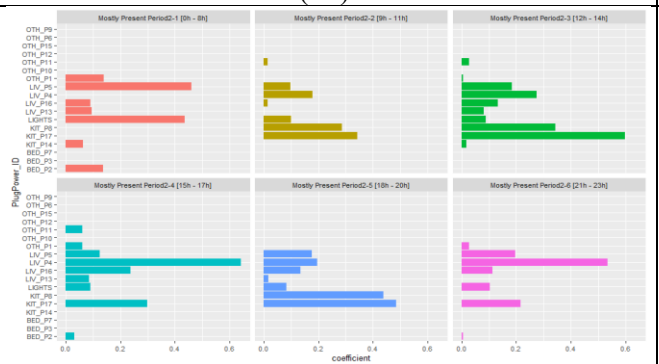
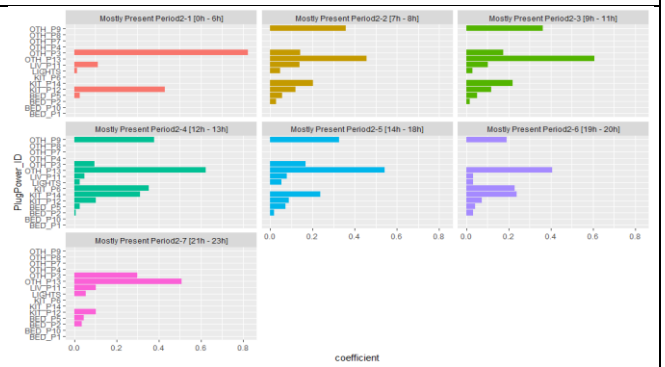
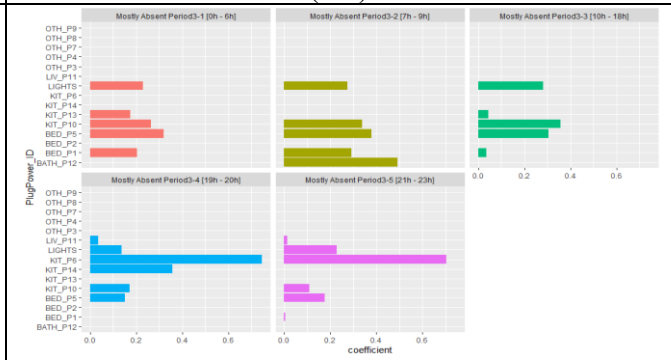
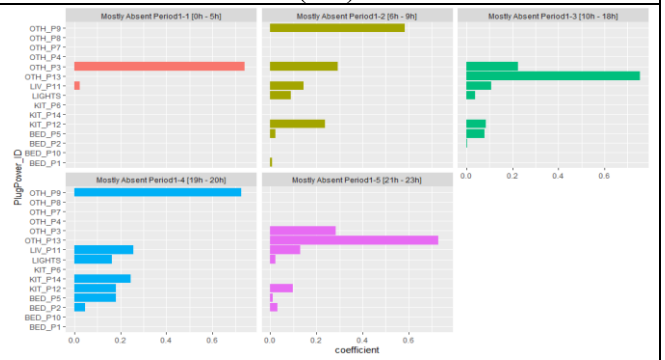
	Apartment 112	Apartment 152	Apartment 162
Day-time Absence cluster	 (a.I)	 (b.I)	
Mostly Present cluster	 (a.II)	 (b.II)	 (c.II)
Mostly Absent Cluster	 (a.III)		 (c.III)

Figure 5-18. 39 regression models obtained for each period within all occupancy clusters in apartment 112, 152, and 162

Figure 5-18 demonstrates the estimated coefficient values of all the LASSO regression models; in total, 39 models are trained for the three apartments. Out of these 39 models, 16, 11, and 12 belong to apartments 112, 152, and 162, respectively. Since the models are trained separately for each period, comparing the variables' estimated coefficients of different periods cannot be considered a valid comparison. To properly analyze the coefficients obtained for each period and understand the variable importance in different periods, the relative ratios of coefficients in each time-period should be compared to the ratios in another period. Within each period, the variables are ranked based on their magnitude of consumption and their correlation with total energy usage during that period, so plugs with higher coefficients can better explain the variations in total energy consumption. Therefore, if a plug ("Plug n") has a high rank within a certain period ("Period I") while its rank drops within another period ("Period II"), it does not necessarily show that the consumption of "Plug n" in "Period II" is lower than "Period I". This drop only implies that the consumption value of "Plug n" compared to other plugs is lower in "Period I" than it is in "Period II". The R^2 of the models range from 0.85 to 0.97. This shows that even after eradicating and shrinking the coefficients of some plug variables, the variability in total consumption can be well-explained using the remaining plug variables.

Based on the obtained models shown in Figure 5-18, some of the important outcomes are as follows:

The common feature in all of the 39 models is that the estimated coefficient of plug variables with multicollinearity is zero or insignificant (see coefficients of OTH_P4, OTH_P7, OTH_P8 in apartment 112, OTH_P6, OTH_P9, OTH_P10 in apartment 152, and OTH_P4, OTH_P7, OTH_P8 in apartment 162). This proves that the plug variables belonging to the first category of energy consumption pattern can be handled by LASSO regression, and their coefficients will be set zero since these variables do not lead to variations of the total energy consumption. Other plugs can be found with all-time zero coefficients such as: BED_P2 and OTH_P9 in apartment 112, BED_P3, BED_P7, OTH_P12, and OTH_P15 in apartment 152, and BED_P1 and BED_P10 in apartment 162; these plugs are associated with insignificant energy consumption values and are very sparse (see Table 5-2, Table 8-1, and Table 8-2).

Furthermore, in the “mostly present” cluster of apartment 152 (b.II), the rank of KIT_P17, which is related to kitchen-related activities, has surpassed other plugs during period 2-3 [12 pm - 2 pm]. Additionally, since KIT_P8 has the 2nd rank during this period, it can be interpreted that occupants tend to use kitchen appliances and attend to kitchen-related activities during this period. In other apartments, the equivalent noon period in “mostly present” clusters (a.II, c.II) (i.e., period 2-3 [11 am - 1 pm] in apartment 112 and period 2-4 [12 pm - 1 pm] in apartment 162) is associated with kitchen-related activities as well. For instance, in apartment 112, KIT_P6 is insignificant in most of the time-periods, but it appears in the 1st rank during period 2-3 in the “mostly present” cluster (a.II). The same inference is true for KIT_P6 in period 2-4 of the “mostly present” cluster in apartment 162 (c.II). This plug has become 3rd during period 2-4 [12 pm - 1 pm] in the “mostly present” cluster, while its coefficient has rarely appeared in other periods. In the same period, KIT_P14 has the 4th rank, and its coefficient has a relatively high magnitude among other plugs, which further proves the existence of kitchen-related activities during period 2-4 in the “mostly present” cluster of apartment 162. Based on the approximate similarity in the timing of period 2-3 [11 am – 1 pm] in Figure 5-18 (a.II), period 2-3 [12 pm – 2 pm] in Figure 5-18 (b.II), and period 2-4 [12 pm – 1 pm] in Figure 5-18 (c.II), these periods are called as the “noon period”. Since the “noon period” in “mostly present” clusters of all three apartments is associated with kitchen-related activities, and the mentioned periods have high energy consumption in all three apartments (see Figure 5-17 (cluster2, period 2-3), Figure 8-4 (cluster2, period 2-3), and Figure 8-6 (cluster2, period 2-4)), the households should be noticed about kitchen-related appliances during noon. While the period after the mentioned “noon period”s in all three apartments have a lower consumption (see Figure 5-17 (cluster2, period 2-4), Figure 8-4 (cluster2, period 2-4), and Figure 8-6 (cluster2, period 2-5)), and it can be seen from section 5.2.2 that during these periods occupants are usually present at home (see Figure 5-4 (b), Figure 5-5 (b), and Figure 5-6 (b)). Therefore, occupants can shift some of the kitchen-related consumptions from the “noon period”s to the next low-consumption periods. In general, with a good knowledge about occupants' schedule and the type of activities taking place during each period, it is possible to check the feasibility of appliance scheduling interventions. Additionally, kitchen plugs have high ranks within the time-periods from 7 pm to 8 pm in every cluster of all apartments (See the

following variables' coefficients during the periods from 7 pm to 8 pm in all 39 models depicted in Figure 5-18: KIT_P6, KIT_P14 in apartment 112, KIT_P8 and KIT_P17 in apartment 152, and KIT_P6 and KIT_P14 in apartment 162).

Earlier, it is found that the periods from 7 pm to 8 pm have high consumptions in all apartments (see Figure 5-17, Figure 8-4, and Figure 8-6). For load shifting, suggestions regarding shifting kitchen-related activities can be given to the households during the noon and evening high-consumption periods. To see the impact of kitchen-related activities on the mean energy usage within these high-consumption periods, the mean usage of noon and evening periods (identified as high-consumption periods in CPD analysis) are recalculated with the exclusion of the top-ranking kitchen-related plugs. The obtained results show that:

- Removing KIT_P6 from period 1-4 [7 pm – 8 pm], period 2-3 [11 am – 1 pm], period 2-5 [7 pm – 8 pm], and period 3-4 [7 pm – 8 pm] in apartment 112 (Figure 8-4 in Appendix D) can reduce the energy usage within the mentioned periods by 19% on average (the reduction in period 1-4, period 2-3, period 2-5, and period 3-4 is 26%, 10%, 27%, and 13%, respectively.).
- Consider the noon and evening high-consumption periods in apartment 152 (period 1-4 [7 pm – 8 pm], period 2-3 [12 pm – 2 pm], and period 2-5 [6 pm – 8 pm] in Figure 5-17). It is revealed that excluding KIT_P8 can decrease the mean energy usage of the mentioned periods by 9% on average, while eradicating KIT_P17 can lead to an average reduction of 20% within the mentioned periods.
- Shifting KIT_P14 from period 1-4 [7 pm – 8 pm], period 2-4 [12 pm to 1 pm], and period 2-6 [7 pm – 8 pm] in apartment 162 (Figure 8-6 in Appendix D) can result in an average reduction of 5%, 10%, and 17% during the mentioned periods, respectively.

The mentioned reduction percentages can be presented to the occupants, and based on the type of appliances associated with the mentioned plugs, households can consider moving the operation of the respective plugs to another period that has a lower energy consumption. The scheduling is dependent on the type of appliances. For example, if the energy consumption of the mentioned kitchen-related plugs is associated with appliances like

dishwasher, rice cooker, etc., shifting their operation is more viable compared to appliances like microwave oven whose operation time has a great impact on occupant's comfort. According to (Zhou et al., 2016) (Zhao et al., 2013), appliances can be categorized into two groups naming, non-schedulable[4] and schedulable[5], based on their dependency on humans' manual control. Since users' comfort is highly dependent on the operation time of non-schedulable appliances (e.g., hairdryer, microwave oven, etc.) (Zhou et al., 2016), changing the timing of activities related to non-schedulable appliances might be difficult.

Another interesting result of regression models is related to bathroom activities during the mornings in all clusters of apartment 112 (in regression models of apartment 112 in Figure 5-18, see coefficients of BATH_P12 in period 1-2 [7 am - 10 am], period 2-2 [8 am - 10 am], and period 3-2 [7 am - 9 am] in "day-time absence", "mostly present", and "mostly absent" clusters, respectively.). The results indicate that the estimated coefficient of BATH_P12 is higher than other plugs during the mentioned periods, while this plug has not appeared in the rest of the periods in apartment 112. It can be concluded that the occupant(s) living in apartment 112 consume a considerable amount of electricity in the bathroom in the mornings.

Furthermore, it can be seen that in apartment 112, LIGHTS coefficients is usually noticeable, especially during the sleeping periods (i.e., period 1-1 [12 am - 6 am] in Figure 5-18 (a.I), period 2-1 [12 am - 6 am] in Figure 5-18 (a.III), and period 3-1 [12 am - 7 am] in Figure 5-18 (a.II)). Similarly, in apartment 152, LIGHTS is one of the most important determinants of total electricity load during the sleeping periods (i.e., period 1-1 [12 - 6 am] and period 2-1 [12 - 8 am] in "day-time absence" and "mostly present" clusters, respectively). Therefore, occupants of apartment 112 and 152 can save energy by paying attention to lighting consumption during the sleeping periods. On the contrary, LIGHTS' coefficient is negligible within the sleeping periods in apartment 162 (i.e. period 2-1 [12 am - 6 am] in Figure 5-18 (c.II) and period 1-1 [12 am - 5 am] in Figure 5-18 (c.III)). In this case, households of apartments 112 and 152 need to make sure the lights are off throughout the nights. LIGHTS are also important in the morning high energy consumption

[4] Some examples of non-schedulable appliances are computer, printer, microwave oven, television, hairdryer, etc.
[5] Some examples of schedulable appliances are washing machine, tumble dryer, rice cooker, air conditioner, dishwasher, water heater, etc.

period in the “day-time absence” cluster of apartment 152 (Figure 5-18 b.I, period 1-2 [7 am – 8 am]). It can be seen that LIGHTS are assigned to a high coefficient and ranked third after two living room-related appliances (LIV_P4 and LIV_P5). Household of apartment 152 needs to take advantage of the daylight throughout this period to save energy consumed for lights.

The obtained regression models can have several benefits:

1. Regarding energy reduction, the regression models show which types of activities have the highest priority for energy reduction at each time of day. In this way, occupants can be more cautious about the type of activities that consume a considerable amount of electricity at each time of day
2. Regarding load shifting purposes, it is possible to see which types of activities have the most contribution during high consumption periods and check the possibility of shifting those activities to a low consumption period (identified by change point method).

6. Conclusions, limitations, and future work

6.1. Conclusions

This study deals with the application of metering infrastructures on household-level energy analyses. It can be recognized from the literature that although the drivers of end-use energy consumption have been broadly investigated in the previous works, the factors of in-day energy consumption, which contributes to the determination of the shape of daily load profiles, have not gained enough attention. These temporal drivers are the determinants of energy consumption at each time of day, and recognizing them leads to an improved energy management in the residential sector. The increasing availability of high-resolution data of residential households, which is collected over a long period, creates the opportunity to execute analyses to discover individual households' diverse routines and identify the impact of temporal factors of the shape of daily load profiles. In this study, a systematic methodology framework is developed to investigate the temporal and contextual factors of households' energy consumption, such as occupants' activities and occupancy. Firstly, a time-series clustering method, called K-shape, is implemented on the daily time-series of occupancy, which enables the discovery of distinctive occupancy schedule patterns of a household. Secondly, for days grouped within the same occupancy cluster, the change point detection (CPD) method is applied on daily time-series of energy consumption to determine the low- and high-energy consumption periods in each occupancy cluster. Lastly, the LASSO regression method is utilized to discover the most influencing activities on energy consumption throughout the periods obtained from CPD analysis. A detailed look into the results indicates that:

- Single household's presence routines can be captured using two or three clusters, and K-shape clustering can extract all of these diverse presence routines for each household. The obtained occupancy clusters can be explained using time-variables such as season or weekdays.
- It is shown that distinctive occupancy clusters are characterized by unique load profiles. This uniqueness mainly appears in the earlier hours than in the evenings. Based on the obtained results, the evening peak timing is similar between occupancy clusters of different households. However, the difference between the load profiles of distinctive occupancy clusters mainly manifests in earlier hours, especially during the noon. Furthermore, change point detection found the hours denoting a change in the shape of daily load profiles, and

it has been shown that these hours are specific to each occupancy cluster. So, occupancy patterns can be recognized as one of the important temporal factors of the load profiles.

- Additionally, analyzing one-year data of occupancy and energy consumption in three residential apartments indicated that during days when apartments are mostly occupied (i.e., "mostly present" occupancy clusters), the energy consumption during the noon is expected to be higher than the annual average consumption. The results indicate that for almost 60 to 70% of the days grouped in "mostly present" clusters, the energy consumption of higher than annual average usage is recorded at 12 pm. On the other hand, the probability of high-energy consumption in the noon is less than 25% in other occupancy clusters. This result further proves that the occupancy schedule of a day is an important temporal factor of the load profile of that day. Based on the results obtained from the three apartments, identifying the occupancy pattern of a day as the "mostly present" can increase the chance of witnessing high-energy consumption during the noon by more than 60%.
- Furthermore, LASSO regression results of all the apartments reveal the importance of kitchen-related activities during the noon peak period in the "mostly present" clusters. It is shown that kitchen-related activities are influencing factors of energy usage during the evening peak hours (7 to 8 pm) as well.
- The average energy usage during the high-consuming noon and evening periods can be reduced by 5 to 27 % (based on the results of all three apartments) if the top kitchen-related activities are shifted to another time. The obtained knowledge can be presented to the households as suggestions for peak energy reduction, and households can manage to shift high-energy consuming activities based on the flexibility of their requirements and the possibility of scheduling the appliances.

Some of the contribution of this study are as follows:

- The impact of peak demand of several households might be mitigated during some hours of the day since the peak demand hours can be different among households. However, this peak demand can also be intensified during the common peak hours, and the application of this work is of greater importance during these common peak demand hours. The findings of this study can help households to consume energy in a more cost-efficient manner. If the common peak demand periods are revealed to the households, they can manage their

energy bills by reducing their consumption during the peak demand hours. This reduction can be performed by shifting the operation of some appliances from high-demand periods to off-peak periods. Knowing the occupancy schedule of a household and the key activity drivers of energy usage during each energy consumption period allows for personalized energy consumption reduction interventions and improved comprehension of residential consumers' flexibility regarding load shifting and appliance scheduling programs. Discovering the comparative contribution of activity factors to the total energy consumption can also raise occupant's awareness about their routine activities that cause great variations to the energy consumption at each time of day. So, occupants can focus on the high-consuming activities and attempt to reduce the energy consumption of those activities, especially during peak demand hours. Furthermore, the obtained temporal factors (e.g., occupancy clusters and the frequent change points when the energy consumption usually increases or decreases) can be used as inputs of energy prediction models of a household and improve the accuracy of estimations on an hourly basis. It has been seen that knowing the occupancy cluster of a day can contribute to the determination of the most probable peak hours during that day. (Singh et al., 2012) identified occupancy and historical peak load of households as significant determinants of energy predictions on an hourly basis. They further stipulated that although the physical features like temperature are helpful for weekly or daily load predictions, these features are not as effective predictors of hourly load predictions as historical-based data of occupancy and historical peak demand. Plus, it is shown that the occupancy clusters can be explained using time-variables such as seasons and weekdays. Therefore, using occupancy schedule clusters as input of energy prediction models can enhance the accuracy of energy estimations in residential apartments on an hourly, daily, or even monthly basis. Moreover, the methodology framework of this study has been implemented on data sets of three households, and it is shown that the framework can be generalized to different households with different presence schedules and different energy consumption routines. Customized load shifting programs and controlling strategies can be suggested to each household using the proposed data-driven framework. The methodology framework can be utilized for various climate conditions and building characteristics since occupancy and load profiles exist for all households. It is also shown that LASSO regression can be implemented for

plug variables having diverse energy consumption patterns. So, the framework can be generalized to households having a variety of appliances with different energy consumption patterns.

6.2. Limitations of the current study

One of the limitations faced in this study deals with the characteristics of K-shape clustering. As mentioned in section **Error! Reference source not found.**, K-shape is shift-invariant (see Figure 4-2 (c)). Therefore, as long as two time-series have similar shapes, they will be clustered together, even if their phases are different. This quality of K-shape contributed to the ignorance of noisy values and small shifts in the phases of time-series, which makes the method more robust and accurate. However, the significant shifts can also be ignored, as it is shown in Figure 5-6 (a); this can pose challenges to the identification of the usual unoccupied periods from the occupied periods. However, with the change point detection method applied to daily load profiles after occupancy clustering, the usual high-energy consumption periods can still be identified and differentiated from regular off-peak hours. Another limitation of the current work is that each plug variable cannot be associated to a specific appliance. Each plug variable represents the energy usage from an outlet in the apartment, and it is evident that more than one device can be plugged into an outlet. Therefore, guessing the appliances based on the consumption pattern of plug variables is challenging. This limitation might reduce the feasibility of feedback aimed to request occupants to shift some appliances' operations from one period to another. As appliances can be categorized as schedulable and non-schedulable based on their dependency on manual control and occupant's comfort, knowing the type of appliances is necessary for practical appliance scheduling and load shifting interventions. Ignoring appliance types can reduce the practicality and feasibility of energy interventions. Providing practical feedback to occupants necessitates analyzing the operation time of each specific appliance individually in order to give accurate recommendations to shift the schedulable appliances' operation to another time. The presented energy feedback in this study is more useful in notifying occupants about the energy usage of the different appliances within a specific room during high-consumption periods. The feedback and suggestions of this work cannot specifically point to certain appliances. Another issue can be pointed to the privacy concerns of the households being monitored by motion detection sensors. This issue can be resolved using other variables such as indoor environmental factors and occupant-related variables

showing interactions with lights and windows to generate occupancy profiles. (Panchabikesan et al., 2021) demonstrated that occupancy level can be predicted using indoor CO2 and relative humidity, along with energy consumption data.

6.3. Future work

This study emphasizes the role of long-term sensor-collected data and data analysis methods in discovering the temporal determinants of load profiles in residential apartments. However, the vast opportunity provided by data analysis is capable of achieving much more in energy and occupant behavior studies. In this section, some of the potential research opportunities for future works are suggested:

- Based on the results of this study, occupancy schedule patterns are important determinants of the shape of load profiles. Therefore, the prediction of occupancy patterns can lead to a more accurate estimation of energy usage at different times of day. It has also been shown that occupancy patterns can be explained using time-variables such as season and weekdays. Finding the occupancy patterns can hint at the hours at which energy usage usually increases or decreases. Therefore, discovering the drivers of occupancy patterns and predicting the occupancy patterns using the recognized drivers such as time-variables (i.e., season, weekday, etc.) can be useful for building automation purposes.
- The diverse occupancy patterns for each household are discovered using one-year occupancy data. It is also important to further investigate the sufficient length of the data collection period to capture these diversities in the presence routines of each household. As mentioned in section 5.2.2, each household's data collection period may differ based on the relationships between occupancy patterns and time-variables such as season and weekdays. For example, for the household living in apartment 152, occupancy patterns demonstrate strong relations to the weekdays (see Figure 5-8 (a)). On the other hand, for the household living in apartment 162, the diverse occupancy schedules can be explained by the change of seasons Figure 5-9 (b). It can be concluded that a data-collection period of less than one month is capable of capturing the diversity in presence routines of household 152, while this period should cover different seasons (probably a year) for household 162, since occupancy patterns in apartment 162 change seasonally.

- The manual test and trial process to find the optimum number of change points (Q) and the threshold to define frequent change points in the CPD step (which is practiced in (Li, Panchabikesan, et al., 2019)) is not in line with the automation purposes of the HEMSs. One possible solution is sensitivity analysis. Sensitivity analysis is recommended to find out how sensitive are the final results (relative frequencies of hours and determination of the frequent change points where energy consumption change significantly depicted in Figure 5-16, Figure 8-3, and Figure 8-5) to the selection of Q. The lower the sensitivity, the higher the CPD method's robustness in case of determination of the regular peak hours.
- The identified Occupancy profiles, frequent change hours, and the knowledge about the influencing activities can be incorporated into the control systems. Investigating the effect of the mentioned factors on the control strategies should be addressed in future works. For instance, the feasibility of appliance scheduling and its actual impact on peak and electricity cost reduction needs to be measured when the scheduling is put in action.

7. References

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering - A decade review. *Information Systems*, *53*, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Ashouri, M., Haghighat, F., Fung, B. C. M., & Yoshino, H. (2019). Development of a ranking procedure for energy performance evaluation of buildings based on occupant behavior. *Energy and Buildings*, *183*, 659–671. https://doi.org/10.1016/j.enbuild.2018.11.050

Beaudin, M., & Zareipour, H. (2015). Home energy management systems: A review of modelling and complexity. *Renewable and Sustainable Energy Reviews*, *45*, 318–335. https://doi.org/10.1016/j.rser.2015.01.046

Buttitta, G., Neu, O., Turner, W. J. N., & Finn, D. (2017). *Modelling Household Occupancy Profiles using Data Mining Clustering Techniques on Time Use Data*. http://www.buildingsimulation2017.org/; http://www.ibpsa.org/?page_id=962

Buttitta, G., Turner, W. J. N., Neu, O., & Finn, D. P. (2019). Development of occupancy-integrated archetypes: Use of data mining clustering techniques to embed occupant behaviour profiles in archetypes. *Energy and Buildings*, *198*, 84–99. https://doi.org/10.1016/j.enbuild.2019.05.056

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Chen, J., & Gupta, A. K. (2012). *Parametric statistical change point analysis : with applications to genetics, medicine, and finance*. *53*(9), 1689–1699. https://doi.org/https://doi-org.lib-ezproxy.concordia.ca/10.1007/978-0-8176-4801-5

Chicco, G., Napoli, R., & Piglione, F. (2006). *Comparisons Among Clustering Techniques for Electricity Customer Classification*. *21*(2), 933–940.

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

D'Oca, S., & Hong, T. (2015). Occupancy schedules learning process through a data mining

framework. *Energy and Buildings*, *88*, 395–408. https://doi.org/10.1016/j.enbuild.2014.11.065

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Diao, L., Sun, Y., Chen, Z., & Chen, J. (2017). Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. *Energy and Buildings*, *147*, 47–66. https://doi.org/10.1016/j.enbuild.2017.04.072

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*(3), 32–57. https://doi.org/10.1080/01969727308546046

Fan, C., Xiao, F., Li, Z., & Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy and Buildings*, *159*, 296–308. https://doi.org/10.1016/j.enbuild.2017.11.008

Gajowniczek, K., & Zabkowski, T. (2017). Electricity forecasting on the individual household level enhanced based on activity patterns. *PLoS ONE*, *12*(4), 1–27. https://doi.org/10.1371/journal.pone.0174098

Gilani, S., & O'Brien, W. (2017). Review of current methods, opportunities, and challenges for in-situ monitoring to support occupant modelling in office spaces. *Journal of Building Performance Simulation*, *10*(5–6), 444–470. https://doi.org/10.1080/19401493.2016.1255258

Grunewald, P., & Diakonova, M. (2018). Flexibility, dynamism and diversity in energy supply and demand: A critical review. *Energy Research and Social Science*, *38*(September 2017), 58–66. https://doi.org/10.1016/j.erss.2018.01.014

Hoerl, A. E., & Kennard, R. W. (1970). *Ridge Regression: Biased Estimation for Nonorthogonal Problems*. *12*(1), 55–67. http://www.jstor.com/stable/1267351

Huebner, G., Shipworth, D., Hamilton, I., Chalabi, Z., & Oreszczyn, T. (2016). Understanding

electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes. *Applied Energy*, *177*, 692–702. https://doi.org/10.1016/j.apenergy.2016.04.075

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. In *Springer Texts in Statistics*. http://books.google.com/books?id=9tv0taI8l6YC

Jones, R. V., Fuertes, A., & Lomas, K. J. (2015). The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings. *Renewable and Sustainable Energy Reviews*, *43*, 901–917. https://doi.org/10.1016/j.rser.2014.11.084

Kirschen, D. S. (2003). Demand-side view of electricity markets. *IEEE Transactions on Power Systems*, *18*(2), 520–527. https://doi.org/10.1109/TPWRS.2003.810692

Kwac, J., Flora, J., & Rajagopal, R. (2014). Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid*, *5*(1), 420–430. https://doi.org/10.1109/TSG.2013.2278477

Kwac, J., Flora, J., & Rajagopal, R. (2018). Lifestyle Segmentation Based on Energy Consumption Data. *IEEE Transactions on Smart Grid*, *9*(4), 2409–2418. https://doi.org/10.1109/TSG.2016.2611600

Lavin, A., & Klabjan, D. (2015). Clustering time-series energy data from smart meters. *Energy Efficiency*, *8*(4), 681–689. https://doi.org/10.1007/s12053-014-9316-0

Li, J., Panchabikesan, K., Yu, Z., Haghighat, F., Mankibi, M. El, & Corgier, D. (2019). Systematic data mining-based framework to discover potential energy waste patterns in residential buildings. *Energy and Buildings*, *199*, 562–578. https://doi.org/10.1016/j.enbuild.2019.07.032

Li, J., Yu, Z., Haghighat, F., & Zhang, G. (2019). Development and improvement of occupant behavior models towards realistic building performance simulation: A review. *Sustainable Cities and Society*, *50*, 101685. https://doi.org/10.1016/j.scs.2019.101685

Liang, X., Hong, T., & Shen, G. Q. (2016). Occupancy data analytics and prediction: A case study. *Building and Environment*, *102*, 179–192. https://doi.org/10.1016/j.buildenv.2016.03.027

Liu, H., Yao, Z., Eklund, T., & Back, B. (2012). *Electricity Consumption Time Series Profiling: A Data Mining Application in Energy Industry*. 52–66.

McKenna, E., Higginson, S., Grunewald, P., & Darby, S. J. (2018). Simulating residential demand response: Improving socio-technical assumptions in activity-based models of energy demand. *Energy Efficiency*, *11*(7), 1583–1597. https://doi.org/10.1007/s12053-017-9525-4

McLoughlin, F., Duffy, A., & Conlon, M. (2015). A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy*, *141*, 190–199. https://doi.org/10.1016/j.apenergy.2014.12.039

Nilsson, A., Wester, M., Lazarevic, D., & Brandt, N. (2018). Smart homes, home energy management systems and real-time feedback: Lessons for influencing household energy consumption from a Swedish field study. *Energy and Buildings*, *179*, 15–25. https://doi.org/10.1016/j.enbuild.2018.08.026

O'Brien, W., Gunay, H. B., Tahmasebi, F., & Mahdavi, A. (2017). A preliminary study of representing the inter-occupant diversity in occupant modelling. *Journal of Building Performance Simulation*, *10*(5–6), 509–526. https://doi.org/10.1080/19401493.2016.1261943

Ozturk, Y., Senthilkumar, D., Kumar, S., & Lee, G. (2013). An intelligent home energy management system to improve demand response. *IEEE Transactions on Smart Grid*, *4*(2), 694–701. https://doi.org/10.1109/TSG.2012.2235088

Panchabikesan, K., Haghighat, F., & Mankibi, M. El. (2021). Data driven occupancy information for energy simulation and energy use assessment in residential buildings. *Energy*, *218*, 119539. https://doi.org/10.1016/j.energy.2020.119539

Paparrizos, J., & Gravano, L. (2015). K-shape: Efficient and accurate clustering of time series. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, *2015-May*, 1855–1870. https://doi.org/10.1145/2723372.2737793

Pereira, P. F., & Ramos, N. M. M. (2018). Detection of occupant actions in buildings through change point analysis of in-situ measurements. *Energy and Buildings*, *173*, 365–377.

https://doi.org/10.1016/j.enbuild.2018.05.050

Pereira, P. F., & Ramos, N. M. M. (2019). Occupant behaviour motivations in the residential context – An investigation of variation patterns and seasonality effect. *Building and Environment*, *148*(October 2018), 535–546. https://doi.org/10.1016/j.buildenv.2018.10.053

Ren, X., Zhang, C., Zhao, Y., Boxem, G., Zeiler, W., & Li, T. (2019). A data mining-based method for revealing occupant behavior patterns in using mechanical ventilation systems of Dutch dwellings. *Energy and Buildings*, *193*, 99–110. https://doi.org/10.1016/j.enbuild.2019.03.047

Richardson, I., Thomson, M., & Infield, D. (2008). A high-resolution domestic building occupancy model for energy demand simulations. *Energy and Buildings*, *40*(8), 1560–1566. https://doi.org/10.1016/j.enbuild.2008.02.006

Robert Tibshirani. (1996). *Regression Shrinkage and Selection via the Lasso*. *58*(1), 267–288. http://www.jstor.com/stable/2346178

Roberts, N., & Thatcher, J. B. (2009). Conceptualizing and Testing Formative Constructs: Tutorial and Annotated Example. *Data Base for Advances in Information Systems*, *40*(3), 9–39. https://doi.org/10.1145/1592401.1592405

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sardá-Espinosa, A. (2019). Time-Series Clustering in R Using the dtwclust Package. *The R Journal*, *11*(1), 22. https://doi.org/10.32614/rj-2019-023

Satre-Meloy, A. (2019). Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models. *Energy*, *174*, 148–168. https://doi.org/10.1016/j.energy.2019.01.157

Satre-Meloy, A., Diakonova, M., & Grünewald, P. (2020). Cluster analysis and prediction of residential peak demand profiles using occupant activity data. *Applied Energy*, *260*(September 2019), 114246. https://doi.org/10.1016/j.apenergy.2019.114246

Schwartz, L., Wei, M., Morrow, W., Deason, J., Schiller, S. R., Leventis, G., Smith, S., Leow, W. L., Levin, T., Plotkin, S., Zhou, Y., & Teng, J. (2017). Electricity end uses , energy efficiency , and distributed energy resources baseline. *Energy Analysis and Environmental Impacts Division Lawrence Berkeley National Laboratory*, *January*, 77.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://www.jstor.org/stable/2958889

Singh, R. P., Gao, P. X., & Lizotte, D. J. (2012). On hourly home peak load prediction. *2012 IEEE 3rd International Conference on Smart Grid Communications, SmartGridComm 2012*, 163–166. https://doi.org/10.1109/SmartGridComm.2012.6485977

Standby Power Summary Table. (2018). https://standby.lbl.gov/data/summary-table/

Sugawara, E., & Nikaido, H. (2019). EIA energy outlook 2020. *Antimicrobial Agents and Chemotherapy*, *58*(12), 7250–7257. https://doi.org/10.1128/AAC.03728-14

Teeraratkul, T., O'Neill, D., & Lall, S. (2018). Shape-Based Approach to Household Electric Load Curve Clustering and Prediction. *IEEE Transactions on Smart Grid*, *9*(5), 5196–5206. https://doi.org/10.1109/TSG.2017.2683461

Thieblemont, H., Haghighat, F., Moreau, A., & Lacroix, G. (2018). Control of electrically heated floor for building load management: A simplified self-learning predictive control approach. *Energy and Buildings*, *172*, 442–458. https://doi.org/10.1016/j.enbuild.2018.04.042

Torriti, J. (2020). Temporal aggregation: Time use methodologies applied to residential electricity demand. *Utilities Policy*, *64*(March), 101039. https://doi.org/10.1016/j.jup.2020.101039

Touzani, S., Ravache, B., Crowe, E., & Granderson, J. (2019). Statistical change detection of building energy consumption: Applications to savings estimation. *Energy and Buildings*, *185*, 123–136. https://doi.org/10.1016/j.enbuild.2018.12.020

Vassileva, I., Wallin, F., & Dahlquist, E. (2012). Analytical comparison between electricity consumption and behavioral characteristics of Swedish households in rented apartments. *Applied Energy*, *90*(1), 182–188. https://doi.org/10.1016/j.apenergy.2011.05.031

Viegas, J. L., Vieira, S. M., Melício, R., Mendes, V. M. F., & Sousa, J. M. C. (2016). Classification of new electricity customers based on surveys and smart metering data. *Energy*, *107*(2016), 804–817. https://doi.org/10.1016/j.energy.2016.04.065

Vijayapriya, T., & Kothari, D. P. (2011). Smart Grid: An Overview. *Smart Grid and Renewable Energy*, *02*(04), 305–311. https://doi.org/10.4236/sgre.2011.24035

Wei, Y., Xia, L., Pan, S., Wu, J., Zhang, X., Han, M., Zhang, W., Xie, J., & Li, Q. (2019). Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks. *Applied Energy*, *240*(February), 276–294. https://doi.org/10.1016/j.apenergy.2019.02.056

Widén, J., & Wäckelgård, E. (2010). A high-resolution stochastic model of domestic activity patterns and electricity demand. *Applied Energy*, *87*(6), 1880–1892. https://doi.org/10.1016/j.apenergy.2009.11.006

Xiao, F., & Fan, C. (2014). Data mining in building automation system for improving building operational performance. *Energy and Buildings*, *75*, 109–118. https://doi.org/10.1016/j.enbuild.2014.02.005

Yang, J., Ning, C., Deb, C., Zhang, F., Cheong, D., Lee, S. E., Sekhar, C., & Tham, K. W. (2017). k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy and Buildings*, *146*, 27–37. https://doi.org/10.1016/j.enbuild.2017.03.071

Yao, R., & Steemers, K. (2005). A method of formulating energy load profile for domestic buildings in the UK. *Energy and Buildings*, *37*(6), 663–671. https://doi.org/10.1016/j.enbuild.2004.09.007

Yu, Z., Fung, B. C. M., Haghighat, F., Yoshino, H., & Morofsky, E. (2011). *A systematic procedure to study the influence of occupant behavior on building energy consumption*. *43*, 1409–1417. https://doi.org/10.1016/j.enbuild.2011.02.002

Yu, Z., Haghighat, F., Fung, B. C. M., Morofsky, E., & Yoshino, H. (2011). A methodology for identifying and improving occupant behavior in residential buildings. *Energy*, *36*(11), 6596–6608. https://doi.org/10.1016/j.energy.2011.09.002

Zhang, Y., Bai, X., Mills, F. P., & Pezzey, J. C. V. (2018). Rethinking the role of occupant behavior in building energy performance: A review. *Energy and Buildings*, *172*, 279–294. https://doi.org/10.1016/j.enbuild.2018.05.017

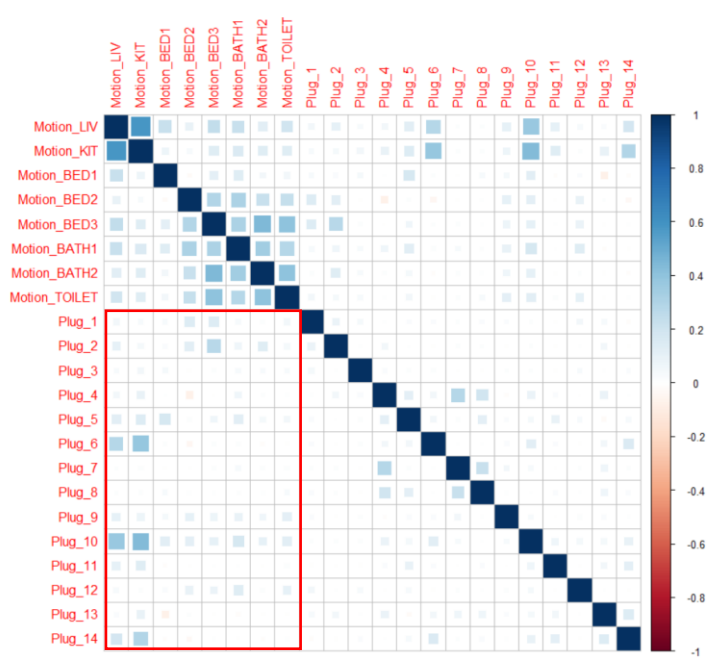
Zhao, J., Lasternas, B., Lam, K. P., Yun, R., & Loftness, V. (2014). Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining. *Energy and Buildings*, *82*, 341–355. https://doi.org/10.1016/j.enbuild.2014.07.033

Zhao, Z., Lee, W. C., Shin, Y., & Song, K. Bin. (2013). An optimal power scheduling method for demand response in home energy management system. *IEEE Transactions on Smart Grid*, *4*(3), 1391–1400. https://doi.org/10.1109/TSG.2013.2251018

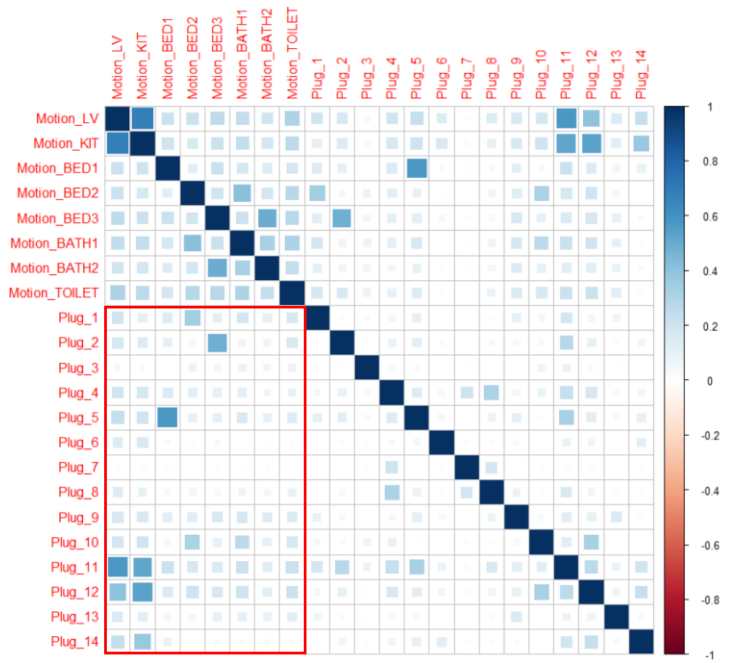
Zhou, B., Li, W., Chan, K. W., Cao, Y., Kuang, Y., Liu, X., & Wang, X. (2016). Smart home energy management systems: Concept, configurations, and scheduling strategies. *Renewable and Sustainable Energy Reviews*, *61*, 30–40. https://doi.org/10.1016/j.rser.2016.03.047

8. Appendix

Appendix A



(a) Apartment 112



(b) Apartment 162

Zonal labels of the apartment 112	Plug Power Variables
Bedroom-related	P1, P2, P5
Livingroom-related	P11
Kitchen-related	P6, P10, P13, P14
Bathroom-related	P12
Others	P3, P4, P7, P8, P9

Zonal labels of the apartment 162	Plug Power Variables
Bedroom-related	P1, P2, P5, P10
Livingroom-related	P11
Kitchen-related	P6, P12, P14
Bathroom-related	-
Others	P3, P4, P7, P8, P9, P13

Figure 8-1. Pearson correlation coefficients between plug variables and motion detection variables of apartment (a) 112 and (b) 162; the tables show the zonal labels of plug variables in each apartment

Appendix B

Table 8-1. Summary statistics of variables in apartment 112

Plug Variables in Apt. 112	Mean (W.h)	SD	25th percentile	50th percentile	75th percentile	Category (based on pattern of consumption)
BED_P1	3.86	16.70	0	0	0	3rd
BED_P2	0.70	1.56	0	0	0	3rd
OTH_P3	2.07	34.40	0	0	0	3rd
OTH_P4	18.45	0.81	18	18	19	1st
BED_P5	65.99	30.83	53	61	71	2nd
KIT_P6	10.22	70.46	0	0	0	3rd
OTH_P7	6.01	0.23	6	6	6	1st
OTH_P8	10.52	0.61	10	11	11	1st
OTH_P9	0.03	0.28	0	0	0	3rd
KIT_P10	33.80	33.51	19	23	32	2nd
LIV_P11	0.50	5.87	0	0	0	3rd
BATH_P12	0.96	16.52	0	0	0	3rd
KIT_P13	30.06	20.39	26	30	32	2nd
KIT_P14	1.70	15.94	0	0	0	3rd
LIGHTS	17.95	34.14	0	0	22	-

Table 8-2. Summary statistics of variables in apartment 162

Plug variables in Apt. 162	Mean (W.h)	SD	25th percentile	50th percentile	75th percentile	Category (based on pattern of consumption)
BED_P1	1.41	6.96	0	0	0	3rd
BED_P2	7.05	16.32	0	0	2	3rd
OTH_P3	10.93	61.87	0	0	0	3rd
OTH_P4	19.38	0.75	19	19	20	1st
BED_P5	17.07	23.10	6	7	20	2nd
KIT_P6	4.31	45.17	1	1	1	3rd
OTH_P7	5.82	0.41	6	6	6	1st
OTH_P8	11.83	0.57	12	12	12	1st
OTH_P9	14.50	80.65	0	0	0	3rd
BED_P10	1.11	2.76	0	0	0	3rd
LIV_P11	36.58	42.69	12	20	36	2nd
KIT_P12	58.94	42.55	38	42	55	2nd
OTH_P13	15.64	115.27	0	0	0	3rd
KIT_P14	10.88	63.33	0	0	0	3rd
LIGHTS	7.51	19.21	0	0	5	-

Appendix C

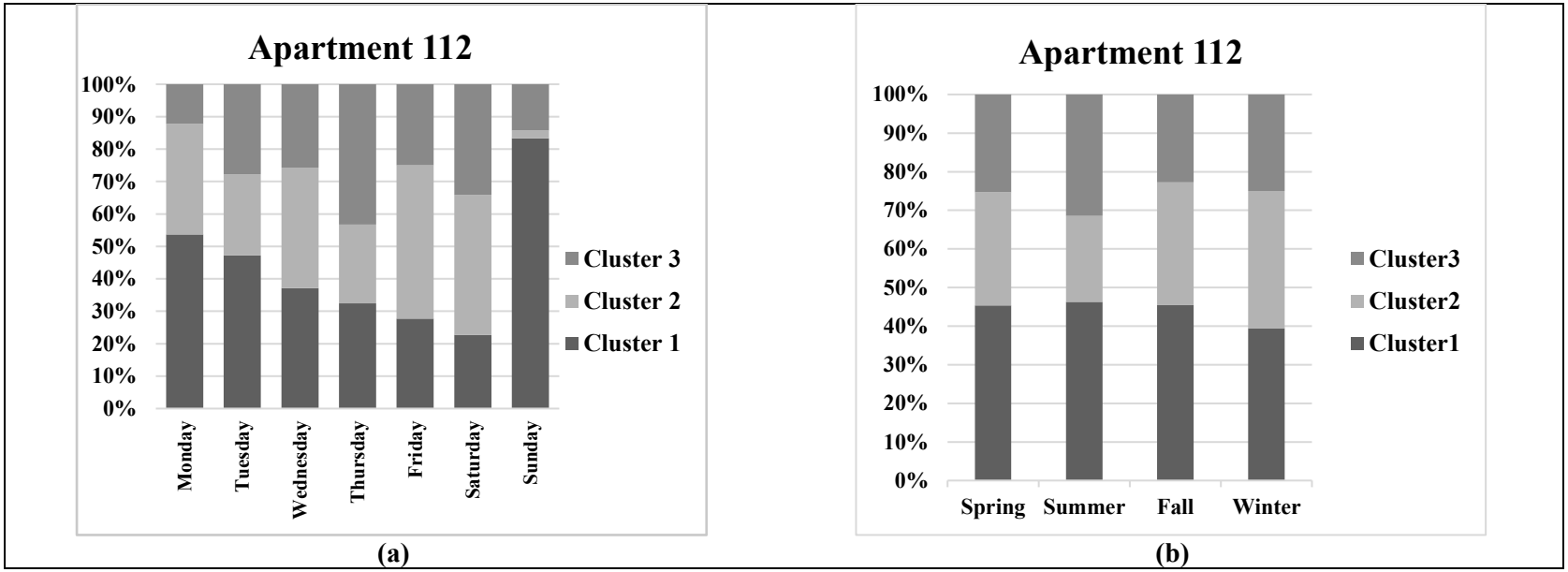


Figure 8-2. Distribution of occupancy patterns among (a) weekdays and (b) seasons in apartment 112

Appendix D

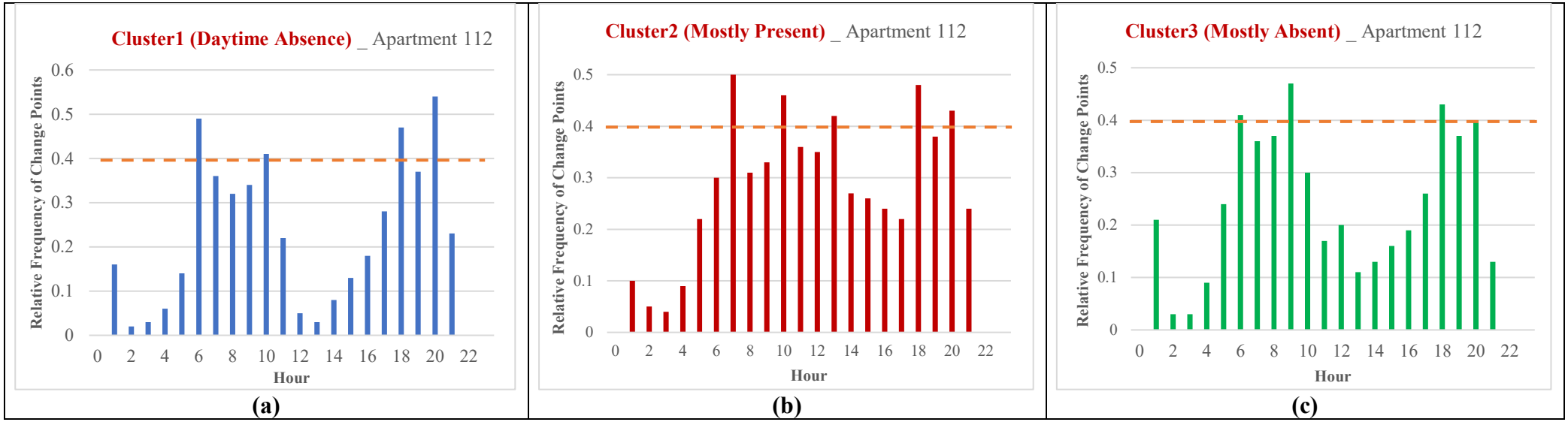


Figure 8-3. relative frequency of change occurrence at each hour in cluster1 (a), cluster2 (b), and cluster3 (c) of apartment 112

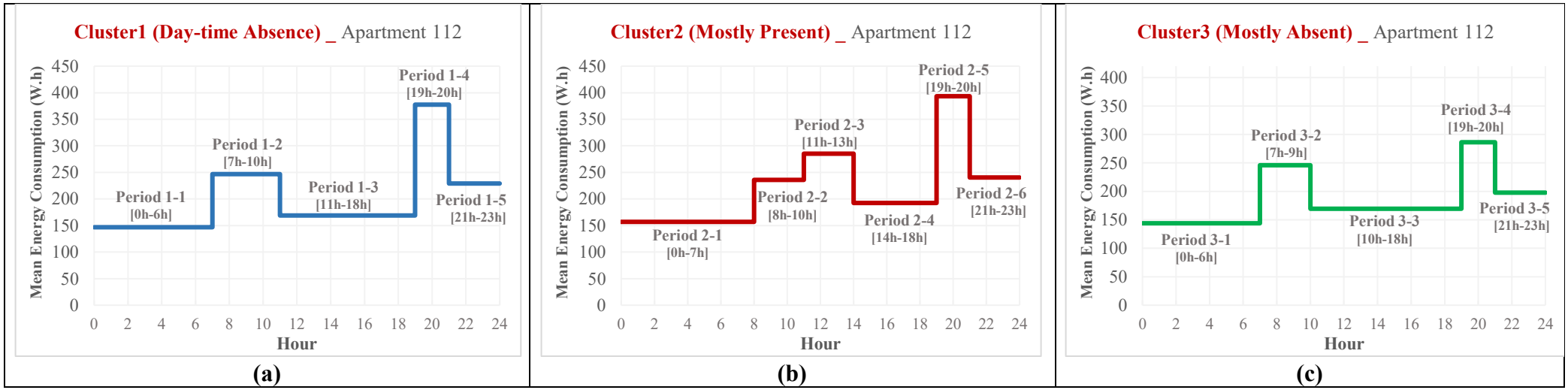


Figure 8-4. mean electricity consumption within each specified period of cluster1 (a), cluster2 (b), and cluster3 (c) in apartment 112

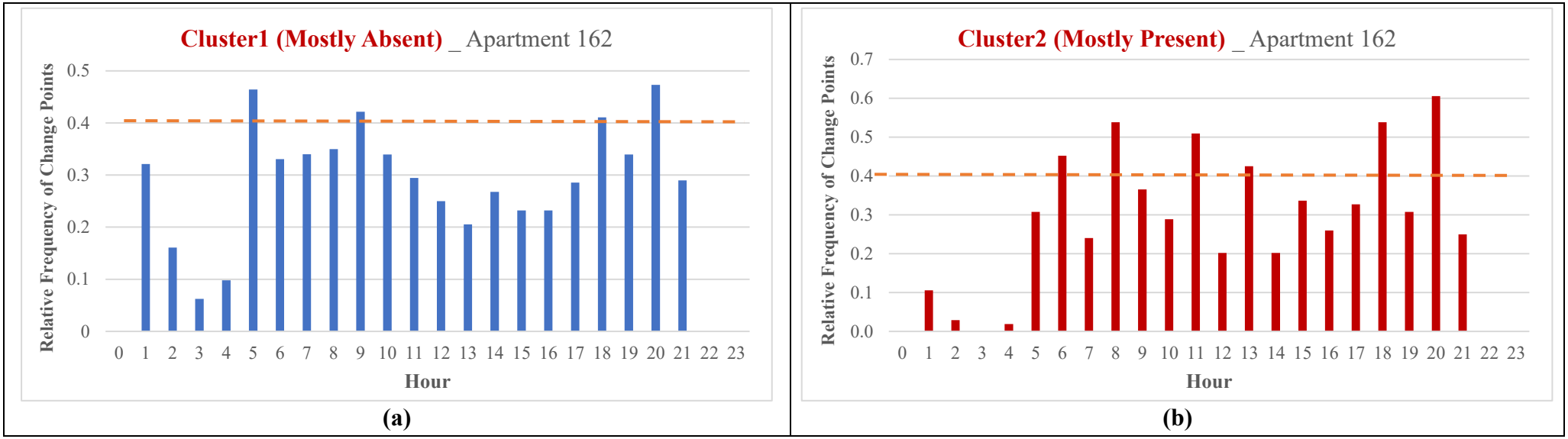


Figure 8-5. relative frequency of change occurrence at each hour in cluster1 (a) and cluster2 (b) of apartment 162

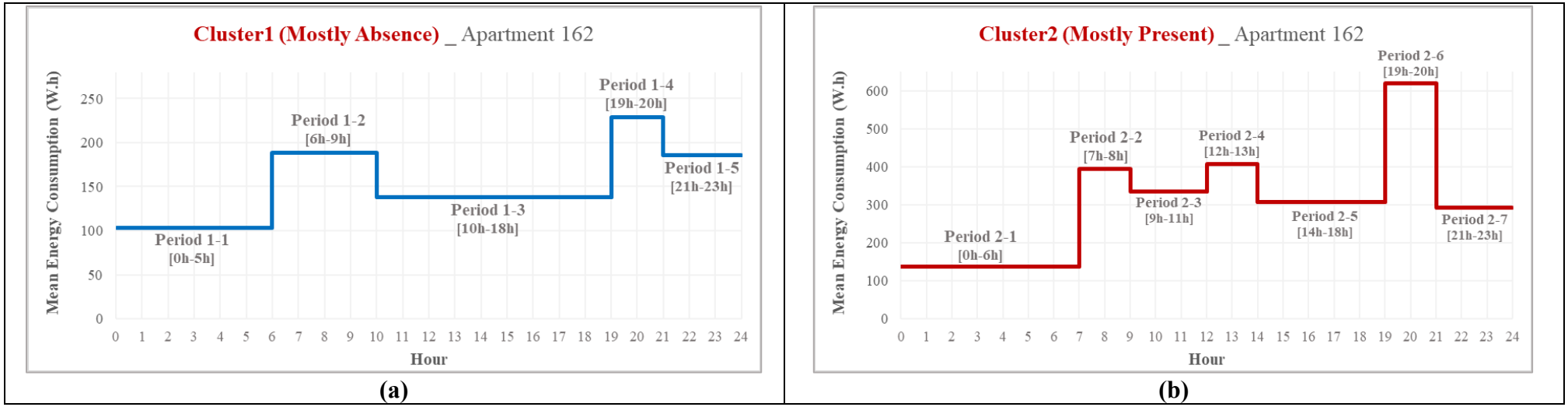


Figure 8-6. mean electricity consumption within each specified period of cluster1 (a) and cluster2 (b) in apartment 162

Appendix E

Table 8-3. Variance Inflation Factors (VIFs)

Apartment 112																	
BED_P1	**BED_P2**	**OTH_P3**	**OTH_P4**	**BED_P5**	**KIT_P6**	**OTH_P7**	**OTH_P8**	**OTH_P9**	**KIT_P10**	**LIV_P11**	**BATH_P12**	**KIT_P13**	**KIT_P14**	**LIGHTS**			
1.023	1.065	1.016	**253.128**	1.132	1.085	**400.813**	**180.315**	1.029	1.586	1.032	1.019	3.066	1.059	1.084			
Apartment 152																	
OTH_P1	**BED_P2**	**BED_P3**	**LIV_P4**	**LIV_P5**	**OTH_P6**	**BED_P7**	**KIT_P8**	**OTH_P9**	**OTH_P10**	**OTH_P11**	**OTH_P12**	**LIV_P13**	**KIT_P14**	**OTH_P15**	**LIV_P16**	**KIT_P17**	**LIGHTS**
2.872	1.098	1.022	1.054	1.080	**40.758**	1.193	1.096	**633.675**	**655.815**	1.087	1.278	2.993	1.154	1.060	1.363	1.233	1.308
Apartment 162																	
BED_P1	**BED_P2**	**OTH_P3**	**OTH_P4**	**BED_P5**	**KIT_P6**	**OTH_P7**	**OTH_P8**	**OTH_P9**	**BED_P10**	**LIV_P11**	**KIT_P12**	**OTH_P13**	**KIT_P14**	**LIGHTS**			
1.221	1.253	1.056	**393.943**	1.316	1.016	**438.685**	**316.675**	1.369	1.306	1.861	2.776	1.244	1.244	1.490			

Appendix F

Table 8-4. Cluster validation indices summary (adopted from (Satre-Meloy et al., 2020))

Index	Equation	Optimal case
Calinski-Harabasz (Caliński & Harabasz, 1974)	$$CH(q) = \frac{trace(B_q)/\,(q-1)}{trace(W_q)/\,(n-q)}$$ • W_q: the within-cluster dispersion matrix for data clustered into q clusters • B_q: the between-cluster dispersion matrix for data clustered into q clusters • q: number of clusters • n: number of data points	To be maximized
Davies-Bouldin (Davies & Bouldin, 1979)	$$DB(q) = \frac{1}{q}\sum_{k=1}^{q} \max_{k \neq l}\left(\frac{\delta_k + \delta_l}{d_{kl}}\right)$$ • $K, l = 1, \dots, q$, both indicates cluster number, and q is the number of clusters • d_{kl}: the distance between centroids of cluster $C_k$and C_l • δ_k: the average distance between each data point of cluster k and the centroid of cluster C_k	To be minimized
Dunn (Dunn, 1973)	$$Dunn(q) = \min_{1 \le i \le q}\left\{ \min_{1 \le j \le q,\, j \neq i}\left\{\frac{d_{ij}}{\max_{1 \le k \le q}\{\delta_k\}}\right\}\right\}$$ • $i, j, k = 1, \dots, q$ • d_{ij}: the distance between centroids of cluster $C_i$and C_j • δ_k: the average distance between each data point of cluster k and the centroid of cluster C_k	To be maximized
Silhouette (Rousseeuw, 1987)	$$Silhouette = \frac{\sum_{i=1}^{n} S(i)}{n}, Silhouette \in [-1, 1]$$ • $S(i) = \frac{b(i)-a(i)}{\max\{b(i),\ a(i)\}}$ • $a(i)$: The mean distance between a sample (the ith data point) and all other points in the same cluster • $b(i)$: The mean distance between the ith data point sample and all other points in the next nearest cluster • n: total number of observations (data points)	To be maximized

Appendix G

Quartile method find outliers using the following equations:

$$lower\ outliers = Q_1 - IQR$$

$$upper\ outliers = Q_3 + IQR$$

- Q_1: first quartile
- Q_3: third quartile
- IQR: inter quartile range ($Q_3 - Q_1$)

Knowing the median is the second quartile (Q_2), the first quartile (Q_1) is the value between the median and the minimum value, and the third quartile (Q_3) is the value between the median and the maximum value.