Data Mining Frameworks for Energy Consumption Reduction of Existing Buildings

Milad Ashouri Sanjani

A Thesis

In the Department

Of

Building, Civil, and Environmental Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy
(Building Engineering)
at
Concordia University
Montréal, Québec, Canada

September 2019

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Milad Ashouri Sanjani

Entitled: Data Mining Frameworks for Energy Consumption Reduction of Existing Buildings

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

Doctor Of Philosophy (Building 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. Akshay Kumar Rathore

____ External Examiner
Dr. Zhiqiang Zhai

____ External to Program
Dr. Amin Hammad

____ Examiner
Dr. Andreas K. Athienitis

____ Examiner
Dr. Fuzhan Nasiri

____Thesis Co-Supervisor
Dr. Fariborz Haghighat

____Thesis Co-Supervisor
Dr. Benjamin C.M. Fung

Approved by
Dr. Michelle Nokken, Graduate Program Director

November 4, 2019 ____
Dr. Amir Asif, Dean
Gina Cody School of Engineering & Computer Science

ABSTRACT

Data Mining Frameworks for Energy Consumption Reduction of Existing Buildings

Milad Ashouri, Ph.D.

Concordia University, 2019

Many technical solutions have been developed to reduce buildings' energy consumption, but limited efforts have been made to adequately address the role or action of building occupants in this process. On the other side, Building Management System (BMS) monitors the performance of buildings by recording the data to improve the building operation, control systems and maintenance. Usually, BMS produces a large volume of data throughout the year including information with regard to patterns of energy use, occupant behavior, etc. The availability of this huge data has created an opportunity to extract information to improve the building energy performance through leveraging powerful data analytic tools.

The objectives defined in this thesis lie in developing methodologies to find energy saving opportunities by analyzing data coming from occupants' energy consumption. Three tasks are defined in this thesis. The first task is to provide a recommender system to alert the occupants to take certain measures in order to reduce their energy consumption through end-use loads. Therefore, the quantification of potential savings is provided upon following recommendations. The proposed methodology is also capable to detect the energy saving measures performed by occupants. The second task focuses on a systematic comparison procedure between the buildings to make the occupants aware of their rank among other buildings and hence give them clues on how to improve their performance. The third task focuses on developing a framework to create a reference building acting as a reference for a given building. Therefore, the given building can be compared against its reference building. Potential savings are given to the given building along with directions how to achieve them. The results show successfulness of developed methodologies in finding energy saving opportunities through modifying occupant behavior.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Professor Fariborz Haghighat for his valuable comments in energy and buildings domain and funding my Ph.D. Program. I also express my gratitude to Professor Benjamin Fung for his amazing updated knowledge in artificial intelligence, and data mining domain that have guided me through the Ph.D. program. It is their guidance and support that has allowed me to mature as a Ph.D.

I thank my committee members, Dr. Nasiri and Dr. Athienitis, for providing valuable outsiders perspective to my research during my research proposal and comprehensive exam. Their advice and direction were greatly appreciated.

I thank all the Concordia Energy and Environment group members, past and present, who have contributed to both scientific and nonscientific discussions with a special thanks to those who have become life-long friends.

I would like to thank the Department of Building and Civil Engineering at Concordia University. It has been a joy to work within the department and that is in large part because of the friendly work environment.

Finally, I thank my family and friends who have always supported me. There have been laughs and tears along the way and I am very thankful to have shared those moments with them.

Contribution from Authors

Development of Building Energy Saving Advisory: A Data Mining Approach

Milad Ashouri	Data cleaning, analysis, writing, editing, and proofing
Fariborz Haghighat	Research supervisor, funding, editing, commenting, and proofing
Benjamin Fung	Research supervisor, editing, commenting, and proofing
Amine Lazrak	Commenting, and editing
Hiroshi Yoshino	Providing the dataset

Development of a Ranking Procedure for Energy Performance Evaluation of Buildings based on Occupant Behavior

Milad Ashouri	Data cleaning, analysis, writing, editing, and proofing
Fariborz Haghighat	Research supervisor, funding, editing, commenting, and proofing
Benjamin Fung	Research supervisor, editing, commenting, and proofing
Hiroshi Yoshino	Providing the dataset

Systematic Approach to Provide Building Occupants with Feedback to Reduce Energy Consumption

Milad Ashouri	Data cleaning, analysis, writing, editing, and proofing
Fariborz Haghighat	Research supervisor, funding, editing, commenting, and proofing
Benjamin Fung	Research supervisor, editing, commenting, and proofing
Hiroshi Yoshino	Providing the dataset

TABLE OF CONTENTS

List of Figures

List of Tables

1. Introduction

1.1 Background and Motivation

Energy consumption is a 21st-century global concern that is reaching a crisis point. This phenomenon is indirectly connected to others such as global warming, ozone layer depletion, climate change, carbon dioxide (CO2) accumulation, dependence on oil crisis, population growth, and drought. The normalized primary energy consumption and CO2 emission increased by 50% and 11%, with an average annual raise of 1.4% and 0.3%, respectively, over the period of 1990–2016, according to the International Energy Agency (IEA) [1] (see Figure 1-1). The global contribution of buildings to energy consumption has increased steadily by up to 20–40 % in developed countries [2]. According to Natural Resources Canada [3], more than 30% of the total secondary energy is used by residential and commercial buildings. These reports indicate the necessity for manipulating energy consumption in buildings for a sustainable future.

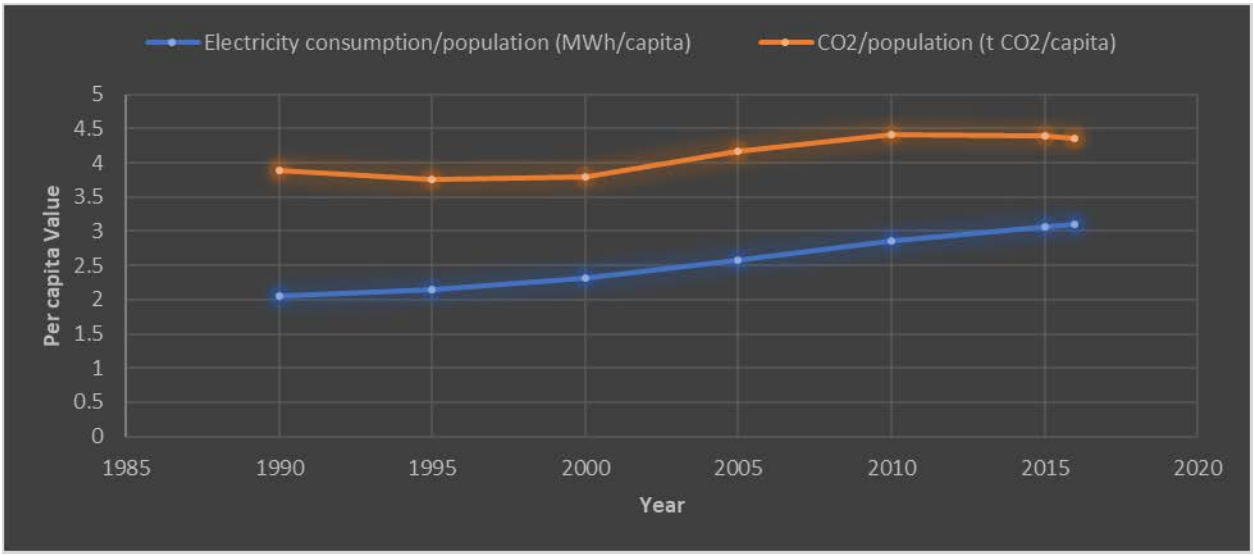

Figure 1-1 Primary energy consumption, CO_2 emissions and population growth over 1990–2016 (World data). Reports from IEA [1].

Although the IEA has reported that over the last two decades global energy efficiency has increased through measures such as renewable and green technologies, the clear link between energy consumption, population growth, and economic development endangers these improvements. Consequently, globalization, living style enhancement, and developing communication networks increase energy needs and create consumption patterns that may endanger future generations by exhausting the fossil fuel supply and increasing environmental damage [2].

Thoroughly understanding the major factors affecting energy consumption is necessary for applying energy reduction strategies. These factors comprise four major categories:

1) Building characteristics

   Building characteristics refer to all physical characteristics of the building (e.g., wall material and thickness, insulation, window-to-wall ratio, orientation, floor and wall area, etc.). Modern buildings are usually designed to achieve the maximal natural cooling, heating, and radiation while minimizing losses.

2) Occupant characteristics

   These account for occupants' presence, activities (what), and operation (how). Occupants can interfere significantly with a building's overall energy consumption [4]. Lifestyles are crucial in defining heating/cooling set points, indoor environmental quality required, window opening/closing behaviors, lighting, and other factors affecting consumption. Degree of education and energy costs are also among the factors that indirectly affect residents' behavioral patterns.

Occupant behavior refers to the presence of users, their interaction with appliances (TV, refrigerator, etc.), and user changes to lighting, air conditioning, temperature set points, and other energy-related settings.

3) System efficiency

   Building services systems and operations are also key factors in determining building energy consumption. These services include space heating/cooling and hot water supply, pumps, fans, and other heating and cooling system components. Apart from that, the efficiency of home appliances (e.g., oven, microwaves, washing machines, lamps, etc.) is also a consideration. Currently, no tool is available to notify building owners when aging, inefficient appliances require replacement.

4) Climatic conditions

   Climatic conditions are crucial when investigating the energy consumed by building systems and through occupant behavior. Outside temperature, solar radiation, humidity, and wind velocity are important factors. For example, in wintertime, the lower outside temperature increases heat load demand and encourages people to stay inside more, whereas shorter days increase lighting demands.

Because these four categories affect building energy consumption either directly or indirectly, understanding them thoroughly is a crucial task. However, simulation software packages must still overcome the challenge posed by the difference between designed/simulated energy consumption and actual building energy consumption, a difference that may arise from the complexity of these factors (e.g., uncertainty in climatic conditions, complexity of occupant behavioral patterns). Meanwhile, distinguishing among the main influential factors, identifying how each one contributes to overall energy consumption, and determining how they interact with one another also present a challenge for researchers. Until these factors and their interrelationships are clearly understood, energy consumption reduction strategies cannot be feasibly implemented.

One emergent area of applications science involves the development of data analysis tools to extract useful knowledge, provided that the data contain actual information about these influential factors. Data science refers to data mining, machine learning, statistics, data visualization, and other data analysis methods. To be more specific, “data science is the interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured” [5].

Modern buildings, particularly those in the public and commercial sectors, are increasingly equipped with monitoring tools (e.g., thermostats, sensors) to collect real-time data from building equipment to control their operation. This monitoring system is referred to as building automation system (BAS) [6]. This data consist of measurements from heating, ventilating, and air conditioning (HVAC) systems; ambient conditions; electricity consumption; lighting, noise; security systems; vertical transportation systems; and other systems necessary for a building’s operation. With advances in information systems technology, researchers have taken advantage of the availability of data spanning periods of months or years for knowledge extraction and prediction purposes. Analysis of these data provides novel methods to reduce building energy consumption and prevent system failures. According to the European Copper Institute [7], the potential energy savings from advanced building automation technologies may reach 22% by 2028. Applications of data mining in this domain can help reduce energy consumption by finding patterns of energy use, modifying occupant behavior, understanding occupancy schedule, predicting energy performance indices (e.g., energy rating), and detecting and diagnosing faults.

Building-related data exist in one or all the following forms:

1) Climatic data (outside temperature, solar radiation, wind velocity, humidity, etc.)
2) Operational data of HVAC systems (fans, supply air temperatures, mass flow rates, etc.), energy consumption data (end use loads of home appliances, electricity bills, etc.), and indoor environmental quality (IEQ; indoor temperatures, CO_2, human comfort, etc.)
3) Physical parameters (wall thickness, insulation, etc.)

These data include abundant information about building design, operation, and maintenance and can be used to help reduce building energy consumption. However, data mining is a relatively new science in the domain of building engineering; thus, few efforts have been made to apply data analysis tools to building-related data. However, a few data analysis frameworks (a series of integrated data analysis techniques for extracting information) exist to effectively mine the building-related data. Moreover, occupants and how they use a building's appliances appear to comprise one main energy leakage source. These gaps have motivated this research to contribute to data mining applications in building engineering by creating more data analysis frameworks to deal with the complexity of building-related data and especially occupant behavior.

1.2 Problem Statement

Various statistical data analysis processes have been applied in building studies along with simulations and modellings. Here, the goal is to establish a meaningful, appropriate relationship among energy consumption and influencing factors to reduce energy consumption more effectively and feasibly in the long run. However, simple statistical methods cannot capture the increased amount of data generated through the complex interaction of building systems and especially through occupant behavior; thus, they are inadequate for identifying correlations in the data that could improve performance. Therefore, problems such as accurately identifying the behavior of occupants by analyzing their data are still being addressed at a basic level and have not been answered thoroughly by traditional data analysis methods. The challenges to be considered are as follows:

- How can we accurately monitor the performance of occupants, based on their capabilities, without interference from other factors, to give them reliable, prioritized recommendations?
- How can we give occupants potential reductions in energy consumption upon following the recommendations so they can see the effect of their measures as a motivation?
- How can we report occupants' savings resulting from energy consumption in a way that would motivate them to increase their savings?
- How can we develop a process to assess the performance of a set of buildings based on their occupants' energy-saving awareness?
- How can we develop a process to give the building occupants a specific reference building with whose energy consumption they can compare their own?

To address these questions, we must look more deeply at building-related data to extract the information for analysis. This is only possible through the development of a systematic data analysis framework that makes it possible to generalize the methodology to be applied.

1.3 Data Analysis Tools and Methods

Analysis of recorded data is the core purpose of this research. Data analysis refers to using data mining (DM) to scrutinize the data in novel ways for information extraction. DM is the process of data analysis bringing together different techniques from statistics, machine learning, and pattern recognition to automatically extract hidden and unexpected interrelationships and patterns of interest from the dataset [8]–[12]. In other words, DM is "the analysis of large observational datasets to find unsuspected relationships and to summarize the data in novel ways so that data owners can fully understand and make use of data" [13]. DM has extensively been used in industry [12], but its application in the building energy field is relatively new. DM is applicable to real datasets, which yield reliable results and accurate predictions. Examples are in [12] and [14]. In addition, uncertainties regarding the boundary conditions (ambient temperature, solar radiations, etc.) and occupant behavior are already included in a real dataset [11], and the use of DM-generated models is often less time-consuming than physical systems themselves [15]. However, expert knowledge must be integrated into this combination of multidisciplinary DM approaches to maximize benefit.

1.4 Objectives and Purposes

The goal of this research is to develop a new framework for data analysis in the building engineering domain with a specific focus on occupants' role in energy consumption. The frameworks proposed here can help building owners or managers monitor building energy performance, estimate potential savings, and implement more profitable measures. A DM framework involves a series of consecutive methodologies applied to the data. Although each methodology been applied to various ranges of datasets in the building engineering domain, novel integration of them can yield pioneer insights regarding the data. To demonstrate the proposed frameworks, they will be applied to real datasets of 80 buildings located in different climatic conditions in Japan. The methodologies developed will surely answer the key problems outlined in chapter 1.2. This study's objectives can be summarized as follows:

1. Develop a methodology to recommend to occupants about how to reduce their end-use loads energy consumption and prioritize the recommendations:
    - Occupants of a building may not be aware of how much energy they use regarding a specific end-use load (e.g., hair dryer, oven, or room light) during the day.

Moreover, they use different appliances for certain tasks (e.g., cooking, laundry, or entertainment). By analyzing the energy consumption profiles of all individual appliances during a given period (for example, 1–6 months), we may notice correlations between appliances: for example, while the kitchen lights are on, the oven is used and the television is off. By searching through the data for these correlations and examining them, we may find instances of energy wastage and potentially generate recommendations for improving occupant behavior.

- An intended outcome is an accurate actionable quantitative report for occupants regarding potential energy savings upon following the recommendations. A report of achieved savings may motivate them to take more energy-saving measures. Chapter 3 provides more details.

2. Develop a systematic procedure to rank the energy performance of a group of buildings based on occupant behavior:

   Comparing energy efficiency between similar buildings is an effective approach for evaluating how efficiently building occupants are operating. This way, the residents of each building would know where they stand in relation to others and can be motivated to take energy-saving measures. In other words, a building's occupants can learn from one another, and observing that a similar building is consuming less energy would persuade them to reduce their own consumption. For example, if occupants of a single building see that the occupants of other similar buildings are using less energy to provide the indoor environment they require, they might be persuaded to do the same. However, the existence of several factors in energy consumption patterns of occupants (e.g., number of occupants, floor area, and house type) makes a simple comparison of the energy consumption of several buildings infeasible. Moreover, such comparisons do not account for the activities that occupants have performed to save energy and the degree of their energy consciousness. Thus, such simple comparisons are not yet effective. As many influencing factors as possible should be considered to make a fair comparison. In addition, if the procedure is well designed, it can reveal potential saving opportunities for building occupants. Chapter 4 provides more details.

3. Develop a systematic approach to provide building occupants with feedback to reduce energy consumption:

   This objective involves developing a methodology that would generate a reference building (RB) for any given building, regardless of its climate or characteristics, that would enable occupants of a given building to evaluate their own energy efficiency. In comparing two buildings regarding the same factors, we may notice, for example, that one of the buildings has lower energy expenditure toward a certain end-use load. This shows that some occupants are aware of certain activities or behaviors that could help reduce energy consumption. For example, the occupants of one building may use little energy for kitchen appliances, HVAC operation, or both. However, occupants may overuse some appliances such as televisions (or entertainment appliances more generally) that make them high-energy consumers regarding those specific appliances. This enables the creation of a building that has all the positive characteristics of low-consumption buildings so we can inform the occupants of the building in question of how much energy they can save by copying the RB. Chapter 5 outlines the RB concept in more detail.

The rest of this thesis investigates the tasks outlined in problem statement section 1.2. First, a complete literature review covers the state of the art in the domains of DM and building engineering. To address the challenges, this thesis defines three tasks, for which Chapters 3 through 5 provide detailed explanations, respectively, along with conclusions and future research directions. Chapter 6 provides a summary and closing remarks, along with overall limitations and future work.

2. Literature Review

2.1 Data

The data used in data mining (DM) are of great importance since modeling, analysis and decision making depend strongly on them. In an array of data set, each row is called a sample, tuple or record and each column is called a variable, feature or attribute. Data pre-processing and cleaning is to make sure that the data does not contain any missing value or outlier. Researchers have used various tools such as Euclidean distance [16], Hotelling T^2 method [17], and lower and upper quartile [11] to address this problem. The existence of outliers and missing values might bias the results. Highly relevant variables should not be used together since in some DM techniques, such as clustering analysis, specific aspects covered by these variables may be overrepresented in the clustering solution [18]. As an example, Wang [19] showed that using total number of degree days and total number of rooms for a multi linear regression (MLR) problem was not accurate since they were correlated linearly to the other involved factors. In addition, all relevant variables should be included in the DM process; otherwise, some important combinations of values or hidden patterns may be missing. Thus, feature/variable selection is an important preprocessing step.

Generally, recorded variables fall into one of the following two categories: climatic data (e.g. ambient air temperature, ambient relative humidity, etc.); and operational data, including those of the building (e.g. presence of occupants, temperature, humidity, flow rate, pressure, power, control signals, states of equipment, etc.). The climatic conditions greatly affect the energy consumption of buildings; therefore, climatic variables should be included in the DM dataset, but the operational data is different and depends on the objective of the study. The size of the data (both variables and samples) should be decreased (but at the same time be representative of the whole system) in order to reduce the computational time, simplify the problem, and reduce the cost of data gathering and storage, which are important in some applications [20]. Minimizing of the data size can be achieved by either reducing the number of records or reducing the number of variables. The number of records is usually selected to cover at least one year of operation to account for all the seasons (climatic conditions), while the number of variables differs in different applications. To reduce the number of variables, Pearson formula may be used to identify the critical variables, which are highly correlated with the variables to be modeled (outputs). An example was given in [21]. Another approach is to select the variables playing an important role in the energy balance and momentum balance of the system [17]. For HVAC systems, generally, temperature, flow rate, humidity and pressure are used to develop simulation and fault detection models [15], [17], [22]. Some other methods, such as principal component analysis (PCA), have also been applied for variables [23], [24] A PCA is defined as a linear transformation of the original variables. The original values can be represented in a new space with reduced dimensions, which is very useful in cases with many variables or noisy data, without much information loss. An example on total electrical energy consumption of households was presented by Abreu et al. [25].

2.2 Building Energy Performance Improvement

2.2.1 Discovering patterns of operation

Finding the patterns of operation plays an important role in building performance improvement. Generally, pattern recognition refers to finding specific trends in the data. Patterns of energy use are valuable knowledge for improving energy performance, i.e., design, optimization, maintenance, and fault detection. They can also be applied for energy use forecasting. Cluster analysis is an efficient tool for finding the patterns by grouping the data into k separate subsets.

Fan et al. [6] found the typical operation patterns of a building cooling system by clustering the data of energy consumption over a year. The results suggested two power consumption patterns; weekends and weekdays. A similar work [26] suggested three patterns using the same technique. D'Oca et al. [27] investigated the patterns of window opening and closing as an important factor in energy consumption of 16 office buildings using regression analysis and clustering. These patterns are interesting to be incorporated into simulation packages to produce more accurate load calculations. As an example, patterns of window opening duration during the day (long, medium or short intervals) affect infiltration and consequently the energy consumption. In addition to finding patterns, clustering helps isolate the effect of influential factors in energy consumption. For instance, do Carmo et al. [28] clustered hourly heat load data of 139 single family detached houses into three separate groups. The goal was to eliminate the effect of weather conditions in order to isolate the effect of household and building characteristics on the thermal load demand. Pattern discovery also helps find daily routines for each household. This can be compared to the best, worst or "normal" consumption for changing habits. Abreu et al. [25] developed a framework to extract the daily routines of households and patterns of energy consumption during the year. They identified the recurrent behaviors during the day (daily routines of households) by applying the PCA on the daily electricity energy consumption data. Moreover, Abreu et al. [25] extracted patterns of energy consumption using clustering as unoccupied period "baseline", cold weekend days, cold working days, and hot and temperate working days. Other examples of pattern discovery using clustering are occupancy schedule [29], energy consumption of domestic appliances [11] and indoor air quality [30].

Clustering along with other techniques have also been used to rate buildings in terms of energy performance. Wang et al. [19] applied a multi-criteria benchmarking method to rank 324 single family dwellings according to their energy performance indicator. TOPSIS method was used to rank the building energy performance. It was revealed that over three years, more than half of the buildings have a TOPSIS score above 0.50 showing an efficient performance. Also, clustering (K-means) was applied to TOPSIS results in order to rate the buildings from "excellent" to "unsatisfactory". This provided a framework to study the energy performance of the buildings and offered a rating system based on which instructions to building users can be given.

DM is also helpful to find associations between building variables. Sometimes these associations may reveal hidden and useful knowledge because they cannot be discovered through simulation programs. Examples are Yu et al. [11] who applied association rule mining on energy consumption of domestic appliances. Association rule mining extracted the highly correlated activities (such as TV [High]→ Ventilator [High]), which may be unexpected and are useful to guide occupants for energy saving recommendations. However, the extracted rules are qualitative (low and high). Fan et al. [6] applied quantitative association rule mining on a building cooling system to reveal the associations between its components. This can be used for maintenance, design optimization or recognizing the faulty conditions. Another example is finding associations between window opening and closing behaviors by D'Oca et al. [27]. The techniques of clustering and association rule mining can be used together for framework development in a step-by-step process [6], [8], [11], [26], [27], [31]. The process usually involves data pre-processing, pattern discovery, knowledge discovery, knowledge interpretation and selection.

2.2.2 Occupancy and occupant behavior

Identification and modification of occupant behavior is an important task in building energy data mining as the occupants have a major impact on the building energy consumption through appliances usage: hot water, lighting, window opening/closing, etc. However, due to the lack of efficient methods for identifying the role of occupant behavior in buildings as well as its inherent complexity, design expectations of building energy efficiency and its actual operation outcomes differ substantially [32].

Together with recent technological improvements in building industry such as design and operation of building systems [33], a vast number of studies are seen in the literature regarding occupant behavior. Two factors are common among the works. The first factor is related to stochastic and complex nature of occupant behavior and the second factor lies in interrelation between other influencing parameters in building energy consumption such as climate, building materials and characteristics, and economics. These two factors make it difficult to model the exact occupant behavior. With abundance of data in building through sensors (to measure human interactions such as movement. CO2 detection, windows, blinds), Wi-Fi signals, control system (for HVAC and lighting), in-depth analysis of occupant behavior has been enabled [34].

Currently, building simulation software can only incorporate occupant activities in a pre-defined manner such as occupancy presence, utilization of appliances, etc. Data mining can address this problem. Yu et al. [11] proposed a framework to isolate the effect of occupants from other factors of building energy consumption and also to reduce the overall energy consumption. D'Oca et al. [27], [32] extracted the pattern of use of windows by occupants in 16 office buildings as an indicator of occupant behavior. Also, they identified the most influential factors on window opening and closing behavior using a logistic regression. This information about each office can be used in simulation tools as well. Occupancy presence and scheduling (number of occupants in each hour of the day) is also modeled using data mining methods. Predicting the existence of occupants could affect the building energy modeling due to the internal heat gains of occupants

(Liang et al. [29], D'oca et al. [29], [35], and Sun et al. [36]). Cluster analysis and decision tree were used by Liang et al. [29] to improve the accuracy of the occupancy schedule. Liang et al. [29] showed that although ASHRAE standard 90.1 [37] provides an acceptable general estimation of occupancy pattern, it does not fit into any specific building. However, a whole-year data can be used as a reference very easily avoiding mathematical analysis (clustering, decision tree, etc.). The comparison should be carried out in order to see if there are any improvements in the proposed method. Sun et al. [36] showed that working overtime (presence of occupants in an office after the normal working hours) can directly influence the status of HVAC equipment; therefore, cooling energy consumption. They presented a stochastic model for working overtimes during weekdays based on the measured occupancy data from an office building. Results indicated that number of occupants and duration of their presence followed binomial and exponential distributions, respectively.

Pattern discovery also helps to optimize the operation of HVAC and reduce the unnecessary loads. Capozzoli et al. [38] used occupancy pattern analysis to reduce the HVAC energy consumption. By grouping similar occupancy patterns in the same thermal zone, they changed the HVAC control strategy to reduce the load while keeping the thermal comfort. Wi-Fi signals can give us a lot of information about the occupancy patterns, and how occupants interact with the building. Wang and Shao [39] used Wi-Fi signals to find occupancy patterns and used association rule mining to find energy wastage instances. It was claimed that up to 26% of lighting energy could be saved through this method. Kastner et al. [40] proposed a web-based intervention plan to encourage occupants to more energy efficient behavior through giving them recommendations (either habit-based intervention or knowledge-based intervention) and recording their actions. Field studies show that cultural differences of different sites influence implementation levels. In another study, Meinke et al. [41] used feed forward information to give the occupants information about their choice of cooling strategy (removing shirt, turning on ceiling fan, turning on air conditioner or tilting the windows) and how they impact energy use and their level of comfort. The result showed that most of the participants changed their action after getting the related information. This shows the effectiveness of giving occupants information and recommendation on their behavior. Knowing the effectiveness of feedback-based systems on household electricity use, Fischer [42] developed a psychological model to evaluate which features of feedback works best for the occupants. The feedback features are frequency, duration, content, breakdown, medium and way of presentation, comparisons, and combinations with other instruments. The results indicate that an effective feedback should be frequent, over a long time, be appliance specific breakdown, and be presented in an appealing way using computerized and interactive tools.

Association rule mining was used by Yu et al. [43] to find the relationship between the power consumption of different appliances (dishwasher, microwave, TV, etc.) and climatic parameters (temperature, relative humidity, etc.) to provide occupants with recommendations on how to reduce the appliance usage and to find energy inefficient behaviors.

The two factors mentioned in the beginning of the section 2.2.2 show the necessity of in-depth occupant behavior analysis to reveal new opportunities toward energy efficient buildings while maintaining the comfort.

2.2.3 Predicting energy performance indicators

Predicting energy performance indices, such as COP, energy efficiency or energy rating, is usually done in three steps: initial exploration, model building, and validation and deployment [21]. Decision tree, as one promising DM tool, has been used to predict energy consumption of HVAC systems [6], [14]. An example of such a decision tree is described in Fan et al. [6]. In this case, given an hour and month of the year, the decision tree can give the range of power consumption of an AHU. Another example can be found in Yu et al. [14] who developed a decision tree to classify the buildings into high and low energy intensities. The required buildings parameters to build the model play a crucial role and should be a complete representative of the building energy consumption. Decision trees can provide some more useful knowledge in addition to prediction. The variables from the top to the bottom of a decision tree are sorted according to their importance. For instance, in the decision tree developed by Yu et al. [14] and Fan et al. [15], annual average ambient temperature and hour of the day are the most important decisive variables, respectively. In addition, the error rate given by each leaf of the tree provides information about the reliability of the decision tree. For instance, Yu et al. [14] found that the number of misclassified labels as "Low" EUI (Energy use intensity) was less than "High". Therefore, the decision tree was more sensitive to "low" EUI buildings. This could be due to the fact that the original data contained more buildings with low energy consumption, which makes the tree more sensitive to that category. Similar interpretations were deduced from the decision tree developed by Liang et al. [29] on occupancy prediction.

Artificial Neural Network (ANN) is a powerful modeling approach used in building energy systems such as heat pumps, air conditioning and refrigeration systems [44]. The measured data in a certain boundary condition and time period are used in order to build a model. When the model is trained using experimental data, it can be used as an estimator instead of using mathematical equations (related to system physics such as mass and energy balances) to predict the system indicators (energy performance, energy efficiencies or physical variables such as temperature, mass flow rate, etc.) at other boundary conditions. However, care should be taken to avoid false predictions. In this context, the data training process plays a determinant role in developing a reliable and efficient model. Training data should be representative of the boundary conditions in which the model is used. For certain applications (static problems) some methods (like experimental design and Latin hyper cube techniques) could be applied to build the training data. For HVAC applications, this issue is more complicated since systems are dynamic and complex (thermal storage for instance has transient behavior during charging/discharging or the starting phase). However, some methods have been suggested for such scenarios [45]. In the framework of DM application in monitored buildings and districts, this is not an issue since usually a very large amount of data is available. Also, in this framework, the model is exploited in the same climate, which was used for training. Thus, the ANN encounters similar patterns to those learned during the learning stage.

Finding all variables, which directly or indirectly affect the performance indices (such as COP, energy efficiency, etc.), is difficult due to interrelations between subsystems and variables. On one hand, if all variables are chosen to train the network, there may exist correlations among them, which may result in less accurate network predictions due to over fitting problems. On the other hand, if an insufficient number of variables are chosen, the network may not have enough information to generate accurate predictions. Chou et al. [21] applied the Pearson correlation among parameters to find the most critical variables associated to COP and then performed different neural network methods on the refrigeration cycle. The application of ANN for predicting the system indicators such as energy consumption, efficiency, COP, cost, etc. is quite extensive. Examples are Belman-Flores and Ledesma [46] in case of a MLP (multi-layer perception) ANN and Bechtler et al. [46] in case of a dynamic neural network and Swider [47] who performed an MLP and RBF (radial basis neural network ANN).

2.3 Remarks and Challenges Regarding Building Energy Performance Improvement

As discussed earlier, knowledge discovery helps to identify the hidden patterns in the data and provides engineers with an insight to predict the behavior of the system, which is useful for energy auditing and modern control systems. However, one major shortcoming of knowledge discovery is the inability of DM to provide useful knowledge by itself. In other words, an expert should interpret and analyze the results of DM, which might be difficult for some cases. Another important issue is the large volume of knowledge or patterns generated by DM, which may overwhelm the expert. A solution to address this challenge is application of interactive data mining. The idea is to ask users to select some patterns, which they are interested in. Then, the system can interactively find more patterns, which are relevant or similar to the chosen ones. Another important challenge is the availability of numerous methods in DM field with its own pros and cons (nearly all of the machine learning techniques are applicable to DM). Selecting the most relevant methods for knowledge discovery requires domain experience. The availability of sufficient data is an issue too. For instance, to extract knowledge from a heating system one needs at least data from the whole heating season to find frequent and reliable patterns. If one extracts results from a limited time such as one week, the data may be misleading (misleading rules) since the frequent rules for one week may not be applied to the whole seasons.

To conclude, the advantages and disadvantages of knowledge discovery are summarized as the following:

- Knowledge discovery via DM helps to identify the hidden patterns in the data,
- It provides engineers an insight to predict the behavior of the system,
- It is beneficial for energy auditing and modern control systems,

- DM is unable to provide useful knowledge without the need of an expert to interpret this knowledge,
- Selecting the most relevant methods for knowledge discovery requires domain experience,
- The conclusions are really specific to each set of data or case study and they cannot be generalized.

As mentioned earlier, predictive tasks utilize the inputs and outputs of one system for training the model in order to predict the energy performance indices. However, some challenges and shortcomings exist:

- Input selection is important in developing a reliable ANN model in order to avoid the over fitting problem. The inputs should directly affect the outputs,
- The time needed to train a network for huge data is exhausting,
- More recently, some advanced neural networks have been introduced, which generate more accurate predictions with less root mean square error (RMSE). For instance, Mocanu et al. [48] applied deep learning method (FCRBM) to predict building energy demand and compared it with support vector machine and neural networks. Results showed less RMSE in most cases. Application of such more advanced methods needs to be investigated in building energy systems, as well, and
- The ANN lacks explanation tools for its knowledge. In most cases, ANN relations cannot be explained physically.

3. Development of Building Energy Saving Advisory: A Data Mining Approach

3.1 Introduction

Buildings' impacts on global energy consumption have been increasing steadily, reaching up to 20-40% in developed countries [2]. According to Natural Resources Canada [3], more than 30% of the total secondary energy is used by residential and commercial buildings. These reports show it is necessary to control energy consumption in buildings to ensure sustainability.

Clear understanding of major influencing factors in building energy consumption is the necessary step when determining energy retrofitting strategies. The influential factors can be divided into 4 major categories: **Building Characteristics,** all physical features of the building such as wall material and insulation; and **Occupant Behavior,** which include their presence, activities (what), and operation (how). Tenants can regulate the overall building energy consumption greatly [4], such that buildings with same physical characteristics have large discrepancies in electricity consumption. The influential factors are changing heating/cooling set points, the indoor environmental quality required, windows opening/closing behaviors, lighting, etc. **System Efficiency and Operation** refers to space heating/cooling and hot water supply, pumps, fans, etc. The efficiency of home appliances (oven, microwave, washing and drying machines, lamps, etc.) should not be neglected either. The fourth factor is **Climatic Conditions,** which refer to outside temperature, solar radiation, humidity, and wind velocity. Simulation software packages consider all these factors to model building performance, yet the designed/simulated building energy consumption differs from the actual one. This challenge may originate in the complexity of these factors (e.g. uncertain climatic conditions or complexity of occupant behavioral patterns).

One recent emerging science is the development of data analysis tools to extract information, and patterns, hidden in data. Data science refers to data mining, machine learning, statistics, data visualization, and various data analysis methods [5]. Building monitoring systems (Building Automation System (BAS)) measure consumption of Heating, Cooling and Air Conditioning (HVAC) systems, ambient conditions, electricity consumption, lighting, noise, security systems, vertical transportation systems, etc. This data includes abundant information about the building's design, operation and maintenance, and can be used to reduce building energy consumption as well as recognizing faulty conditions.

However, data mining is a relatively new science in building engineering; thus, little effort has been made to apply data analysis tools to building industry. Few data analysis frameworks (a series of data analysis techniques integrated together to extract information) exist to mine the building related data effectively (some of them will be mentioned in the literature review). Also, one of the key approaches to reduce building energy consumption can be started from the occupants and how they use the appliances. Huge savings are possible by modifying their behavior [49]. It is the only way to reduce energy consumption without costly fundamental changes such as upgrading building systems, reconstruction, envelope renovation etc.

3.2 Methodology

3.2.1 Overall Approach: Data Analysis Framework

Energy reduction strategies could be achieved by focusing on one or all the mentioned factors. This study aims to improve the building energy efficiency by focusing on occupant behavior. One building with fixed characteristics and system efficiency (equipment) is investigated, and then evaluated based on its energy consumption before and after applying reduction strategies. The difference shows the net effect of these measures. The inputs of the system are detailed energy consumption of the household appliances and weather data. The outputs of the system are prioritized recommendations and quantified reports of energy savings upon following them. Also, the system detects, and estimates energy savings done by occupants. Figure 3-1 shows the overall approach. Three data mining tasks (clustering, association rules mining (ARM, and ANN) are linked successively together. Clustering reduces the effect of weather conditions; ARM finds the appliances correlations in each cluster and ANN builds models on each analyzed rule. Detailed explanations are given in the following sections.

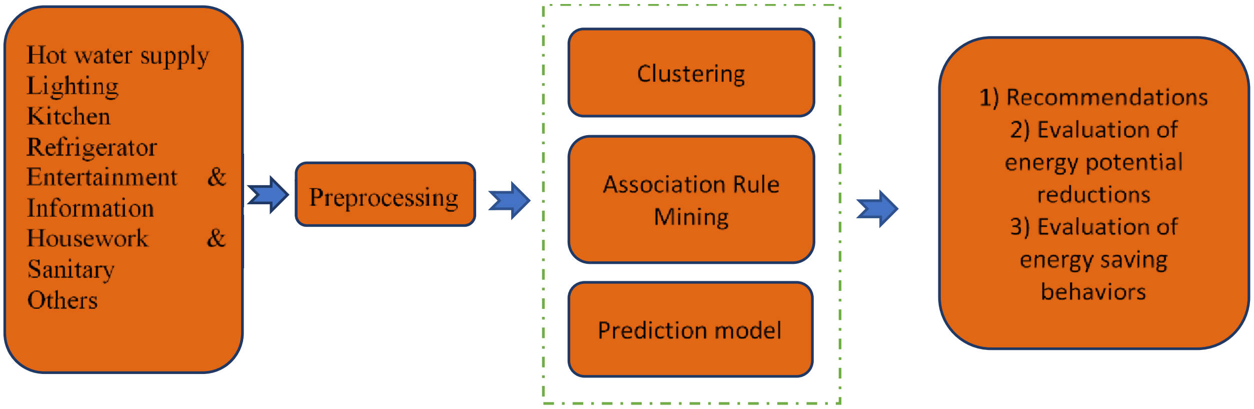

Figure 3-1. The proposed methodology for the recommender system. Inputs are detailed energy consumption of end-use loads and weather data. The outputs are prioritized recommendations and quantification of achieved energy savings along with the potential savings.

Reducing the effect of weather, energy consumption data is grouped by weather conditions and each cluster is analyzed separately. Figure 3-2 illustrates the data mining process. Correlations between different occupant activities are derived using association rule mining, applied on each cluster separately. The extracted rules are analyzed and categorized into three categories (RM, RS, and RN) based on the definitions given in section 3.2.4. The recommender system then uses occupants' previous behavior (the extracted rules) to train neural network models, which becomes the basis for judging whether the occupant behavior improved or needs modification.

The database consists of historical information about occupants' energy use. It must be updated as it registers any behavioral changes such as changes in occupant number, lifestyle, and activities. After the learning phase, the extracted knowledge is applied to a new dataset for investigation (See Figure 3-3). The historical data must cover all weather and behavioral patterns and could be six months, one year, two years, or more depending on the availability and resolution (Every minute, hourly or daily) of the dataset. The recent test data should be the current time to monitor the energy consumption in a real time. However, in this study, results are shown for days due to unavailability of data. This does not affect our approach because the methodology remains the same.

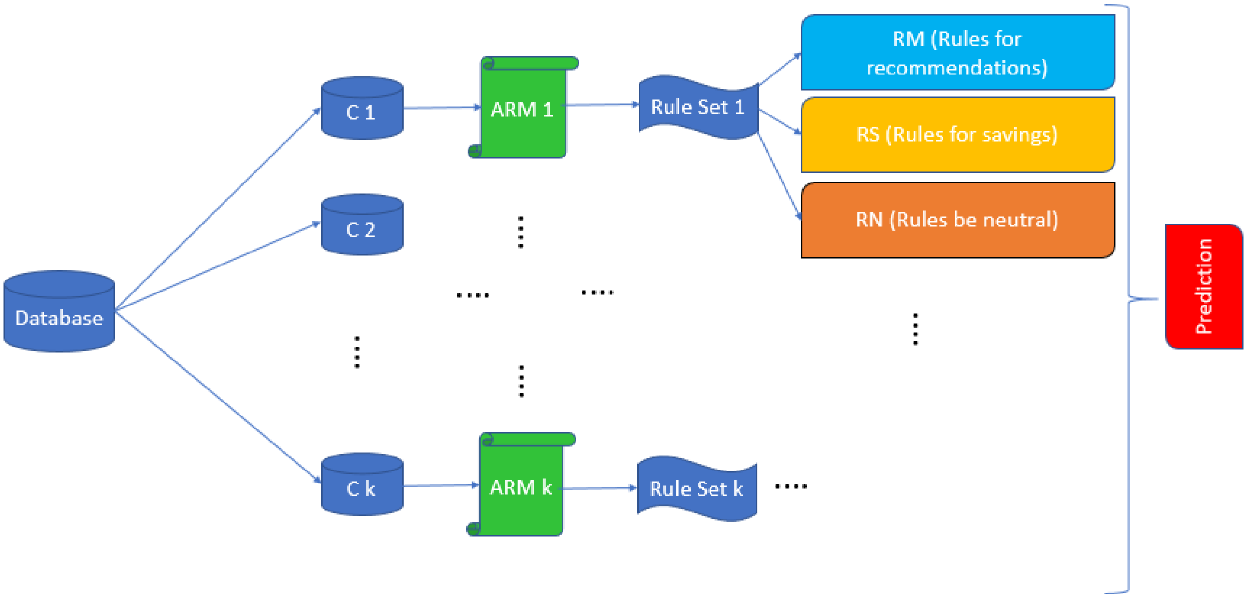

Figure 3-2. The data mining framework for knowledge discovery of occupants. The framework includes clustering, association rule mining, and neural networks to give the occupants useful information about their behavior in energy consumption of end-use services. The ARM (association rule mining), clustering and prediction models are detailed in sections 3.2.3 to 3.2.5.

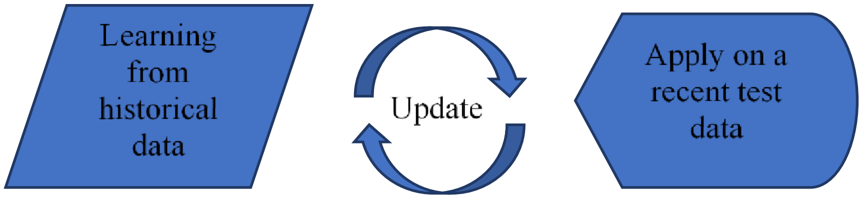

Figure 3-3. The learning and applying process. The learning phase is based on historical data (6-month, one year or more), the applying process is the most recent data (current day, week, or month)

3.2.2 Data Preprocessing

The data was extracted from a project entitled "Investigation on Energy Consumption of Residents All over Japan" [50]. The project was carried out by the Architectural Institute of Japan (AIJ) from December 2002 to November 2004 to evaluate and improve the energy performance of the buildings [50]. Field surveys on energy consumption were carried out in eighty residential buildings in six various locations in Japan including Hokkaido, Tohoku, Hokuriku, Kanto, Kansai, and Kyushu [11], [14]. Table 3-1 shows the survey items corresponding to investigation methods, and time intervals. More information on measurement methods are found in [11].

Table 3-1. Investigation methods and items.

Method	Items	Measurement interval
Field measurements	• Home appliances energy usage (Electricity, Gas, and Kerosene) • Climatic data (e.g. indoor air temperature, humidity, wind speed, etc. 1.1m above ground)	• Daily averaged values • Hourly averaged values (original data resolution: 15 minutes)
Questionnaire	Number of occupants, equipment uses, annual income, etc.	—
Inquiring survey	Building characteristics (building types, area, heat loss coefficient, equivalent leakage area, etc.)	—

Data cleaning is applied to enhance the quality of raw data by removing outliers and inconsistencies while considering missing values. Outlier detection and removal methods used in literature are domain expertise [6], lower and upper quantile (Q) [12], complete case analysis [51], simple moving average method [6] and inference based methods. By using other attributes of a given instance, the chance of predicting a missing record close to its real value is relatively high. These methods are more complex and time consuming but more accurate and reliable. In this study, the outliers are detected using the quantile method and then estimated using a regression model based on other available attributes. To avoid depending on the choice of measurement units and to speed up the learning process of neural networks [26], Min-max normalization was performed on the data as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}(x'_{max} - x'_{min}) + x'_{min} \tag{3.1}$$

where indices min and max refer to minimum and maximum values of each range. This technique preserves the relationship between the initial data.

3.2.3 Clustering

Clustering is used mainly in pre-processing large datasets, identifying outliers, discovering patterns or any data segmentation. Clusters are internally coherent and externally separated. In this study, clustering was used to group data by weather conditions. In regions with high temperature fluctuations, one may obtain several clusters, and only one in tropical regions. The attribute used for clustering is the 24-hour outdoor temperature data. Similarities between records are evaluated using distance-based criteria (e.g., Euclidean or Manhattan metrics). Various clustering algorithms exist in literature that depend on data dimensionality, type, and distribution. K-Means has been used successfully across various fields. This algorithm tries to separate the data points in (k) groups of equal variances to minimize the inter-cluster's sum of squares. It divides a set of n records (X) into k disjoint clusters (c), each described by the mean (μ_j) of the samples (x_i) in the cluster. Means are commonly called “centroids” and are representatives of the corresponding clusters. K-Means has to choose centroids that minimize inertia, or the within-cluster sum of squared [52]:

$$obj = \sum_{i=0}^{n} \min_{\mu_j} \|x_i - \mu_j\| \tag{3.2}$$

Similarly, k-Medoid uses the Euclidian distance to find the nearby data points and cluster them. The only difference between this method and K-Means is that it chooses the center of clusters from the most centrally located data points, not the average. This prevents clusters from being affected by outliers.

Hierarchical clustering belongs to the family of clustering algorithms that builds nested clusters by splitting or merging the dataset. In this study, a bottom-up agglomerative clustering that merges nodes having the least similarity in their Manhattan distance is being used.

Performance evaluation of clustering algorithms is performed using external validation methods when true labels exist (e.g. mutual information, F-measure, etc.) [53]. But, if the ground truth labels are unknown for a dataset (e.g. sample labels are unknown), evaluation must be done using the model itself using metrics such as silhouette index, Dunn Index, Calinski-Harabaz index, and Davies-Bouldin index. These evaluation metrics tests the clusters to see whether they satisfy some assumption that members belonging to the same class being more similar than members of different classes according to some similarity metrics [54]. A higher Silhouette Coefficient score means a model with better-defined clusters. The silhouette coefficient ranges between -1 to 1 (1 means highly dense clustering, and -1 false clustering). The Silhouette Coefficient is defined for each sample as follows:

$$s = \frac{b - a}{max\ (a, b)} \tag{3.3}$$

where a is the mean distance between a sample and all other points in the same class, and b the mean distance between a sample and all other points in the next nearest cluster. The Silhouette coefficient is the mean of all samples in the dataset. Another clustering evaluation method is the Dunn index, the ratio of the minimum cluster distance between observations in separate clusters to the maximum intra-cluster distance. The Dunn Index has a value between zero and infinity, and the highest value gives the best clustering.

3.2.4 Association Rule Mining (ARM)

Association rule mining is the next step of process and is applied on each cluster separately (see Figure 3-2). ARM is an unsupervised learning process and usually used for items frequently associated, meaning that they happen together. It was first used in market basket analysis to identify items frequently bought together. Support and confidence are the validity and certainty of the association rule. Support is an indication of how frequently the itemset appears in the dataset. The support of X with respect to Y is defined as the proportion of the transactions T in the dataset which contains the itemset X and Y. Confidence is an indication of how often the rule has been found to be true. The confidence value of a rule, X→Y, with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y.

There are various thresholds for both indicators that show the effectiveness of the rules. For example, a confidence level of 100% ensures that based on the data, two items are bought together all the times (such as a phone and case, or a laptop and its charger). ARM has been used in diverse fields such as sociology, bioinformatics, and retail [12].

Mathematically, support and threshold are defined as follows:

$$Support\ (X \to Y) = P(X \cup Y) \tag{3.4}$$

$$Confidence\ (X \to Y) = P(Y|X) \tag{3.5}$$

$$Lift\ (X \to Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = \frac{P(Y|X)}{P(Y)} \tag{3.6}$$

Lift demonstrates the correlation between of X and Y. When Lift > 1, there is a positive relationship between premise and conclusion; when Lift < 1, there is a negative correlation. If Lift =1, there is no correlation. Therefore, one only considers rules with minimum support (Sup) and confidence (Conf) threshold, and a Lift > 1. There are two popular algorithms for association rule mining, Apriori and frequent-pattern growth (FP-Growth) algorithm [12]. In this study, the FP-growth is used due to its high speed and wide applicability. To find associations between various activities of occupants, ARM is applied to the dataset. All attribute values are set to high or low based on whether the value is between average and maximum value or between minimum and average value, respectively. For simplicity, attributes are coded (1 to 17) according to Table 3-2. Also, the zero in the right side shows low and 1 shows high consumption. For example, 20 means the attribute number 2 has low energy consumption (the zero shows low energy consumption), while in number 81 the attribute number 8 signifies high energy consumption (1 indicates high energy consumption).

Table 3-2 shows a selection of the result of this process. All attribute names are listed. Further discussion will be presented in section 3.2.

Rules Categorization

There is a set of rules that must be categorized into three distinct groups for further analysis. The groups are **Rules for Modification (RM), Rules for Savings (RS),** and **Neutral Rules (RN)**, and each is defined in the sections below.

Rules for Modification (RM)

Attributes are divided into weather directly modifiable (e.g. refrigerator, lights, etc.) or indirectly modifiable (e.g. temperature or humidity). If any modifiable attributes with LOW value are in the conclusion, this rule is categorized as RM. Inspecting a new dataset (separate from the training data), any record respecting the premise but not the conclusion is found and flagged as inefficient. This is because, based on the rule, the conclusion is HIGH instead of LOW. Therefore, the occupants are informed that they should correct their behavior. Given that the obtained rules are frequent, it means that the occupants have shown such behavior before, and modifying their behavior is fully feasible.

Rules for Savings (RS)

If any attributes (directly modifiable) with HIGH value are in the conclusion, this rule is categorized as rules for savings (RS). By inspecting a set of new data (apart from the training data), any record respecting the premise but not the conclusion is found and identified as efficient. The reason is that based on the found rule, the conclusion should be HIGH, but it is LOW. The occupants are notified that they have saved energy.

Rules be Neutral (RN)

Rules whose suitability cannot be judged are found based on the data. Some rules can be modified to improve occupants' behavior and lower energy consumption, and some rules should be left unchanged.

3.2.5 Prediction Model

Prediction models are meant to build a model based on given inputs and outputs (historical data) that predicts outcomes using a new set of inputs. For example, to predict the price of a house based on information such as area, age, and location it is possible to train an ANN model on a set of area, age, and location data as the inputs and the house price as the output. Artificial neural networks (ANNs) are among such predictive models. Using a neural network model, potential savings and achieved savings are estimated. Let us consider the following example. According to an RM rule ([20, 30, 120]→[80], samples should be [20,30,120,80] (see Table 3-2 for a detailed description of numbers and their meaning). However, some samples are organized as [20, 30, 120, 81]. Based on these samples, occupants are warned that on some certain days the attribute number 8 has had high energy consumption (81) as opposed to the normal behavior. This may show an energy wasting behavior or abnormal use. The next step is estimating savings upon following the recommendation (81 to 80). Given that there is a correlation between the said attributes, the [20, 30, 120, 80] data is used to build a model to predict attribute 8 using attributes 2, 3, and 12 as inputs. After training the model, the records that respect the premise but not the conclusion ([20, 30, 120, 81] form) are plugged in the model. These samples are transformed to [20, 30, 120, 80] using the model. The difference between energy consumption (before and after correction, 81 and 80) will indicate the potential energy reduction for the considered attribute.

The same process is established for RS rules. This means that in general consumption is usually high, but low on certain days, showing energy preservation. The achieved savings are estimated using a similar approach. It is important to note that the procedure is performed for all extracted rules in RN and RS (one ANN model per rule). The cumulative effect of these savings can be a great contribution (that depends on the occupant behavior) to reduce energy consumption of buildings for a period of time.

3.2.6 Prioritizing Recommendations

Recommendations are ranked according to the amount of savings respecting the following recommendations. The amount of saving potentials is the cumulative savings upon correcting a behavior. For instance, if one tries to correct a behavior b [High] → a [Low], where ‘b’ and ‘a’ are energy consumption profiles before and after modification, the cumulative saving will be:

$$\% \, Savings = \frac{sum(b) - sum(a)}{sum(b)} \tag{3.7}$$

Maximum saving is reached when sum (a) = 0 and the minimum is no savings which means sum (a) = sum (b). In this study, any measure producing savings over 25% is considered as a high recommendation.

3.3 Results and Discussion

As mentioned in Figure 3-1, the process is designed to capture any abnormal behavior seen in the recent data based on analysis of historical data. Data is clustered based on outside hourly temperature so the energy consumption data shares similar weather conditions. Association rules find all interesting patterns in data which make the basis for alerting the tenants when some behavior dissimilar to these patterns is seen. Artificial neural networks are used to quantify the energy saving potentials and achieved savings by occupants. Following sections describe the results of each task in more detail.

3.3.1 Clustering Results

The open source software Python [55] and its open source libraries were used in this paper. Python libraries Numpy, Pandas, Matplotlib, and scikit-learn package were used in data mining tasks. Clustering was performed on the dataset consisting of 24-hour outdoor temperature data to group days into similar weather conditions. Three popular algorithms K-Means, K-Medoids, and Hierarchical were used to cluster data. Two clustering evaluation criteria were used for selecting the optimal number of clusters along with the clustering algorithm (See Figure 3-4 and Figure 3-5 for results).

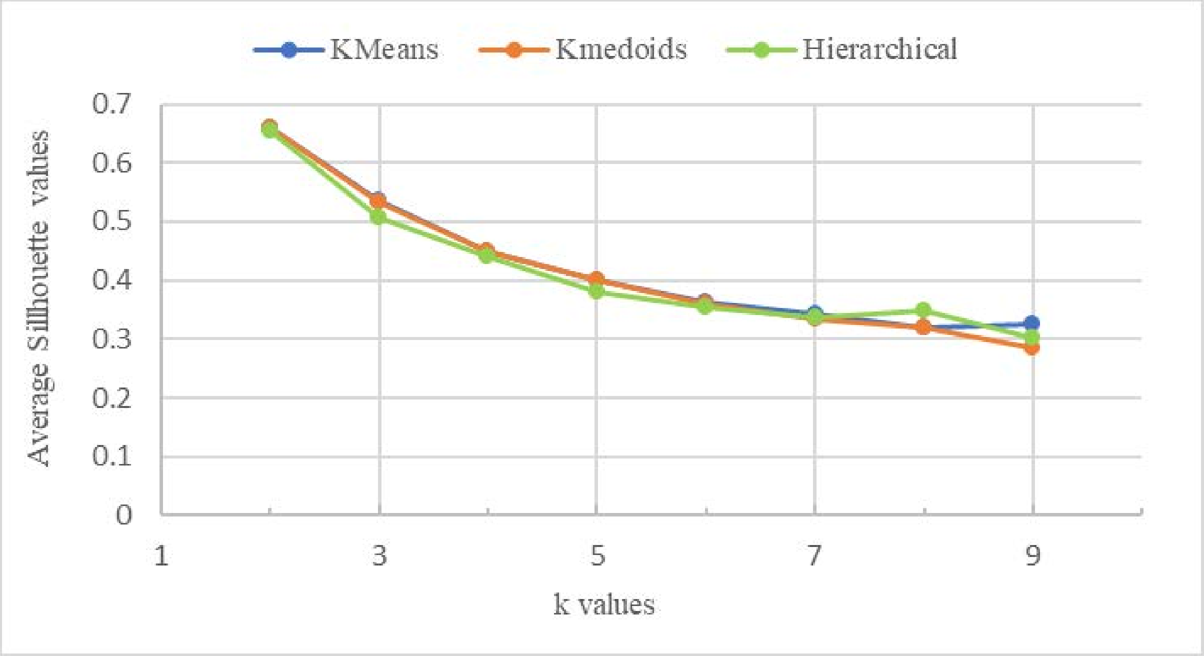

Figure 3-4. Clustering validation results with three different clustering algorithms and silhouette index as the evaluator. K values are the number of clusters. Three clusters roughly give the same silhouette index. The optimum number of clusters is two.

Figure 3-4 and Figure 3-5 show the results of three clustering algorithms when Sillhouette index and Dunn index are used as the clustering criterion, respectively. All three algorithms perform best when the number of clusters is two. K-Means was used as the clustering algorithm in this study.

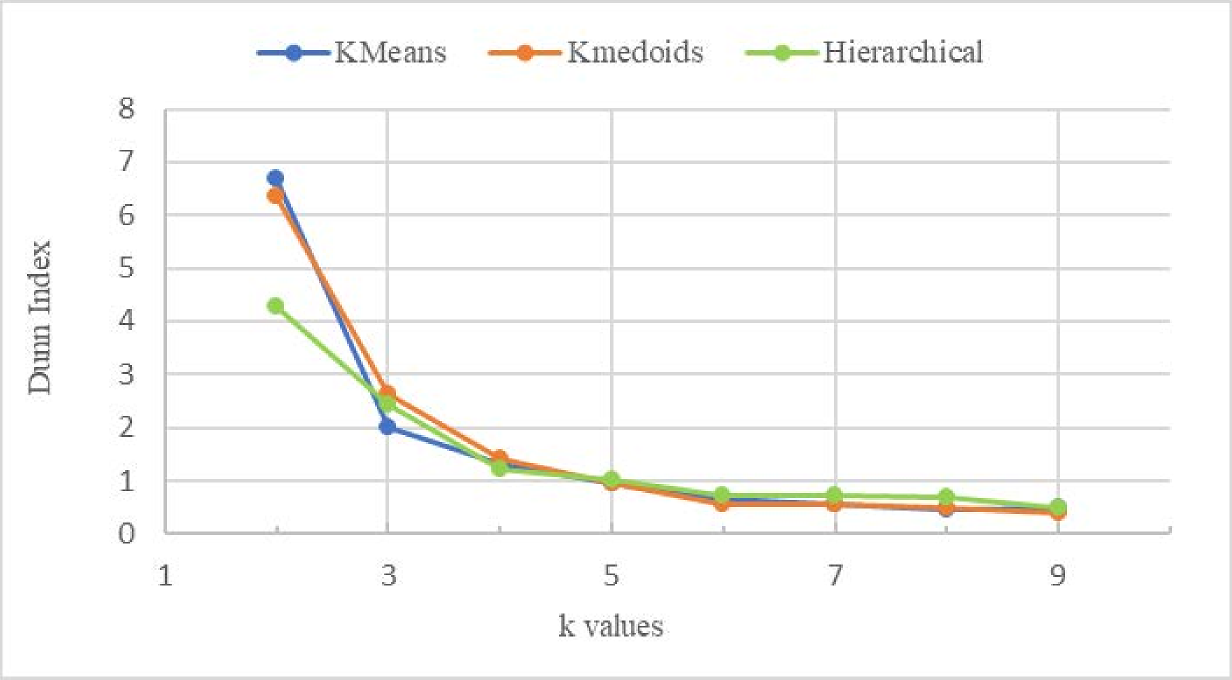

Figure 3-5. Clustering validation results with three different clustering algorithms and Dunn index as the evaluator. K-means works best when number of clusters is two (optimum), but with increasing k the three algorithms show similar performance.

Figure 3-6 shows Silhouette scores. The thickness of each cluster shows the number of data points in that cluster. The dotted vertical line indicates the average silhouette value for clustering. It is important to note that the distribution of data points is somehow uniform. Figure 3-7 shows the Centroids in each cluster. Given the high number of graphs in each cluster (277 profiles in cluster 0 and 270 profiles in cluster 1), a selection of them is depicted in Figure 3-8. It reveals two patterns of outdoor temperature. The daily mean temperature of the first cluster is 19.39, and 2.40 for the second cluster. Therefore, the first cluster used for high temperature days, and the second cluster for low temperature ones. Further analysis of data reveals that the days in first cluster come mostly from months 6 to 9, while those in low temperature cluster come from months 1, 2, and 12. The days in other months (3, 4, 5, and 11) are shared between two clusters. This demonstrates that outdoor temperature is less connected to calendar seasons, meaning that clustering based on calendar days is not accurate. The two clusters show some overlap, which means they are not thoroughly separated; therefore, the effect of weather is only reduced by clustering.

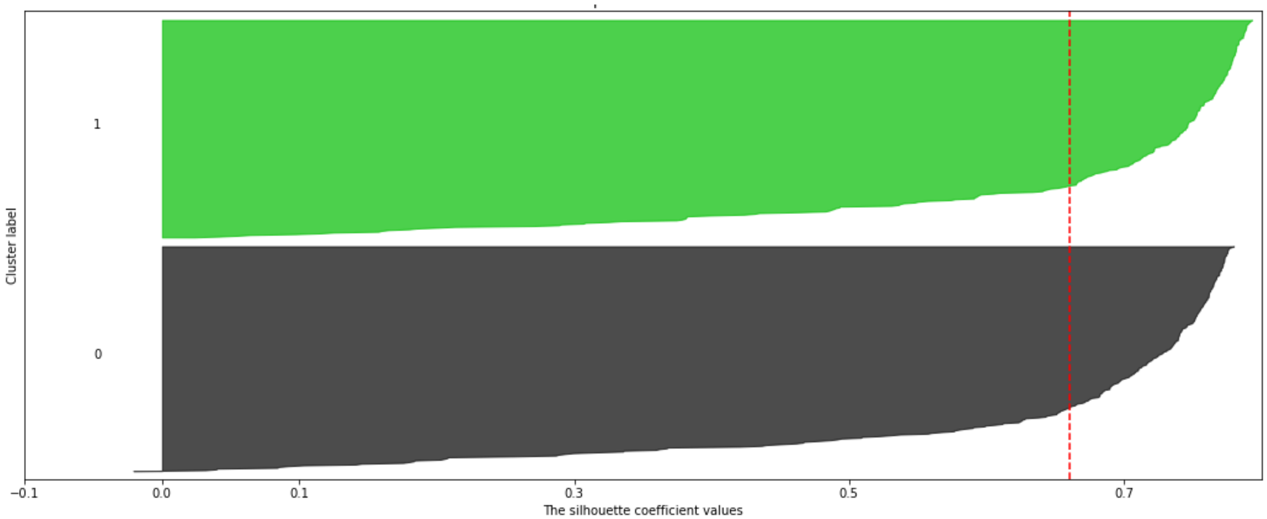

Figure 3-6. The silhouette plot for two clusters. Number of clusters is 2. The number of data points (thickness) is roughly the same in each cluster. The average silhouette value is 0.533 (the dotted red line).

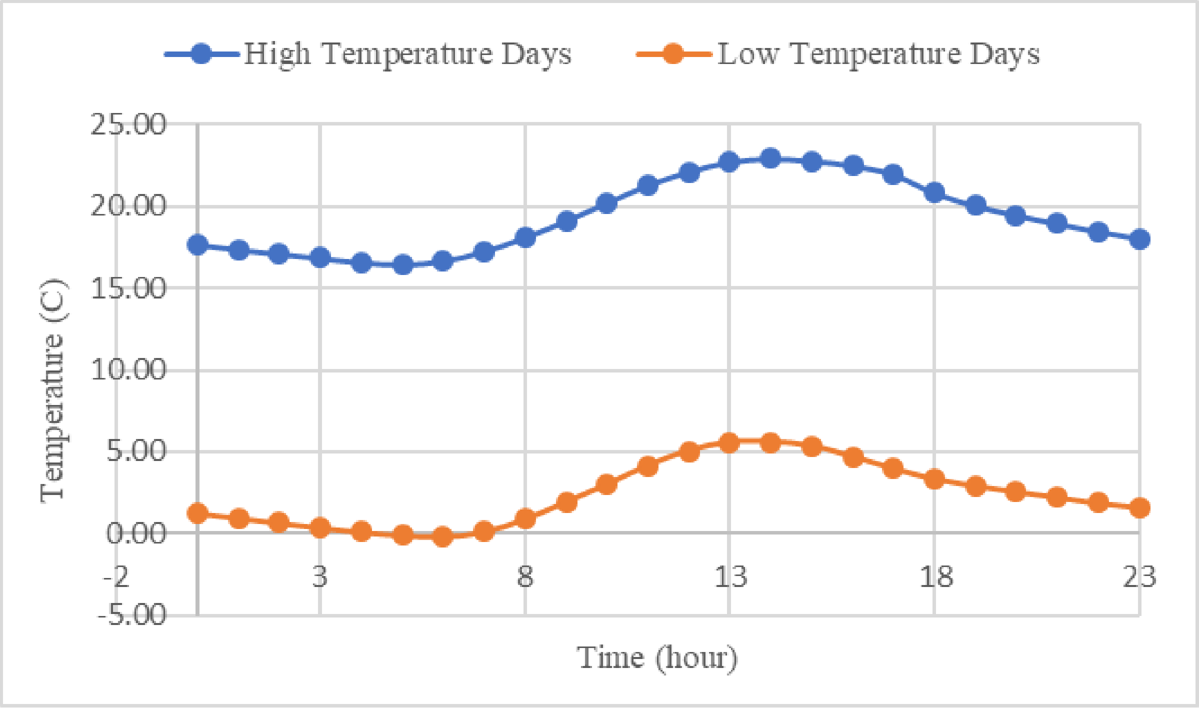

Figure 3-7. Clustering results (centroids) using K-means algorithm and k=2. Attributes are the 24 hours outdoor temperatures. The results clearly show that the clusters are externally separated (High temperature and low temperature profiles). The study of occupant behavior can be performed on each cluster separately, so the days are similar in terms of outside temperature which affects the occupant behavior.

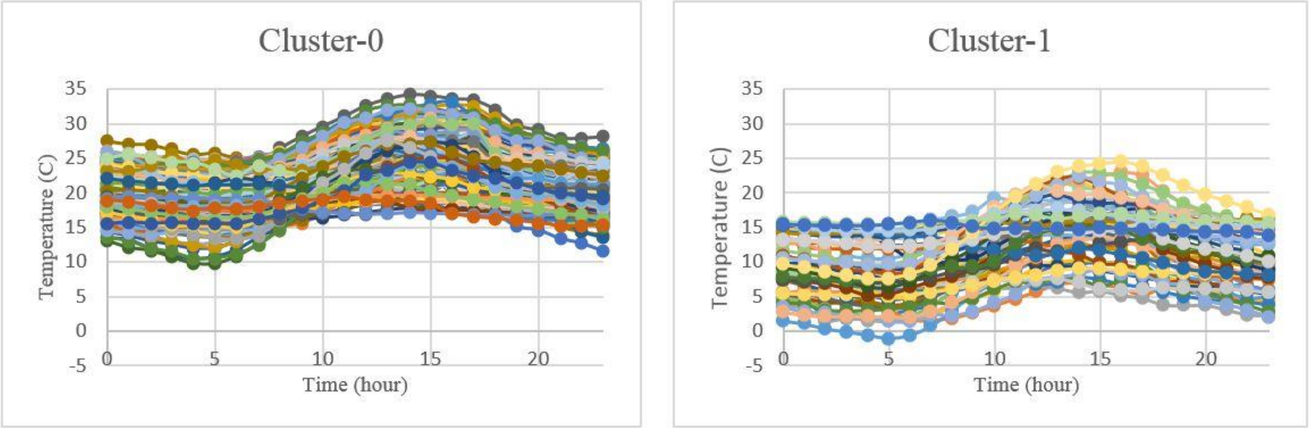

Figure 3-8. Clustering results for a sample data set using K-means algorithm and k=2. Attributes are the 24 hours outdoor temperatures. The results clearly show that the clusters are externally separated.

3.3.2 Association Rule Mining Results

Association rule mining was performed on each cluster separately to determine -correlations between various end-use loads. By inspecting and categorizing the rules, the specific energy inefficient occupant behavior was identified and energy saving recommendations was provided to occupants. After applying ARM to the dataset, 863 rules were extracted from both clusters (465 rules from cluster 0 and 398 from cluster 1). After trying various combinations, a minimum 40% support and 85% threshold were used. The minimum Lift value was set to 1.01 to obtain positive correlations. Table 3-2 shows a selection of association rules. ARM results were generally expected and reasonable meaning that logically makes sense, but many were unexpected and need to be investigated by domain knowledge. For example, according to the first rule in the table, when the refrigerator and electric water heater are off most of the day, the kitchen light is usually off too. This seems logical given the presence of occupants in kitchen when of cooking or using the refrigerator. Similar explanations are given for rule 3 and 6. However, some rules cannot be explained logically. This shows the practicality of association rule mining, which is able to find correlations hidden in the dataset that may not be discovered by other methods. Such rules are the second and the eighth rules in the table. As an example, the eighth rule states that when the living room outlets (Telephone, FAX, lights, and any plug loads connected to living room) are being used while refrigerator is not used frequently, then living room TV should be turned off most of the times. In fact, this rule is not interpretable because it reflects the specific tenants' behavior which may not be a general behavior.

Looking at extracted rules, it is observed that attributes with LOW energy consumption (0 in the rightmost integer) were more frequent than HIGH ones between the rules. This shows building occupants are highly conscious regarding energy-saving measures. In fact, this shows a good opportunity to modify any behavior opposite to these rules.

When analyzing association rules, one should be aware of their corresponding cluster to have a sense of outdoor conditions when giving recommendations. In the case of rules in cluster 0, which is considered high temperature, attributes such as 2, 3 or 7 (electric water heater and lights) should be low most of the time because they are used less frequently in warmer seasons (where days and natural light are longer) compared to colder seasons.

Inspecting the extracted rules allows finding relations between end-use loads. For example, rule number 1 shows refrigerator (low) and electric water heater (low) imply that kitchen lighting be low. Therefore, to reduce the use of kitchen light, occupants may need to use the fridge and electric water heater less frequently. Similar interpretations can be derived using rule number 3 or 5. Although there is always uncertainty involved in these rules, they are still useful to show the correlations.

Table 3-2. A selection of extracted association rules from both clusters.

Rule	Pre[i]	Con[ii]	Supp[iii]	Conf[iv]	Lft[v]	Cat[vi]	List of symbols
1	[100, 30]	70	0.500	0.865	1.157	RM	1= Kitchen - Dishwasher / Dryer [Wh]
2	[50, 71]	20	0.664	0.852	1.077	RM	2= Study Room - Outlet · Electric Light [Wh]
3	[130, 100]	70	0.463	0.856	1.144	RM	3= Electric Water Heater [Wh]
4	[80, 130, 100]	70	0.450	0.850	1.14	RM	4= Entrance, bath, toilet washroom-outlet light [Wh]
5	[20, 70]	100	0.667	0.968	1.117	RM	5= Bedroom · outlet · electric light [Wh]
6	[60, 101]	10	0.545	0.904	1.094	RM	6= kitchen - Electromagnetic cooker [Wh]
7	[10, 101, 141]	71	0.401	1.000	1.088	RS	7= Kitchen - Outlet · Electric Light [Wh]
8	[91, 100]	80	0.526	0.993	1.016	RM	8= Living Room - TV [Wh]
9	[120, 130, 71]	101	0.480	0.950	1.040	RS	9= Living Room outlet (Telephone / FAX, …) [Wh]
							10= Kitchen - refrigerator [Wh]
							11= Washing machine [Wh]
							12= Kitchen - Microwave [Wh]
							13= Kitchen - Rice Cooker [Wh]
							14= Lavatory - Hot water [Wh]
							15= Living room temperature (°C)
							16= Living room humidity (°C)
							17= Bedroom temperature (°C)

[i] Premise
[ii] Conclusion
[iii] Support (min=40%)
[iv] Confidence (min=85%)
[v] Lift (min=1)
[vi] Categorical

3.3.3 ANN and Recommendations Results

Extracted rules were categorized as RM, RS, and RN according to the description detailed in section 2.4. For each RM and RS rule, one prediction model was built and stored. That means 863 ANN models were built and evaluated for 390 RS rules and 473 RM rules. While evaluating ANNs, the network parameters were adjusted for each model; if satisfactory results were not obtained (based on R-squared and sum of mean squared error and other metrics [56]), the rule was rejected. To demonstrate this process, two rules (one from RM and one from RS) were selected and shown in Table 3-3:

Table 3-3. Two example rules for ANN model construction.

Rule No.	Premise	Conclusion	Supp	Conf	Lift	Cat
1	[80, 130, 100]	70	45%	85%	1.14	RM
2	[120, 130, 71]	101	48%	95%	1.04	RS

RM Rules

Rule 1 indicates that when the living room TV, kitchen rice cooker and refrigerator consume low energy during the day, kitchen outlet lights should be switched OFF most of the time. This rule has a confidence level of 85% and a Lift value of 1.14, indicating a strong correlation. This means that in 85% of cases when living room TV, kitchen rice cooker and refrigerator had low energy consumption, kitchen lights were mostly OFF (low consumption). This occupant behavior is frequent. Therefore, it is expected that kitchen lights be OFF when the three loads mentioned are low. This rule is categorized as RM rule and any occupant behavior following the premise (80, 130, 100) but not the conclusion (71 instead of 70) is considered inefficient. The occupant is warned about this. One reason for this may be forgetting to switch off the lights. This recommendation is practical mainly because occupants have shown such behavior frequently; the rule is recurrent and has occurred 48% of the time for the same occupants. Therefore, such rule-based recommendations are feasible and easy to follow.

The second task is to estimate potential savings upon following the recommendation achieved by building the ANN model. This task indicates possible savings and gives occupants the motivation to watch their energy consumption. All data points in the form of [80, 130, 100, 70] are extracted and used for modelling. The last attribute (70) is the model output and the remaining ones are fed to the model as the input. Figure 3-9 shows the normalized profile of inputs and outputs. All data points (i.e. days) belong to the same cluster but are not necessarily successive days. This explains the gap between the graphs. There are a lot of ups and down in the graph, but all of them are in the [0-0.5] range, which indicates low energy consumption.

To map the inputs to the output, a multi-layer feed forward neural network with back propagation was chosen as the model. GridSearchCV from the scikit-learn package was used to optimize network parameters such as solver, number of layers and hidden neurons, regularization parameter, learning rate and activation functions. All combinations were tested, and best results were used for each ANN model. The optimization was performed to get the highest cross-validation accuracy. Given the small size of extracted data sets (each extracted dataset usually have around 200 data points), a 5-fold cross validation was used to tune parameters and test accuracy. However, to analyze model accuracy with unseen data, 10% of data was separated.

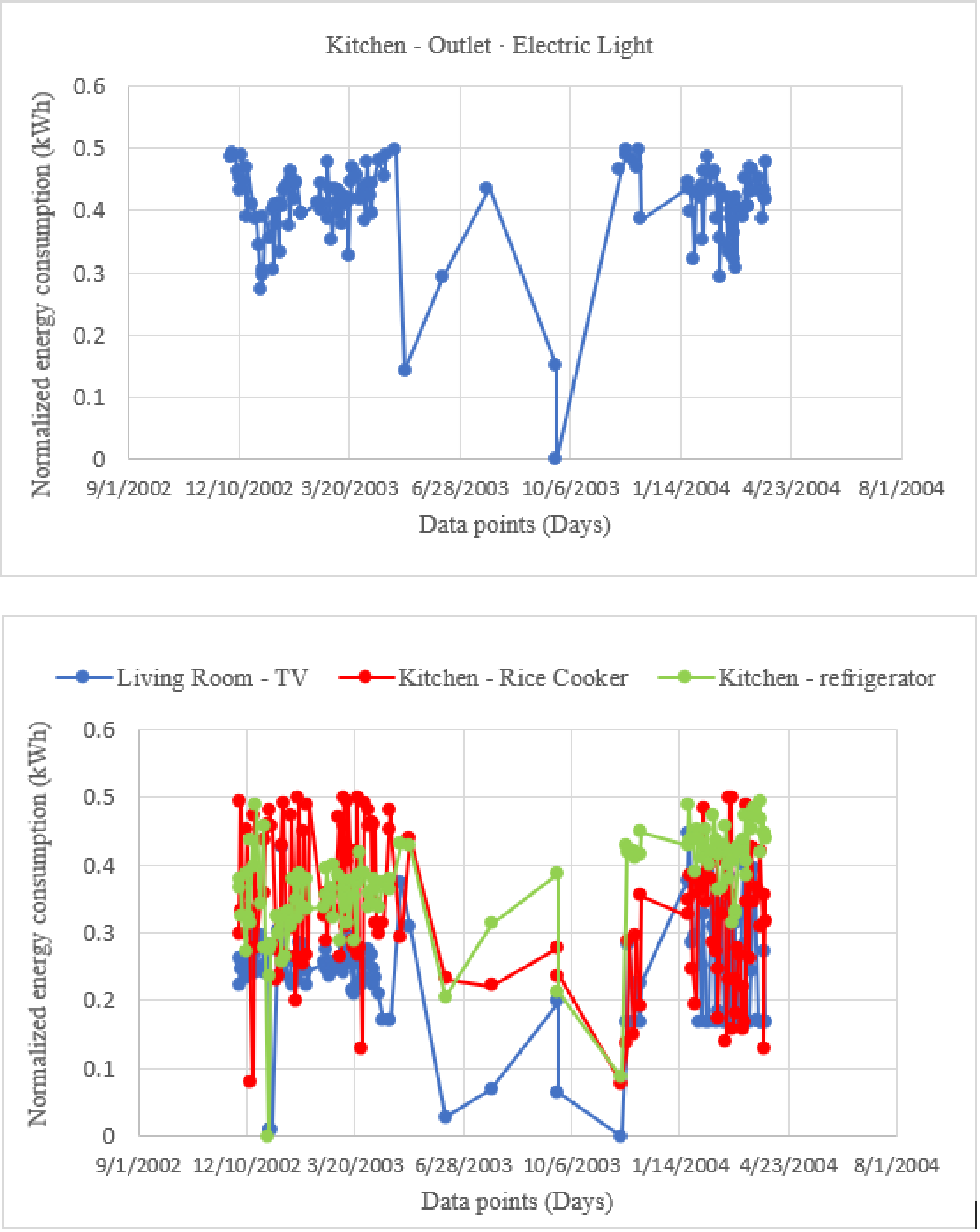

Figure 3-9. Normalized energy consumption of end-use loads based on rule 1. The upper and lower figures show the inputs and output of the ANN model. The data points are extracted from the dataset and are not successive days. Also, the days come from the same cluster.

Figure 3-10 shows the modelling result. The cross-validation error (0.0045) and test error (0.001) are acceptable. However, the R-squared value of Test set is low, which could be attributed to the small size of the dataset and noisy data. Figure 3-11 shows the real and predicted values in the training dataset. It is observable that the network is able to capture the underlying pattern in the dataset. At some points, the network is not able to follow the data pattern; the error is due mainly from the small number of datasets (around 200 data points) and inherent noise in the data.

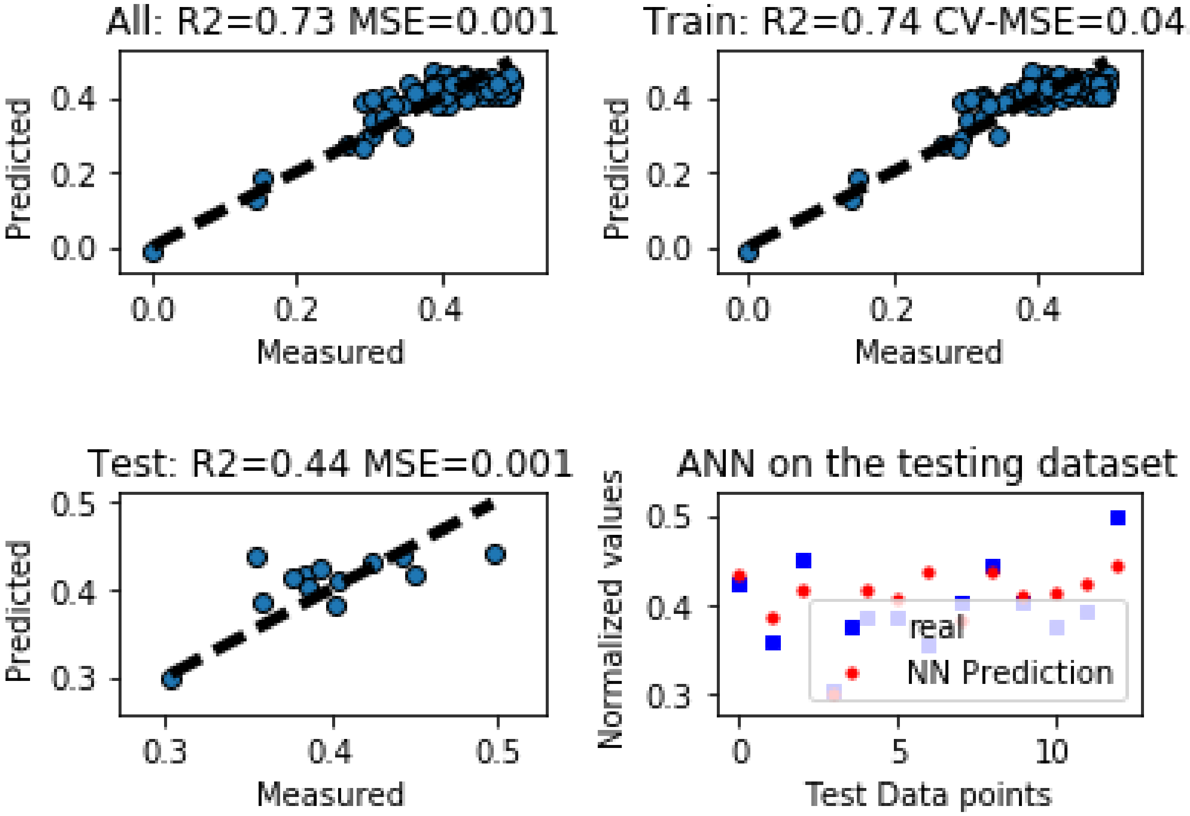

Figure 3-10. Regression fit of the real and modeled output. The y-axis are normalized values for energy consumption. It is seen that the network is able to model the energy consumption of output (kitchen lamp in this case) using other end-use loads as inputs (living room TV, rice cooker and refrigerator in this case). The cross-validation error and test error are 0.04 and 0.001 which are in acceptable range.

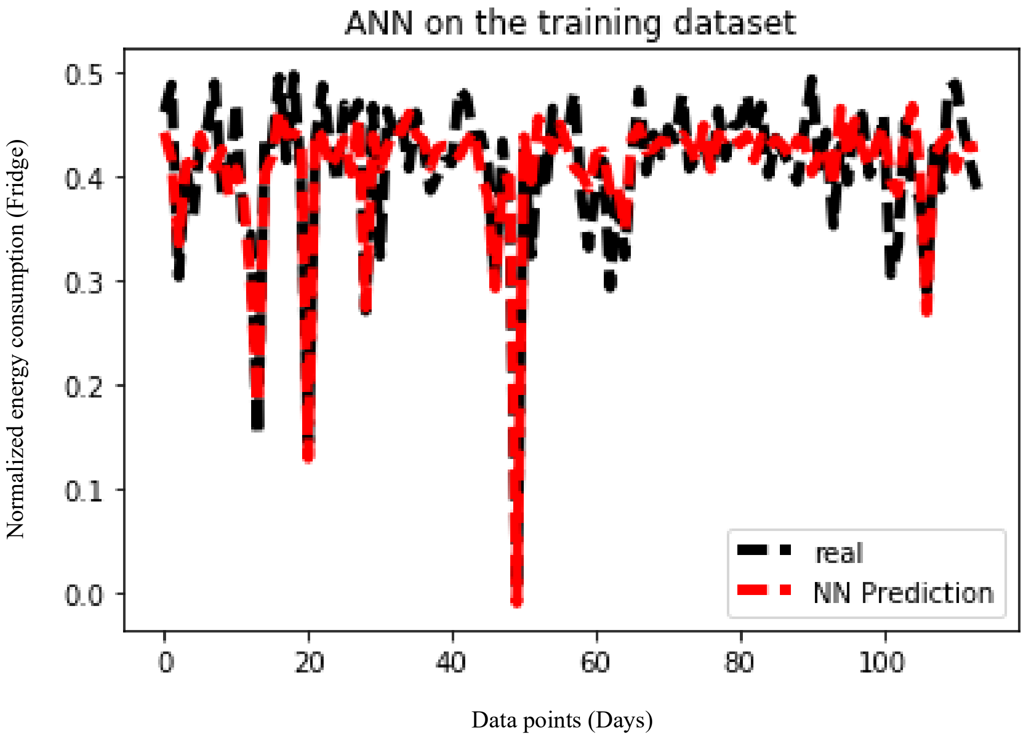

Figure 3-11. Comparison of real and network generated values. The model can follow the trend of the real dataset. Mean squared error is 0.001.

The model is used to estimate potential savings upon following given recommendations. The data samples that follow the premise ([80,130,100] occurring together), but not the conclusion is selected. The premise is fed to the model as input and the output gives us modified values. The difference between modified and real values is the potential savings (See Figure 3-12). Using equation 3.7, cumulative potential savings go up to 19%, which shows a good opportunity to save the energy if occupants follow recommendations and watch their behavior. In other words, this potential saving shows that such system is able to save a considerable amount of energy. One disadvantage of the current approach is that occupants' past actions cannot be reversed to save energy because this new data is also historical. However, if the system expects occupants' behavior online (current data is fed to the system and inspected), when abnormal behavior

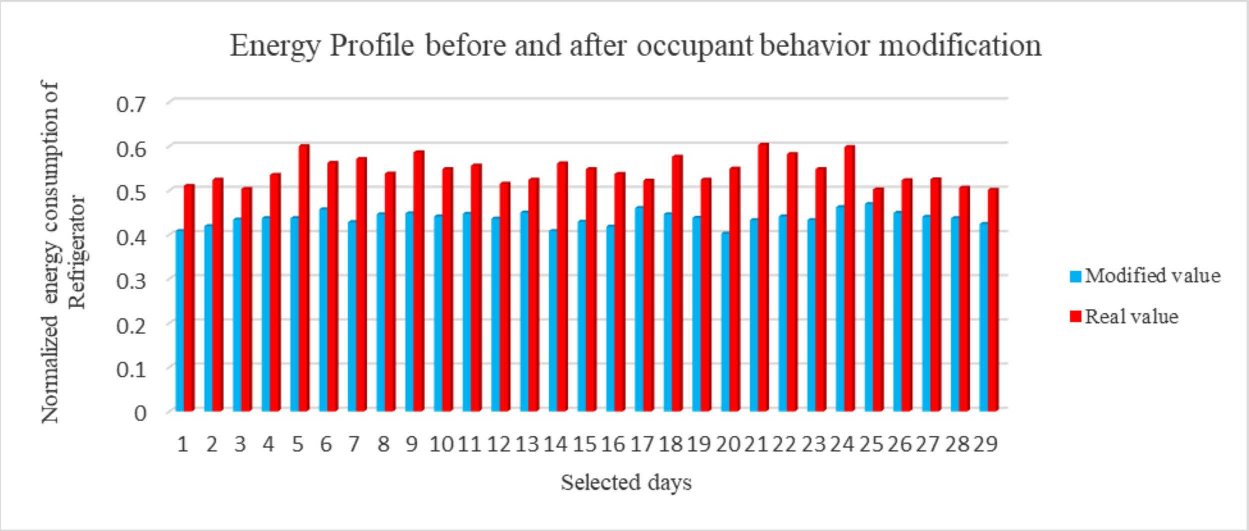

Figure 3-12. Applying the model on the new data set to estimate the potential savings upon following the recommendations. The cumulative sum of difference shows the potential savings. The cumulative savings are up to 19% which shows a good opportunity to save energy.

occurs, the system alerts the occupants right away and prevents energy loss. For example, they could forget to switch lights OFF when leaving home, leaving the TV ON while being absent, leaving the refrigerator door open, or similar actions.

RS Rules

Rule 2 in Table 3-3 implies that when the microwave and rice cooker use low energy during the day (they are OFF most of the time), while the kitchen lamps are mostly ON, then, the data suggests the refrigerator has high energy consumption as occupants frequently open and close it during the day. Such events occurred together 48% of the time and among them, 95% of the time the kitchen refrigerator consumed high energy. This rule is considered as an RS (rule for saving) because the conclusion (kitchen-refrigerator) has high energy consumption. If occupants show the same energy use pattern in the premise and an opposite one in the conclusion, this shows occupants' saving. Importantly, it is assumed that all appliances are working properly (no anomalies exist) and less energy consumption of the appliance is attributed to more energy awareness of the occupants.

To quantify savings, data in the form of [120, 130, 71,101] happening all together is extracted. This dataset is shown in Figure 3-13 and was used for ANN model construction. As the figure shows, the extracted points are days from 2003 to 2004 that respect the mentioned rule, and are not necessarily successive (there are gaps between some points). The normalized energy consumption of refrigerator is always greater than 0.5 which is expected.

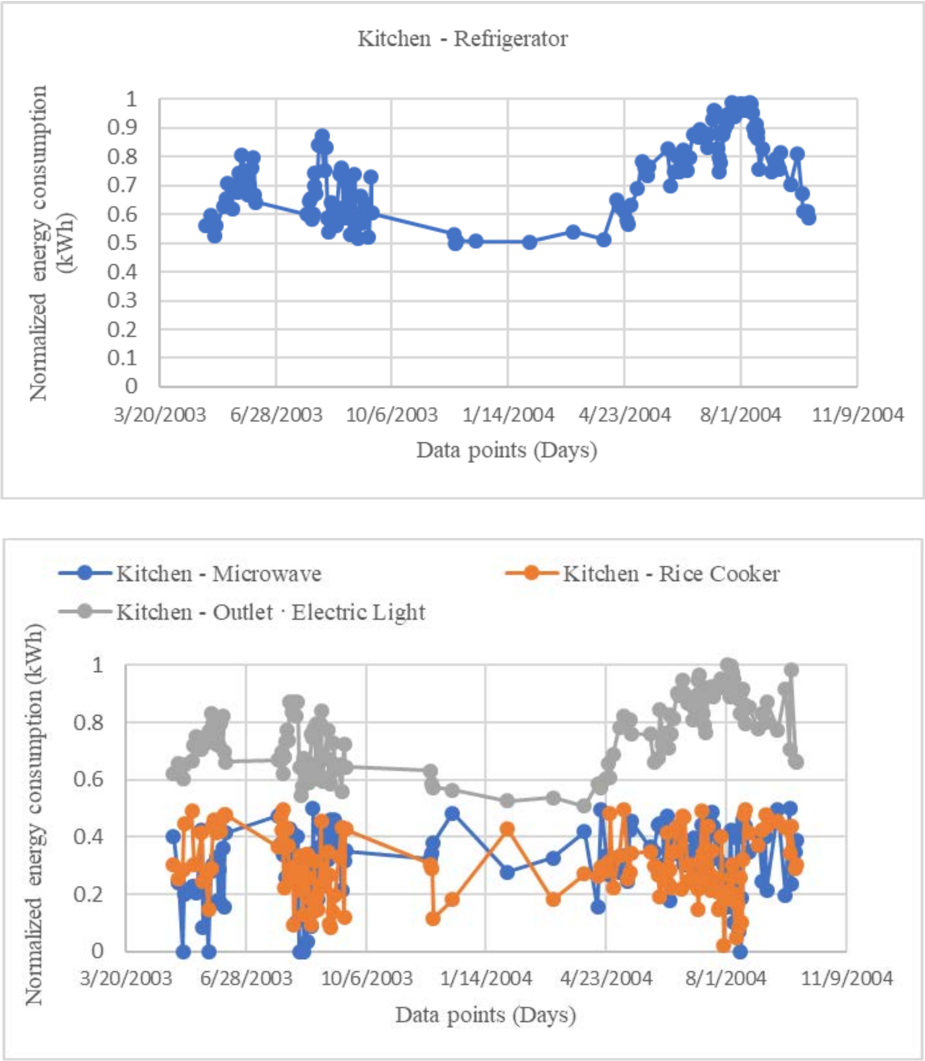

Figure 3-13. Relationships between inputs and the output based on associations. The inputs are shown in the upper figure and the output in the lower part. There is a complicated relation between various end-use loads which are modeled using a neural network model.

Figure 3-14 shows the quality of the fit. The regression fit shows a good performance on the training and testing set. The error is mainly associated to the small number of data points. More data can reduce the error generated by the model (one bad prediction can be seen in train dataset).

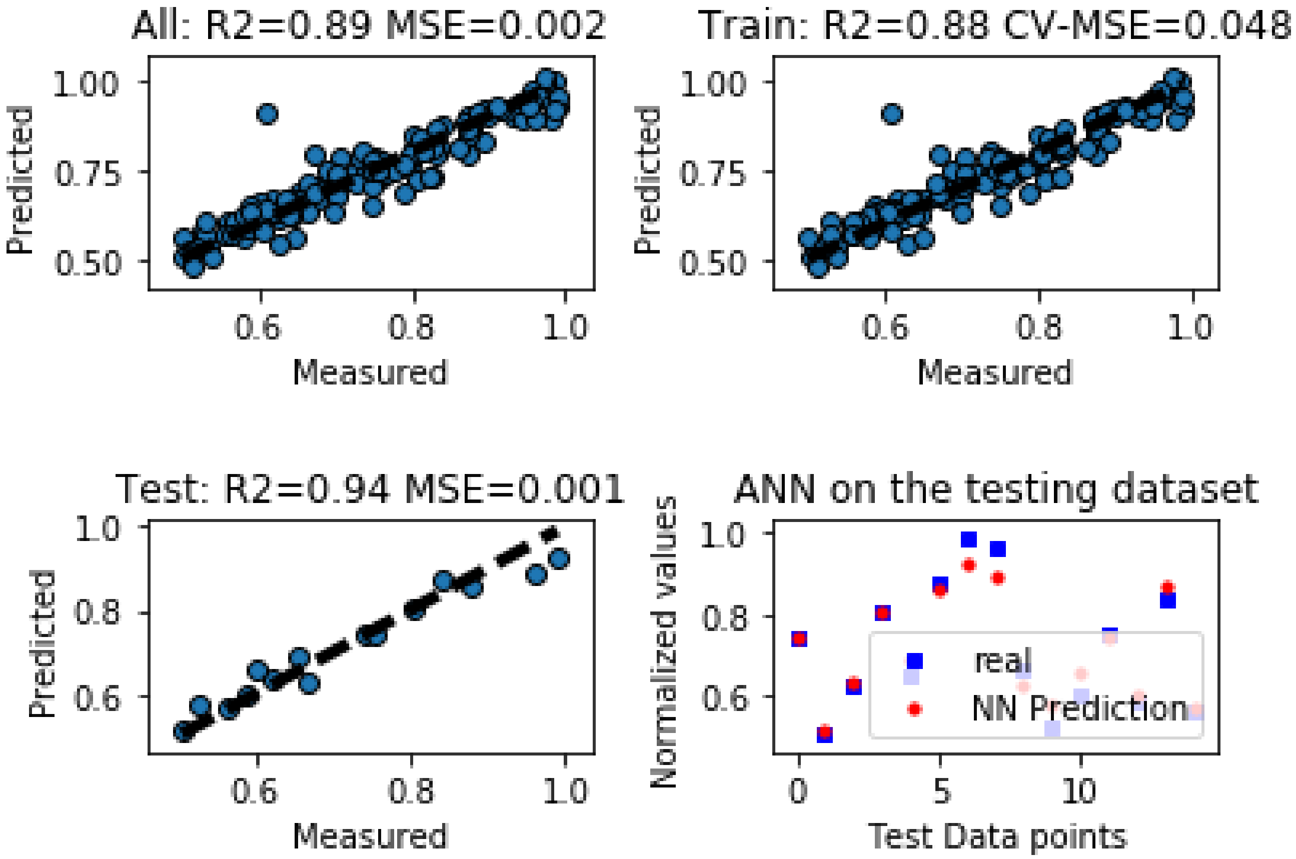

Figure 3-14. Regression fit of the real modeled output. It is seen that the network is able to model the energy consumption of refrigerator using other end-use loads microwave, kitchen lamp, and rice cooker.

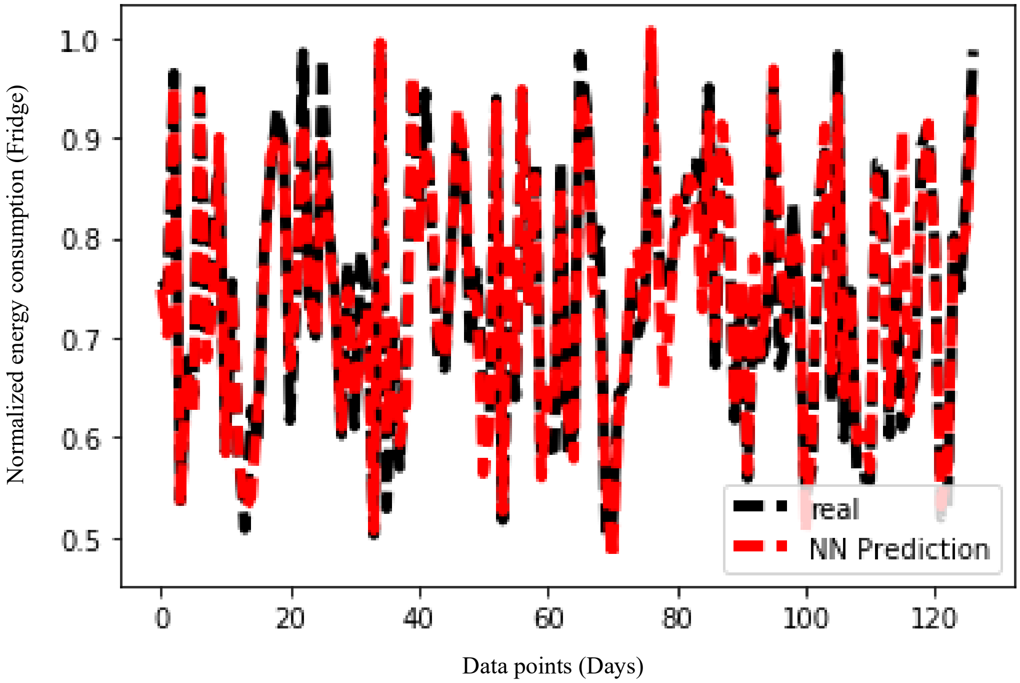

Figure 3-15. Result of training an MLP neural network on the obtained rule. The cross-validation error is 0.04.

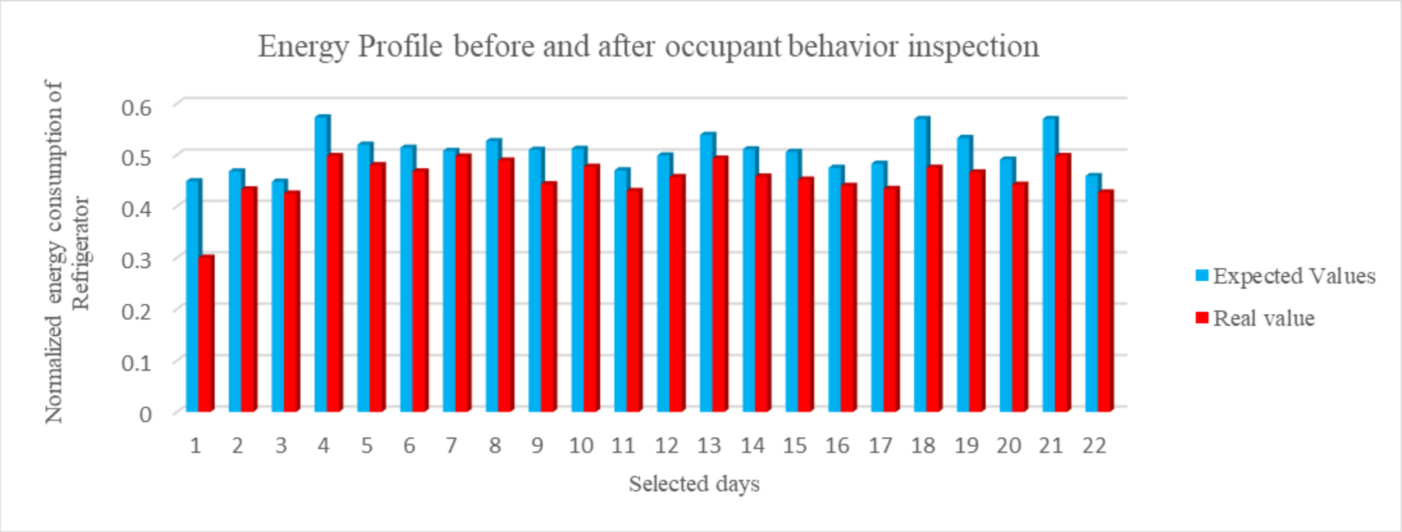

Figure 3-16. Applying the model on the new dataset to estimate the achieved savings by the occupants. The cumulative sum of difference shows the improvements. The cumulative savings are up to 10%.

Figure 3-15 shows the trained model. It is observed that the ANN model is able to follow the pattern of the data satisfactorily. Although there is some discrepancy in the real and predicted values (in the 20th to 40th range), the average error of the network is 0.001 and R2 is 88%.

Once the model accuracy was tested and accepted, the data was inspected and looked for data points showing such behavior in the premise ([120, 130, 71] occurred together) when the conclusion was not respected (100). The inputs were fed to the model and expected behavior was extracted (101). The difference between expected behavior and real behavior (100) shows the savings achieved by occupants. Figure 3-16 shows the results. The blue bars show the expected values generated using the neural networks, while the red bars show the real energy consumed by occupants. The cumulative saving for these 22 days is 10% (calculated by Eq. 7).

Using a similar approach, after evaluating each rule, 29 and 36 rules were catalogued as RM and RS rules, respectively. Table 3-4 shows the result of applying the rules on the dataset. For RM category, potential savings over 25% (for each rule) were flagged as High Recommendations (occupants should take more precautions), and the rest were categorized as Normal Recommendations (The importance of them is not as much as High Recommendations). It also shows that there is a potential saving of up to 21% in total energy consumption (calculations were performed for all rules together). Also, achieved savings were reported to be 12% for occupants for all rules. This motivates tenants to follow more energy saving measures.

A more detailed report that includes the appliances needing more attention is also available, so occupants can focus more on those. For example, considering RM rules, it is revealed that two appliances, kitchen outlet lights and the refrigerator appear in the recommendations more frequently. This shows that these two appliances deserve more attention.

Table 3-4. Result of applying all rules on the dataset. High recommendations have a potential saving of more than 25% each, while the recommendations have lower than 25% potential saving.

	Number of Rules	Recommendations priority	Average of savings per rule	Total savings
RM rules	29	High recommendations: 18 Normal recommendations: 11	High: 26% Normal: 20%	21%
RS rules	36	_	11%	12%

3.4 Conclusions

Although different services in a building may work efficiently, occupants may not be informed enough to fully exploit their opportunities to save energy. This study proposed a new approach for evaluating the energy consumption of different end-use loads and used it to create a recommendation system that would advise tenants how to decrease their consumption. Different data mining tasks were employed in a framework—clustering, association rule mining, and artificial neural networks. The idea was to find frequent patterns in the data and use them as models to inspect occupants' behavior. If an opposite behavior was noticed in the energy consumption pattern for any appliance, tenants were notified and potential or achieved savings reported. According to the rules obtained in the data, a potential saving of 21% is achievable. This demonstrates that there remains a lot of potential for occupant behavior to improve. This highlights the importance of data-based systems to unleash the hidden potential of data for energy savings. The methodology also enables the building management system to report the savings achieved by occupants as a motivation for them to act more consciously. In this case study, the achieved savings were 12%. Further investigation of rules obtained shows the important end-use loads that need more consideration, such as refrigerator and the kitchen lamps in this study. The next step would be developing methodologies involving more than one single building for comparison purposes to overcome the mentioned challenges.

4. Development of a Ranking Procedure for Energy Performance Evaluation of Buildings based on Occupant Behavior

4.1 Introduction

According to reports published by the Natural Resources Canada [3], residential and commercial buildings are a main contributor to total secondary energy use, making up more than 30% of the total. This shows the necessity of energy consumption manipulation in buildings for a sustainable future. Occupants could affect the energy consumption of a building to great extents even if all systems and equipment (end-use loads and heating, ventilation, and air conditioning [HVAC] systems) work perfectly [49]. Recently, there have been many improvements in technological solutions, such as design and operation of building services [33]. Among these, recent research highlights occupant behavior as an important contributor that can increase the energy efficiency of buildings, similar to technological solutions [57].

Generally, the factors influencing the building energy consumption could be divided into four main categories (see Table 4-1).

Table 4-1. Influencing factors in energy consumption

1	Climate (e.g., outdoor temperature, solar radiation)
2	Building-related characteristics (e.g., type, area, heat loss coefficients)
3	Building services (e.g., space heating and cooling, hot water supply)

4	Building occupant activities and behavior (e.g., user presence, activities)

Among these, building occupants' activities and behavior include factors that indirectly affect energy consumption. For example, social and economic factors (energy cost, degree of education, etc.) partly affect the occupants' attitudes toward energy consumption [58]. Indoor environmental conditions are also determined by the occupants; therefore, they are an indicator of occupant behavior. The combined effect of the first three factors on energy consumption is identified using advanced simulation packages that are robust with respect to simulating different scenarios. However, modeling occupant behavior is still a challenge due to its complexity and indirect effects. Additionally, various statistical data analysis processes have been applied to establish meaningful relationships among energy consumption and influencing factors that help reduce energy consumption effectively. However, with the increasing amount of data generated by buildings within the complexity of the systems, especially on occupant behavior, these relationships cannot be captured by simple statistical methods; thus, such methods are inadequate for performance improvement. In this study, the challenge is: "How can we develop a procedure to assess the performance of a group of buildings based on the occupants' behavior?"

The occupants of a single building may not be well informed regarding their energy consumption performance. One solution to this challenge would be developing a procedure for a comparison among occupants of several buildings to show the rank of each building among others and showcase occupants' potential abilities to reduce energy consumption on specific end-use loads. This way, the residents of each building would know their real rank and could be motivated to take energy-saving measures. In other words, the occupants can learn from each other and be persuaded to reduce their consumption because they would observe that a similar building is consuming less energy. For example, if occupants of a single building see that the occupants of similar buildings are using less energy to provide the required indoor environment, they might be persuaded to follow them. However, due to the existence of several factors in energy consumption patterns of occupants (such as number of occupants, floor area, and house type), we cannot simply compare the energy consumption of several buildings. Also, the activities that occupants have performed to save energy and the degree of their energy consciousness are not accounted for. Thus, such simple comparisons are not yet effective. As many influencing factors as possible should be considered to make a fair comparison. Also, if the procedure is well designed, it can reveal potential saving opportunities for occupants of the buildings. In this study, a new framework for energy assessment of a set of buildings is introduced using data-mining techniques. The details are introduced in the following sections.

4.2 Methodology

The proposed method is composed of multiple steps, shown in Figure 4-1. The data from 80 buildings in Japan are collected, integrated, and processed. Outlier detection and removal are applied on the dataset as preprocessing steps. As a result, four buildings are removed due to data deficiency. Outliers are substituted by approximations using regression on other attributes [59]. Each step uses a subset of available features in the data set, which are described in the following sections. The proposed method is composed of two-level ranking, as shown in Figure 4-1 in red boxes. The first level considers the amount of energy usage by occupants after filtering out effects unrelated to occupants. The second level ranks the buildings in terms of achieved and potential savings during the time under investigation. Therefore, the occupants know their specific category in terms of each level. This helps them clearly understand their performance and take suitable measures. The methodology was applied on detailed data of 76 buildings.

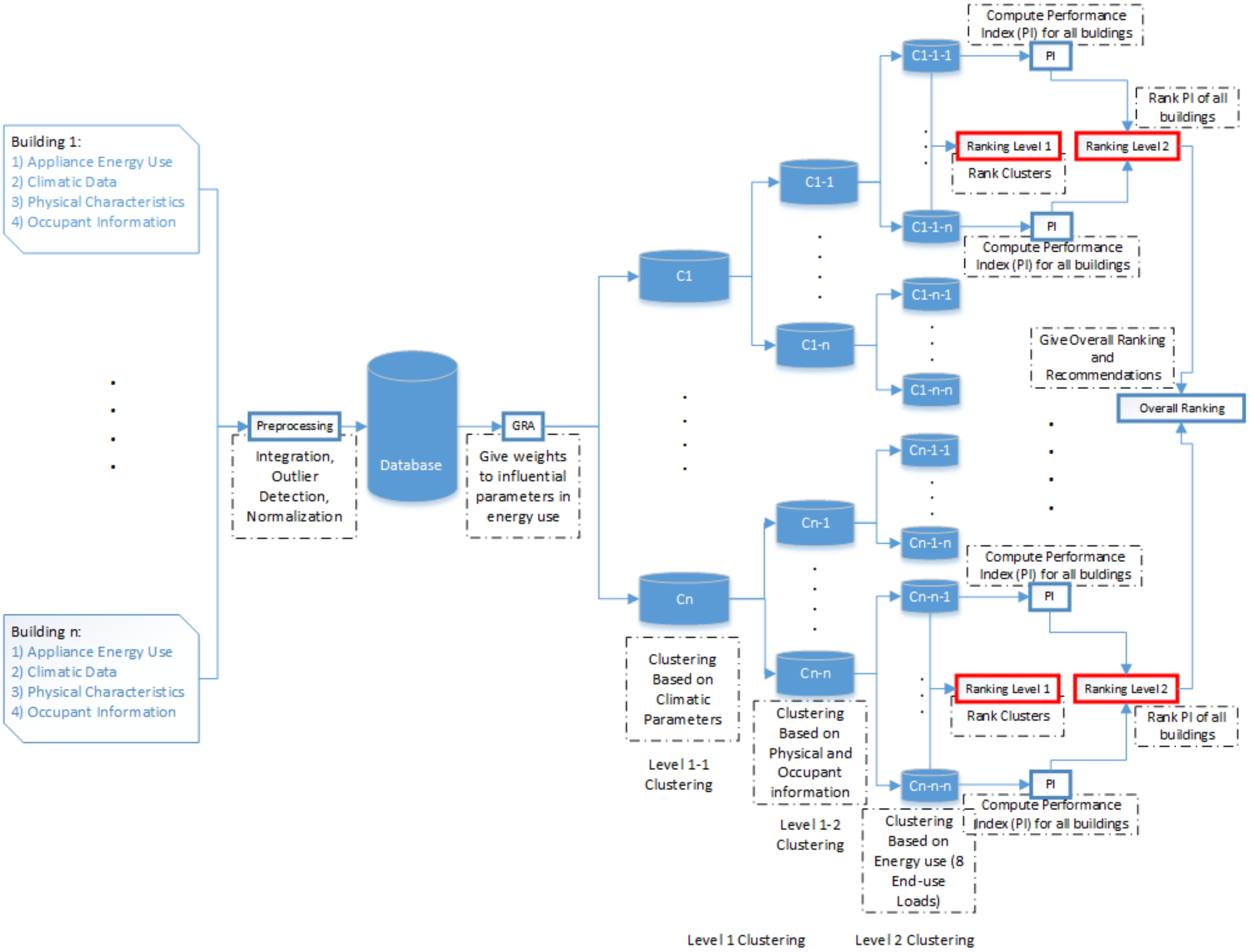

Figure 4-1. Framework development for building performance comparison based on occupant behavior.

4.2.1 Grey Relational Analysis

The influencing factors that affect building energy consumption are listed in Table 4-2. The contribution of each of these factors in overall energy consumption of buildings (energy use intensity, EUI) differs greatly. For example, variation in number of occupants may have larger effects on the total energy consumption than variation in the annual wind speed. Therefore, weights should be assigned to each of these factors. One of the methods that could give such weights is grey relational analysis (GRA), which is applied on the data set as shown in Figure 4-1. GRA tries to identify the causative factors of a defined objective (energy use, in this case) and sort them in terms of their contribution [60]. Considering an n × m data set, y_j denotes the objective sequence (in this case, building energy consumption), and y_i denotes the influencing factors (listed in Table 4-2). Therefore, i: 1,2, …m, and $j: 1,2, \dots P$ (in this case, P = 1), and $k: 1,2, \dots n$ is the index of data point.

$$y_i(k) = \frac{x_i(k)}{\frac{1}{n}\sum_{k=1}^{n} x_i(k)} \tag{4.1}$$

Similarly, $y(k)$ is defined for all objectives. At any data point k, the grey relational grade between $y(k)$ and $y_i(k)$ is defined as:

$$\xi_j(k) = \frac{\min_i \min_k |y(k) - y_i(k)| + \alpha \max_i \max_k |y(k) - y_i(k)|}{|y(k) - y_i(k)| + \alpha \max_i \max_k |y(k) - y_i(k)|} \tag{4.2}$$

where α is the "distinguishing coefficient" and is generally set to 0.5 [59]. The results are calculated for each data point. The average is used as the weight for the corresponding attribute and is known as the relational grade.

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \xi_i(k) \tag{4.3}$$

The relational grades are numerical measures which show the effect of the influencing factors on the objectives. Basically, $r_i > 0.9$ denotes a marked influence, $r_i > 0.8$ shows a relatively important influence, and $r_i > 0.7$ an important one; also, $r_i < 0.6$ denotes a negligible influence [11], [60].

4.2.2 Level 1 Clustering

Clustering analysis tries to group a set of observations by maximizing between-cluster distance and minimizing within-cluster similarities. In other words, it tries to put observations into distinct groups. The buildings are clustered based on influencing factors unrelated to occupant behavior (mentioned in Table 4-2), so buildings in the same group have similar characteristics except for occupant behavior. In this study, clustering is performed several times for different tasks. There are several clustering algorithms, each made for different purposes. They are usually applied on two-dimensional data where each row represents an observation and each column represents an attribute. Clustering mainly involves five main tasks. The first is feature generation, which is the process of choosing appropriate attributes for clustering. This is based on domain knowledge and available data. The attributes chosen for level 1 clustering are described in Table 4-2. Therefore, buildings in the same cluster share similar characteristics in terms of weather conditions, building structure, number of occupants, and building services.

Table 4-2. Representative attributes of the four influencing factors on occupant behavior.

Influencing Factor in EUI	Attribute	Category-Unit	Abbreviation
City climate	Annual mean air temperature	Numerical-°C	T
	Annual mean relative humidity	Numerical	RH
	Annual mean wind speed	Numerical-m/s	WS
	Annual mean global solar radiation	Numerical-MJ/m^2	RA
Building-related characteristics	House type[i]	Categorical	HT
	Building area	Numerical-m^2	BA
	Equivalent leakage area[ii]	Numerical-cm^2/m^2	ELA

	Heat loss coefficient[iii]	Numerical-W/m^{2}K	HLC
Occupant-related characteristics	Number of occupants	Numerical	NO
Building services system and operation[iv]	Space heating and cooling	Categorical	HC
	Hot water supply	Categorical	HWS
	Kitchen equipment	Categorical	KE

[i] The houses are either detached or apartments and are transformed to [0, 1].

[ii] Measured by fan pressurization method.

[iii] Calculated based on building design plans.

[iv] Either electric or nonelectric. They are transformed to [0, 1]. As all space cooling equipment is electric, the value of HC is determined by space heating.

The second task is choosing proximity measures that differ based on the algorithms used. The most widely used measure is the K-means algorithm, given as:

$$\sum_{i=1}^{n} \min_{\mu_j} \|x_i - \mu_j\| \tag{4.4}$$

where x_i is the ith observation and μ_j is the cluster center. Other similar algorithms are the K-medoids and Manhattan distance, Pearson correlation, and cosine similarity algorithms [61]. The third and fourth tasks are applying the algorithm and explaining the results. The last task is to measure the goodness of the clustering, which is done either by external methods (mutual information, F-measure, purity, etc.) or internal measures (such as the Silhouette index or Dunn index). A list of methods can be found in [62].

Prior to clustering the data, some preprocessing is needed to make the data consistent, such as unit conversions, outlier diagnosis, and normalization. For binary attributes, their two states, such as house types, that is, [detached house, apartment], are transformed to [0, 1]. Outlier detection is performed using the quantile method, and the outliers were substituted using regression on other attributes [31], [51].

If all features are used at the same time for clustering, we may end up with some buildings in the same cluster with different climatic conditions, such as outdoor ambient temperature (other features may be very similar, which would put two buildings in the same cluster). However, comparing two buildings with different climatic conditions does not make sense. To make sure that the buildings in the same group are as similar as possible in terms of weather conditions, first, the buildings are clustered in terms of climatic data (temperature, humidity, wind speed, and solar radiation) and then grouped based on other characteristics described in Table 4-2 (level 1-1 and level 1-2 clustering).

4.2.3 Level 2 Clustering

Given that all buildings in the same cluster level 1 share similar characteristics in weather conditions (level 1-1) and building and occupant characteristics (level 1-2), the differences in energy consumption of the buildings of the same cluster (level 1) are due to occupant behavior. The buildings are again clustered in terms of energy use intensities (EUIs), which is an indicator of the occupants' behavior (indicated in Figure 4-1 as level 2 clustering). The detailed attributes are summed up in eight categories:

1) Heating, ventilation, and air conditioning (HVAC)
2) Hot water supply (HWS)
3) Lighting (LIGHT)
4) Kitchen (KITCH)
5) Refrigerator (FRIDGE)

6) Entertainment and information (E&I)

7) Housework and sanitary (H&S)

8) Other end-use loads (OTHER)

This clustering groups the buildings into different energy use levels. The number of clusters is determined based on internal measures, such as the Silhouette index [62]. The clusters are then ranked from highest to lowest EUI. Therefore, the general category of each group becomes known. Also, the main contributors to such differences will be determined from each of the eight attributes described above, which gives the occupants information about how to reduce their overall energy consumption to enter the lower EUI cluster. This constitutes the level 1 ranking.

4.2.4 Performance Index

To evaluate the activities of occupants and whether they have tried to take energy-saving measures, the performance index (PI) is calculated as described below. It is indicated as PI in Figure 4-1 and is applied on each building.

PI is defined as:

$$\mathrm{PI} = \mathrm{AS} - \mathrm{PS} \tag{4.5}$$

AS is the achieved savings and PS is the potential savings. Achieved savings mean that the occupants have tried to lower their energy consumption by taking certain actions to reduce the energy usage of one or more of the eight end-use loads. The potential savings are the amount of energy that could have been saved if the occupants do not increase their energy consumption (opposite to their previous actions which was lower energy usage of a specific end-use load). The process is designed to capture any abnormal behavior seen in the recent data based on analysis of historical data as indicated in Figure 4-2. In this process, the eight end-use loads described above are broken down into more detailed data. For example, KITCH data are broken down to washing machine, dryer, rice cooker, oven, and so on, depending on the available home appliances [63]. Data are then clustered based on outdoor hourly temperature so the energy consumption data share similar weather conditions. Association rule mining finds all the patterns in the data, which forms the basis for alerting occupants when dissimilar behavior to those patterns is seen. The rules are categorized as rules for modification (RMs) and rules for savings (RSs), as shown in Figure 4-3. RMs imply that the energy consumption of an appliance is low. Any behavior opposite from these rules is flagged as potential savings (PS), meaning that there is a potential to save energy by following the recommendation (the RM). In other words, occupants have shown a good behavior regarding an end-use load and any behavior opposite to that is flagged as waste of energy. For example, consider energy consumption of lighting in a room. After analyzing the energy consumptions, the system may extract a rule that at certain times, the light should always be (or most of the times) switched OFF. Any behavior contradicts with this rule (light be switched ON in the mentioned times) is flagged as inefficient and needs consideration and modification. Any waste of energy is considered as potential savings for the occupants. RSs imply that the energy consumption of an appliance is high. Any behavior opposite to these patterns is flagged as achieved savings (AS), which shows that occupants have used less energy than their normal usage. Artificial neural networks are used to quantify the energy-saving potential and achieved savings by occupants. The flowchart of the process is shown in Figure 4-2. More details are provided in [51]. Based on this definition, if a building has low potential for improvement and high savings achievements, it is considered a very sgood building regarding its occupants' energy awareness. Buildings in the same cluster are compared and ranked based on PI. This comparison gives the occupants of a building an idea of their place regarding their efforts to save energy and motivates them to improve their performance using the clues in level 2 clustering. This makes up the basis of a level 2 ranking. More insights are given in the Results and Discussion section. The described process is independent of level 1 ranking. Therefore, the occupants of a single building are informed about their ranks in both levels and can take suitable actions.

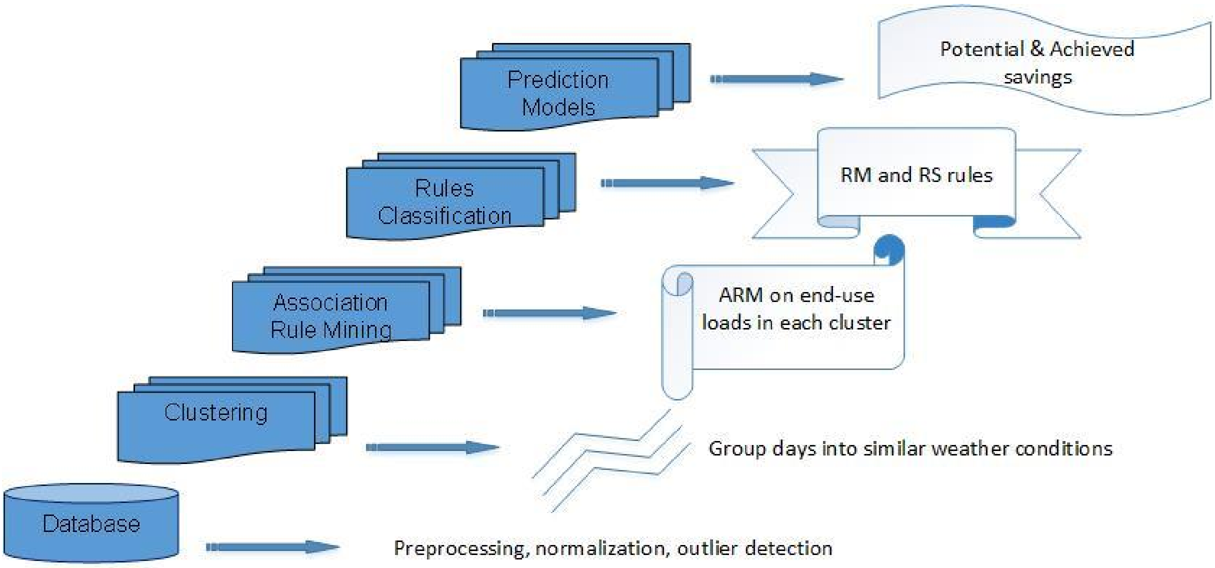
Prediction Models
Potential & Achieved savings
Rules Classification
RM and RS rules
Association Rule Mining
ARM on end-use loads in each cluster
Clustering
Group days into similar weather conditions
Database
Preprocessing, normalization, outlier detection

Figure 4-2. The data-mining process for rule extraction among home appliances to find potential and achieved savings, along with performance index.

[A,B,C,...] → X[low]   RM Rule

[D,E,F,...] → Y[High]   RS Rule

Figure 4-3. Rule categorization process [51].

4.3 Results and Discussions

4.3.1 Grey Relational Analysis

Accumulated annual energy use intensity of buildings in 2003 was selected as the objective variable in Grey Relational Analysis (GRA). Because the EUI already contains information about building area, this factor is not considered in GRA. Results are shown in Table 4-3, in which temperature, relative humidity, wind speed, and solar radiation are functions of time and region and were therefore averaged over 12 months for each district. The rest of the variables are fixed and were calculated using the whole data set. `

Table 4-3. Grey relational analysis of influencing factors on energy consumption.

egion Factors	Hokkaido	Tohoku	Hokuriku	Kanto	Kansai	Kyusyu
T	0.799	0.831	0.772	0.737	0.712	0.654
RH	0.620	0.765	0.644	0.732	0.695	0.661
RA	0.683	0.662	0.716	0.641	0.690	0.675
WS	0.584	0.555	0.532	0.601	0.580	0.605
HT	0.617					
ELA	0.490					
HLC	0.780					
NO	0.701					
HC	0.537					
HWS	0.514					
KE	0.551					

The results imply that outdoor air temperature has the greatest contribution to EUI considering only the weather parameters (except for the Kyusyu region, in which RA has the highest contribution). This is more obvious in colder climates such as Hokkaido and Hokuriku. It appears that in warmer climates, the contributions of weather parameters are similar to each other, for example in Kyushu regions all parameters are in the range of 0.600-0.675, while in Hokkaido they are in the range of 0.580-0.800. Among the other seven variables, heat loss coefficient (HLC) and number of occupants (NO) play the dominant roles.

The achieved GRAs for all variables are multiplied by their corresponding variables in the buildings data set so the more influential variables are dominant in clustering the buildings.

4.3.2 Clustering based on weather parameters (level 1-1)

As described in Section 4.2.2, level 1-1 clustering puts all buildings with similar weather conditions in the same group. This way, buildings in the same group share similar characteristics in terms of four weather parameters: outside temperature, relative humidity, solar radiation, and wind speed. The results imply two clusters, which are shown in Figure 4-4. The figure on top shows that one of the clusters contains buildings from only two regions, Tohoku and Hokkaido, which are considered cold regions with less radiation. The figure at the bottom shows the centroids of each cluster. It is seen that lower temperature and wind speed are the dominant factors that put these buildings in cluster C2, while other clusters contain buildings with higher temperature, humidity, wind speed, and solar radiation. The next step divides each of these two clusters further to consider building characteristics, too. The reason the buildings are divided first by weather conditions is the importance of weather parameters in inspecting occupant behavior.

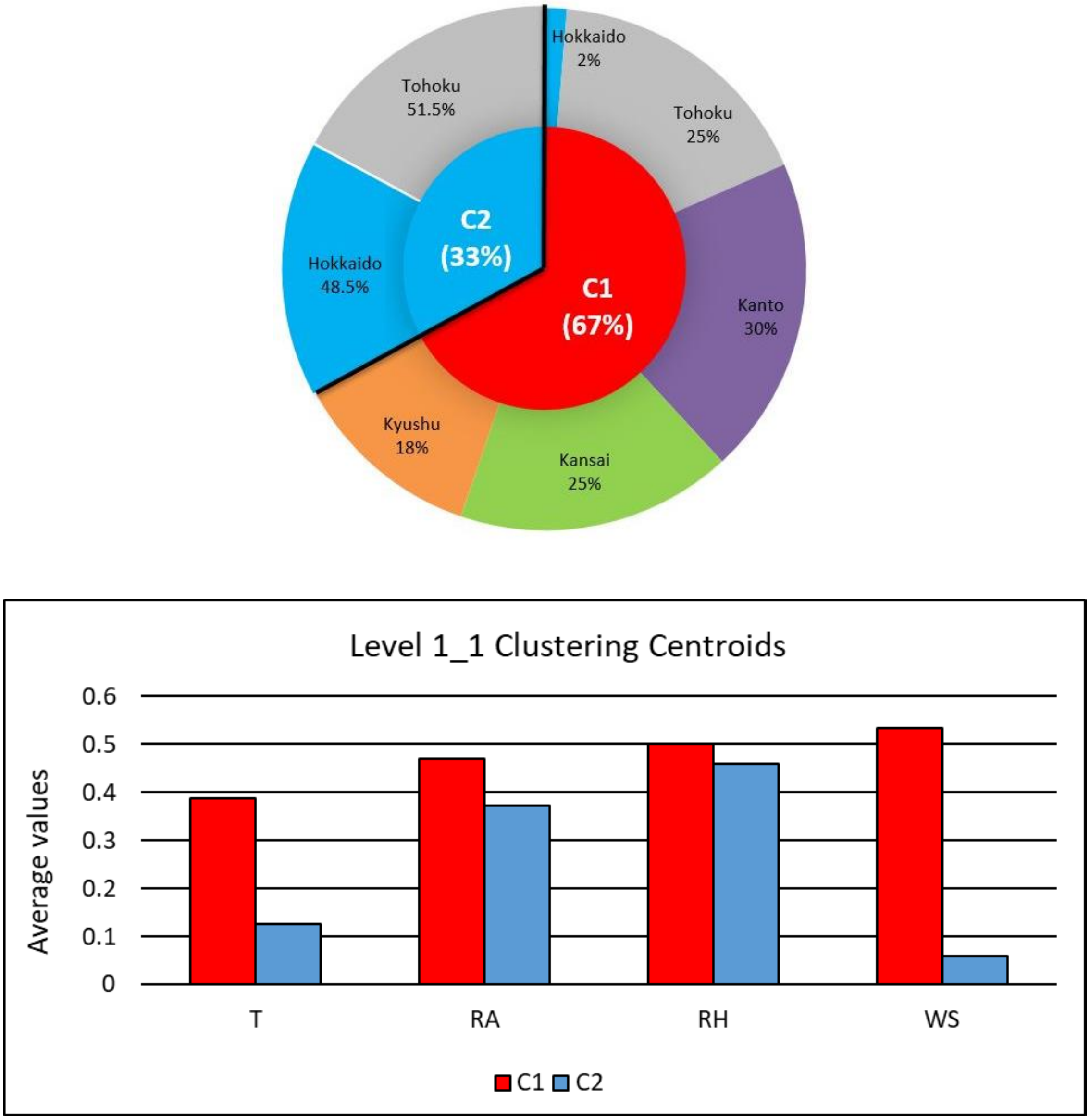

Figure 4-4. Clustering level 1-1 results on statistics and percentages of instances assigned to each cluster.

4.3.3 Clustering based on non-weather parameters (level 1-2)

Level 1-2 clustering was performed on the results obtained from level 1-1, and the cluster centroids are shown in Figure 4-5. Five clusters were obtained.

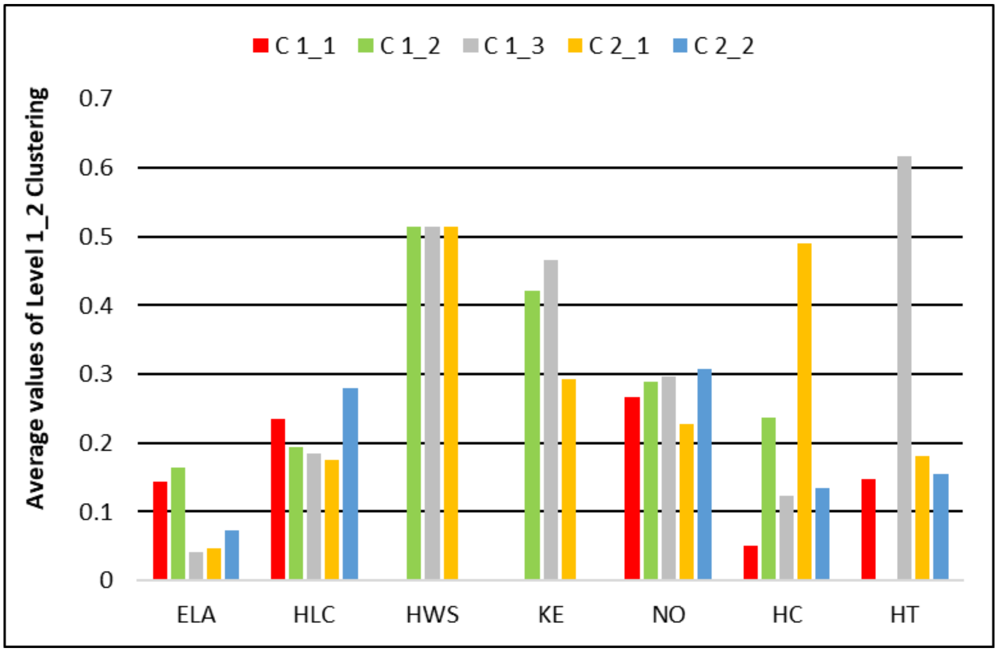

Figure 4-5. Distribution of seven physical and occupant characteristics in level 1-2 clustering.

It is revealed that buildings in clusters C 1_1 and C 2_2 (red and blue bars) share an electric source for kitchen appliance (KE) and hot water supply (HWS), while the opposite behavior is seen in clusters C 1_2, C 1_3, and C 2_1. The high HT value of cluster C 1_3 implies that all buildings in this cluster are apartments, as opposed to cluster C 1_2, in which all buildings are detached houses (HT = 0 for this cluster). Clusters C 2_1 and C 2_2 have lower outdoor air temperature, solar radiation, relative humidity, and wind speed compared to clusters C 1_1, C 1_2, and C 1_3 (based on level 1-1 clustering). From the highest HC of cluster C 2_1, it can be inferred that in cold regions, building owners use a gas-based heating system. It is seen that two variables, HT (house type) and HC (heating/cooling equipment), are dominant in separating the clusters because their values have a high variation.

There may be some overlaps between the clusters, and it is quite possible that buildings in the same cluster are grouped together by the K-means algorithm simply because they have similar characteristics on some non–occupant-related features. However, those dissimilar attributes have opposite effects (they neutralize their effect), which causes the algorithm to put the buildings together in one cluster; otherwise, the buildings are not grouped with each other.

Figure 4-6 shows the distribution of end-use loads in each cluster, along with their corresponding proportions. The difference in EUI between buildings in the same cluster is attributed to differences in occupant behavior. The eight end-use loads of each building were averaged over a year. As shown in Figure 4-6, KITCH and HWS are the two important contributors in cluster C 2_2, so the buildings in this cluster need to focus more on these two end-use loads regarding energy saving. However, FRIDGE and E&I are the two dominant factors of EUI in cluster C 2_1. Also, HVAC, LIGHT, and FRIDGE are the main contributors in cluster C 1_3, while cluster C 1_1 shows a uniform distribution among different end-use loads. This shows that occupants in different clusters show different behaviors regarding the intensities of end-use load usage. The noticeable increase in HWS in cluster C 2_2 may be attributed to the low outside air temperature of the buildings in this cluster (this cluster comes from C 2 which includes colder regions). Also, all buildings in this cluster have electrical heaters (see Figure 4-5) and apparently occupants tend to use them more than kerosene heaters.

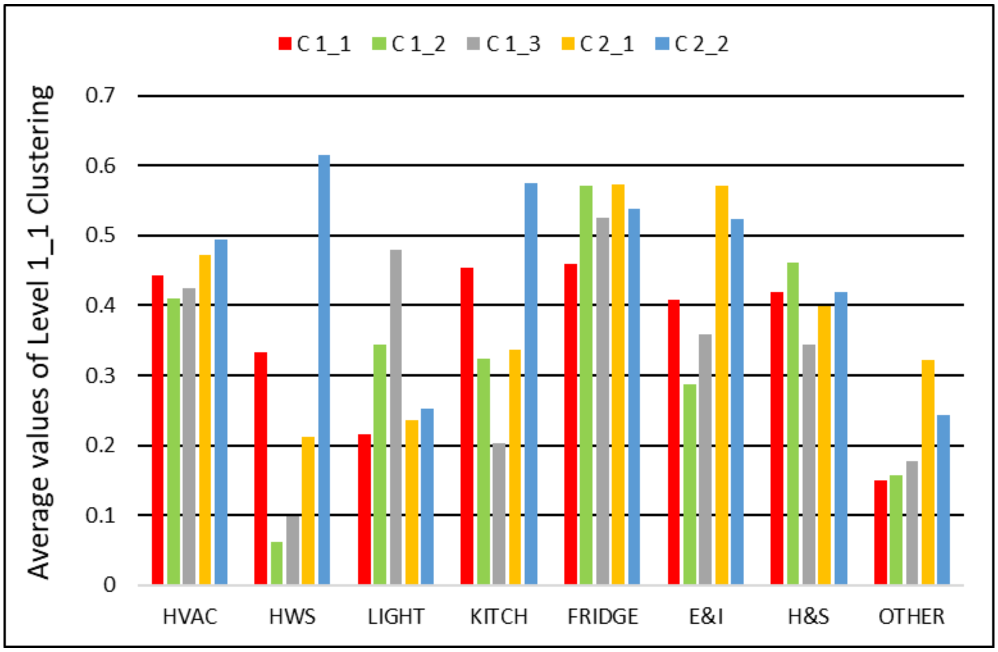

Figure 4-6. Distribution of eight end-use loads in different clusters.

Buildings in the same cluster share similar holistic characteristics, which makes it reasonable to compare them to each other to reveal the occupant effects on building EUI, while buildings in different clusters should not be compared in terms of energy consumption, mainly due to the existence of the influencers listed in Table 4-2.

4.3.4 Level 1 Ranking

To rank the buildings in each cluster to determine which buildings are responsible for the EUI increase, a second clustering was applied on each cluster based on attributes described in Section 4.2.3. The optimum number of clusters according to the Silhouette index [11] was two in all clusters. Thus, the cluster centroid with lower EUI was named Low Energy Consumer, meaning that the buildings in this cluster generally had lower energy usage. The other cluster, on the other hand, was buildings with high energy usage and was named High Energy Consumer. The occupants of these buildings need to modify their behavior in order to reduce their energy consumption.

Figure 4-7 shows the result of the level 1 rankings. Every building of the data set falls into one of the leaves of the graph shown. By following the branches of the curve, some information about the general characteristics of the buildings and weather conditions of the buildings in that cluster can be found. For example, buildings in clusters C 2_1 and C 2_2 are all in cold regions where temperature, humidity, and solar radiation, are low. Buildings in cluster C 2_1 have low equivalent leakage areas and nonelectric hot water supply and space heaters, while buildings in C 2_2 have electric heaters and hot water supply. By looking at clustering level 2, general occupant behavior is extracted. For example, buildings in cluster C 2_1 are clustered further into two groups of high and low energy consumers. Cluster C 2_1_1 is the cluster with higher energy consumption in the majority of end-use loads, such as HVAC, HWS, LIGHT, FRIDGE, E&I, and OTHER. Therefore, the occupants need to focus on these end-use loads. More information about all 10 obtained clusters is shown in Figure 4-7.

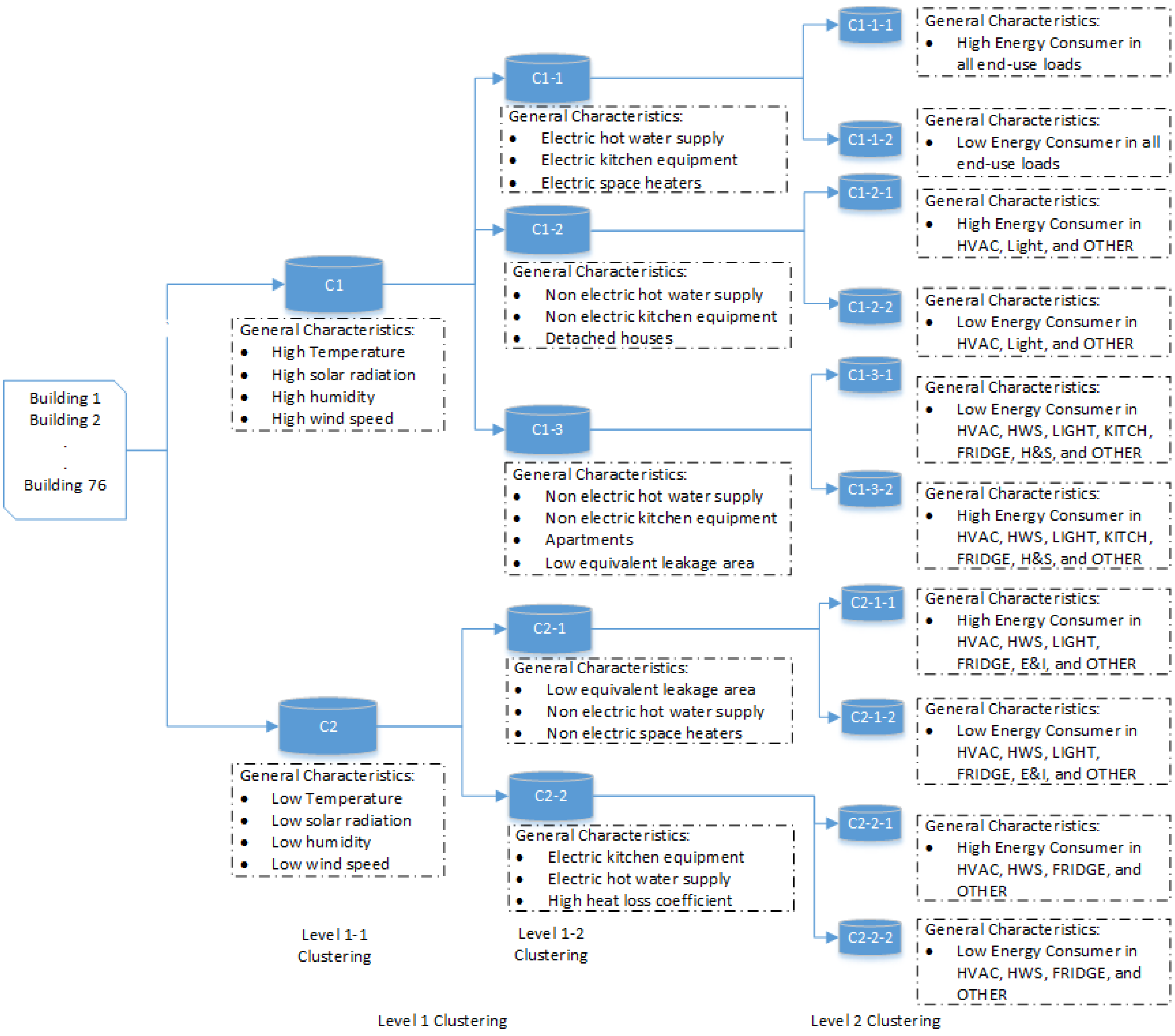
Building 1
Building 2
.
.
Building 76
C1
General Characteristics:
• High Temperature
• High solar radiation
• High humidity
• High wind speed
C2
General Characteristics:
• Low Temperature
• Low solar radiation
• Low humidity
• Low wind speed
C1-1
General Characteristics:
• Electric hot water supply
• Electric kitchen equipment
• Electric space heaters
C1-2
General Characteristics:
• Non electric hot water supply
• Non electric kitchen equipment
• Detached houses
C1-3
General Characteristics:
• Non electric hot water supply
• Non electric kitchen equipment
• Apartments
• Low equivalent leakage area
C2-1
General Characteristics:
• Low equivalent leakage area
• Non electric hot water supply
• Non electric space heaters
C2-2
General Characteristics:
• Electric kitchen equipment
• Electric hot water supply
• High heat loss coefficient
C1-1-1
General Characteristics:
• High Energy Consumer in all end-use loads
C1-1-2
General Characteristics:
• Low Energy Consumer in all end-use loads
C1-2-1
General Characteristics:
• High Energy Consumer in HVAC, Light, and OTHER
C1-2-2
General Characteristics:
• Low Energy Consumer in HVAC, Light, and OTHER
C1-3-1
General Characteristics:
• Low Energy Consumer in HVAC, HWS, LIGHT, KITCH, FRIDGE, H&S, and OTHER
C1-3-2
General Characteristics:
• High Energy Consumer in HVAC, HWS, LIGHT, KITCH, FRIDGE, H&S, and OTHER
C2-1-1
General Characteristics:
• High Energy Consumer in HVAC, HWS, LIGHT, FRIDGE, E&I, and OTHER
C2-1-2
General Characteristics:
• Low Energy Consumer in HVAC, HWS, LIGHT, FRIDGE, E&I, and OTHER
C2-2-1
General Characteristics:
• High Energy Consumer in HVAC, HWS, FRIDGE, and OTHER
C2-2-2
General Characteristics:
• Low Energy Consumer in HVAC, HWS, FRIDGE, and OTHER
Level 1-1 Clustering
Level 1-2 Clustering
Level 1 Clustering
Level 2 Clustering

Figure 4-7. Results of level ranking. Details of clusters and their general characteristics.

Figure 4-8 shows clustering centroids for each end use. By analyzing clustering level 2 results, specific occupant behavior is determined. Some of the implications are as follows:

High energy consumer buildings in cluster C 1_1 (top left graph in Figure 4-8) have higher energy consumption specifically in HVAC, KITCH, and FRIDGE, which implies that building occupants in this cluster should give primary consideration to these activities and bring their energy consumption level to low values. The activities that need more consideration in cluster C 1_2 are FRIDGE, HVAC, and OTHER. The end-use load that deserves attention in all buildings is FRIDGE, because it is the main contributor in nearly all clusters (high energy consumers), and HVAC is a main contributor in all clusters except C 2_1. Also, C 1_1 shows a uniform distribution of energy usage in end-use loads. Blue bars show the centroid of buildings at the low energy consumption level. It is important to mention that sometimes one activity may have a higher portion compared to its corresponding value in the high energy consumers, but the overall energy consumption of occupants (based on Euclidean distance in the K-means algorithm) puts them in the low energy category. Such activities are E&I in cluster C 1_3 and KITCH and H&S in cluster C 2_1. Occupants may focus on these activities to save more energy.

Level 2 clustering gives the building occupants of each cluster (level 1) a basis to reduce their energy consumption by comparing their consumption with similar group. Low energy consumers are encouraged to improve their building's performance by focusing on major end-use loads and comparison with the best building in their section.

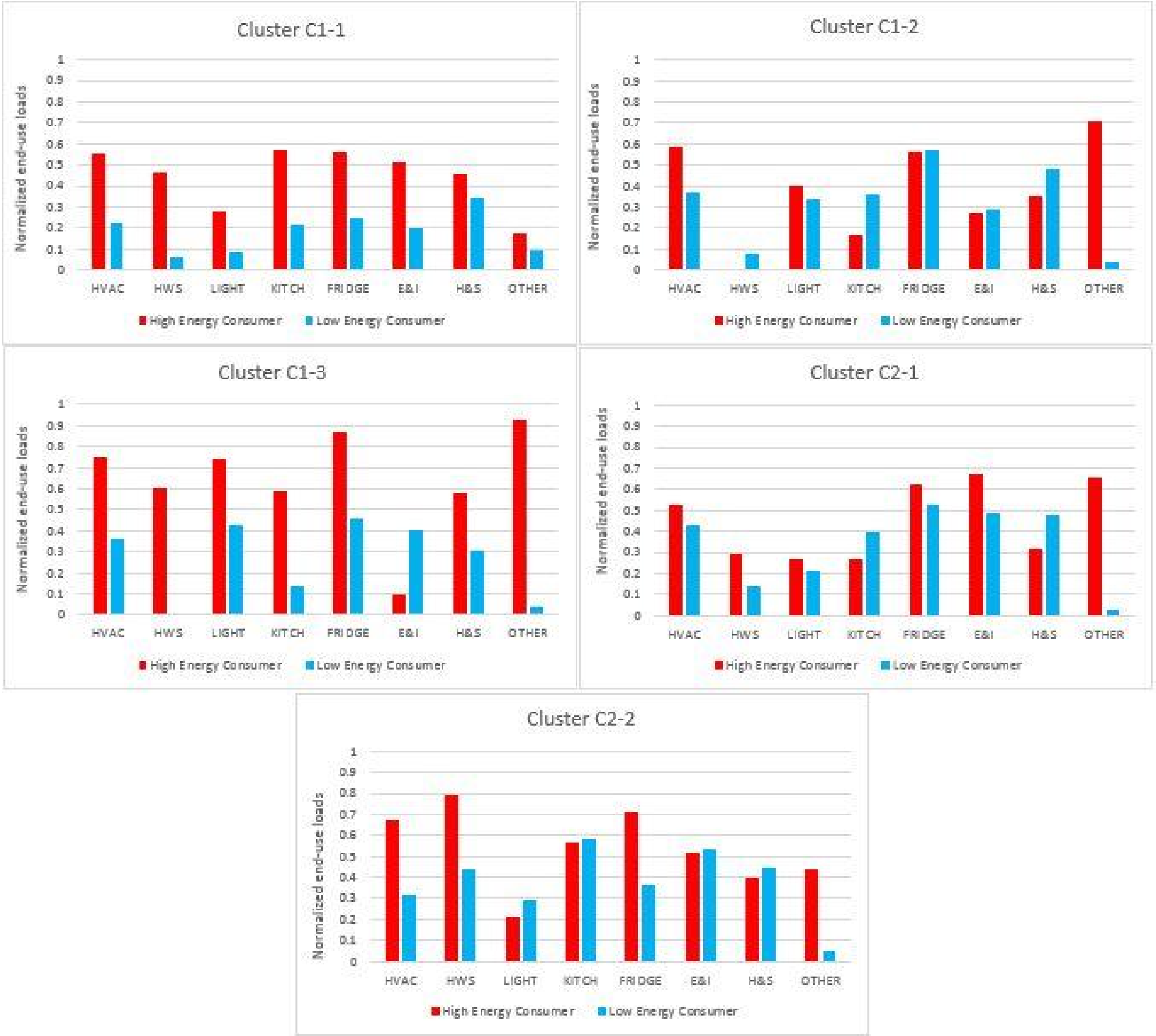

Figure 4-8. Clustering level 2 results. Data in all clusters were clustered again in terms of EUI. Centroids are shown in blue and red bars and are categorized as buildings with either low or high energy consumption, respectively.

4.3.5 Level 2 Ranking

For each building in the same cluster (level 1), the PI was calculated and the results were reported. Buildings with a higher PI have a higher place regarding energy consumption, while occupants of buildings with a lower PI are informed to take suitable actions to reduce their energy consumption level and improve their rank. Potential and achieved savings are calculated based on the extracted rules (RM and RS rules in Figure 4-3). Two sample rules are indicated in Table 4-4. The first rule indicates that when the living-room TV, kitchen rice cooker, and refrigerator consume low energy during the day, kitchen outlet lights should be switched off most of the time (based on this rule, which is derived using historical data). This occupant behavior is frequent; thus, it is expected that kitchen lights should be off when the three loads mentioned are low. This rule is categorized as an RM rule, and any record following the premise ([TV (low), rice cooker (low), refrigerator (low)]) but not the conclusion ([Kitchen light (high)]) is considered inefficient. The occupant is warned about this issue. To estimate the potential savings associated with this waste of energy, an ANN model is built based on the rule in which the input and output are the premise and conclusion, respectively (as shown in Table 4-4). Figure 4-9 and Figure 4-10 show the outputs of the models based on the rules represented in Table 4-4, which are obtained by plugging the values of the premise into the model and getting the output. The recorded (real) values and their differences are also reported in the figure. Figure 4-9 corresponds to the sample RS rule, and Figure 4-10 refers to the sample RM rule. The potential savings are calculated based on the difference between modified and recorded values, and achieved savings are estimated based on the difference between expected and recorded values. This process is repeated for all extracted rules. The cumulative potential savings are reported as a percentage (shown in Table 4-5). The achieved savings is calculated in a similar manner for RS rules.

Table 4-4. Two sample rules for calculation of achieved and potential savings.

Premise	Conclusion	Category
[TV (low), rice cooker (low), refrigerator (low)]	[Kitchen light (low)]	RM
[Microwave (low), rice cooker (low), kitchen light (high)]	[Refrigerator (high)]	RS

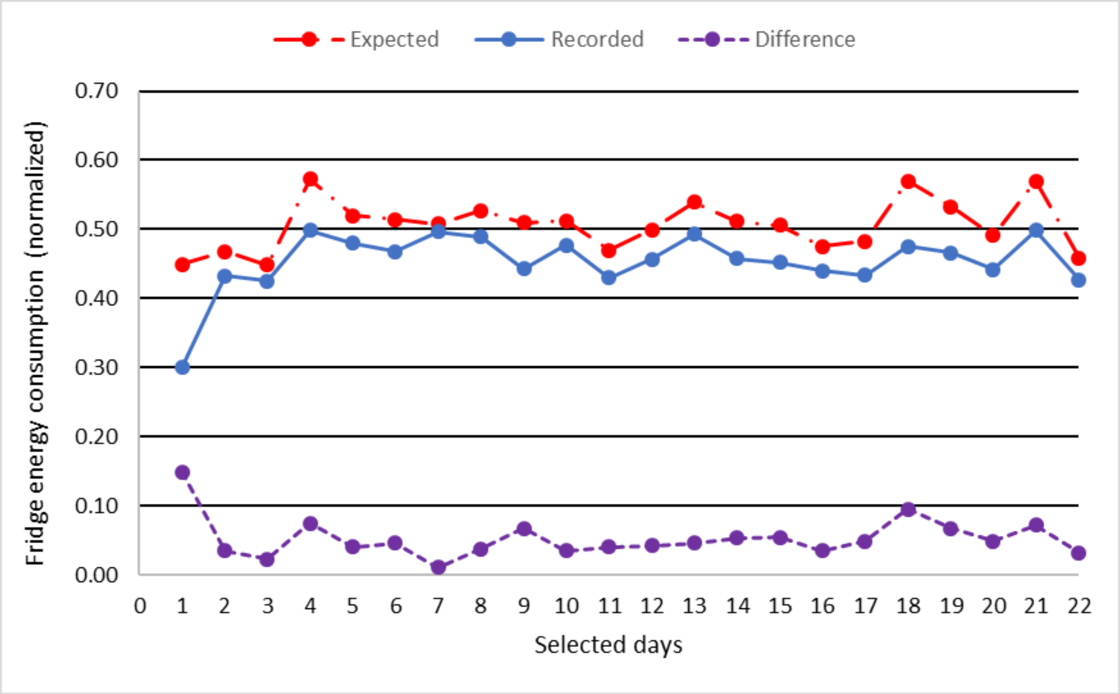

Figure 4-9. Calculation of achieved savings based on a sample RS rule.

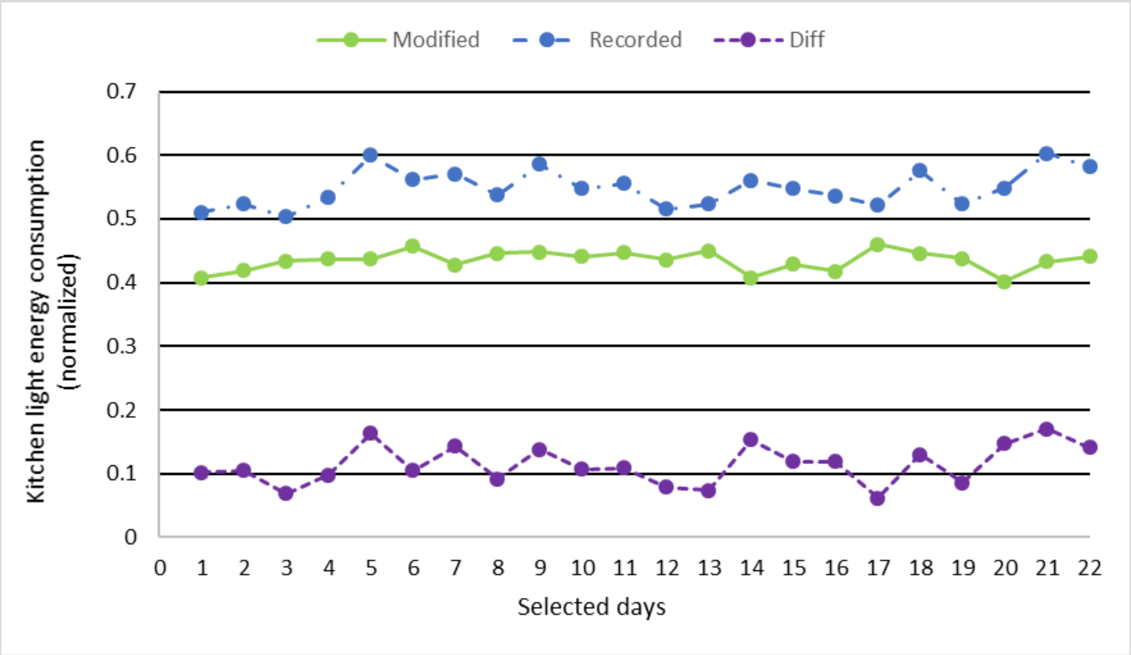

Figure 4-10. Calculation of potential savings based on a sample RM rule.

It is important to mention that in Equation 4.4.5, AS and PS are expressed in terms of percentages, and their subtraction may give negative, zero, or positive values. Negative values mean that the achieved savings are less than the potential savings, zero means they are the same, and positive means the potential savings are lower than the achieved savings, which is the best case. Table 4-5 shows part of the results in cluster C 1_1_1 (high energy consumers; there are 14 buildings in this cluster). High energy consumers (red bars in Figure 4-8) and low energy consumers ((blue bars in Figure 4-8) are ranked separately; therefore, a clear ranking of each building is given to tenants, giving them opportunities to know their place among other buildings and see how to improve their rank. For example, based on Table 4-5, building 3 in cluster C 1_1_1 is a high energy consumer according to clustering level 2 (red bars in Figure 4-8, top left). Therefore, the tenants are advised to try to modify their behavior (especially on HVAC, KITCH, and FRIDGE based on Figure 4-8, top left) to improve their place. This building shows a relatively good performance in terms of PI because it is in second place among the other four buildings, with a PI of –2%. The best building has a PI of 1%. Similar interpretations are possible for other buildings in the data set. Similar results are obtained and can be reported to the occupants of other buildings.

Table 4-5. Part of results of two-level ranking system for buildings in cluster 2 in high energy consumers.

Building No.	Cluster	Level 1 Ranking	Level 2 Ranking	AS	PS	PI
1	C1_1_1	High Energy Consumer	4	12%	21%	–9%
2			3	10%	15%	–5%
3			2	10%	12%	–2%
4			1	11%	10%	1%

4.4 Conclusions

A novel two-level ranking system for a set of buildings was proposed based on occupant behavior and activities. Buildings were first clustered using the K-means method into two levels, levels 1-1 and 1-2, to reduce the effects of non–occupant-related factors and put buildings into separate groups. The differences between the buildings' energy consumption in the same clusters are attributed to occupant roles. A second clustering in terms of eight end-use loads was performed in each group to yield a level 1 ranking for each building (high and low energy consumers). Performance index was defined in terms of achieved and potential savings to determine the amount of savings for each building based on detailed operational data and was named level 2 ranking. Results show that, using the information provided by the two ranking levels, tenants of a certain building are able to understand their performance in terms of energy usage compared to other buildings and get recommendations on how to reduce their energy consumption and improve their rank.

5. Systematic Approach to Provide Building Occupants with Feedback to Reduce Energy Consumption

5.1 Introduction

Occupants' contributions to the energy consumption of buildings could be significant. By monitoring occupants' behavior in this regard, we can find opportunities to save energy. This is important in the sense that modifying occupant behavior is an inexpensive way to reduce building energy consumption, especially if the occupants could benefit financially (reducing energy bills) and without any additional costs. Occupant behavior means occupants' presence, activities, and operations. One of the best ways to assess this behavior is through comparison. Climate conditions, building envelope, and building systems can influence occupant behavior. However, if two buildings with the same types of factors are studied, we may notice, for example, that one of the buildings has lower energy expenditure toward a certain end-use load. This shows that some occupants are aware of their activities or behavior, which could help reduce energy consumption. For example, the occupants of one building may use little energy for kitchen appliances, the operation of heating, ventilation, and air conditioning (HVAC), or both. This can be due to several reasons, such as culture or lifestyle. On the other hand, occupants may overuse some appliances, which makes them high-energy consumers. Based on this, it is possible to create a building that has all the positive characteristics of low-consumption buildings, so we can alert the occupants of the building in question of how much energy they can save by copying the reference building (RB). The idea is shown in Figure 5-1. As expected, buildings have different occupants and characteristics, such as the number of people, age, the level of activities, building thermal characteristics such as wall area, insulations, etc., and weather conditions such as cold and humid, or warm and dry climates. Therefore, one cannot simply compare the high-energy-consuming buildings with those with the lowest energy consumption. The main objective of the current project is to generate an RB so that any other building can be compared with it. The RB should be very similar to the building in question in terms of number of occupants, building characteristics, and weather conditions. In the previous studies, occupant behavior was analyzed using data mining of several existing buildings to reveal opportunities to save energy [1-3]. In this study, the development of a novel methodology to generate an RB from data on existing buildings for performance evaluation is described. The result is a generic, accurate, and scalable tool to evaluate the energy consumption of a given building. In the following sections, an overview of previous works is presented before the current work methodology is introduced.

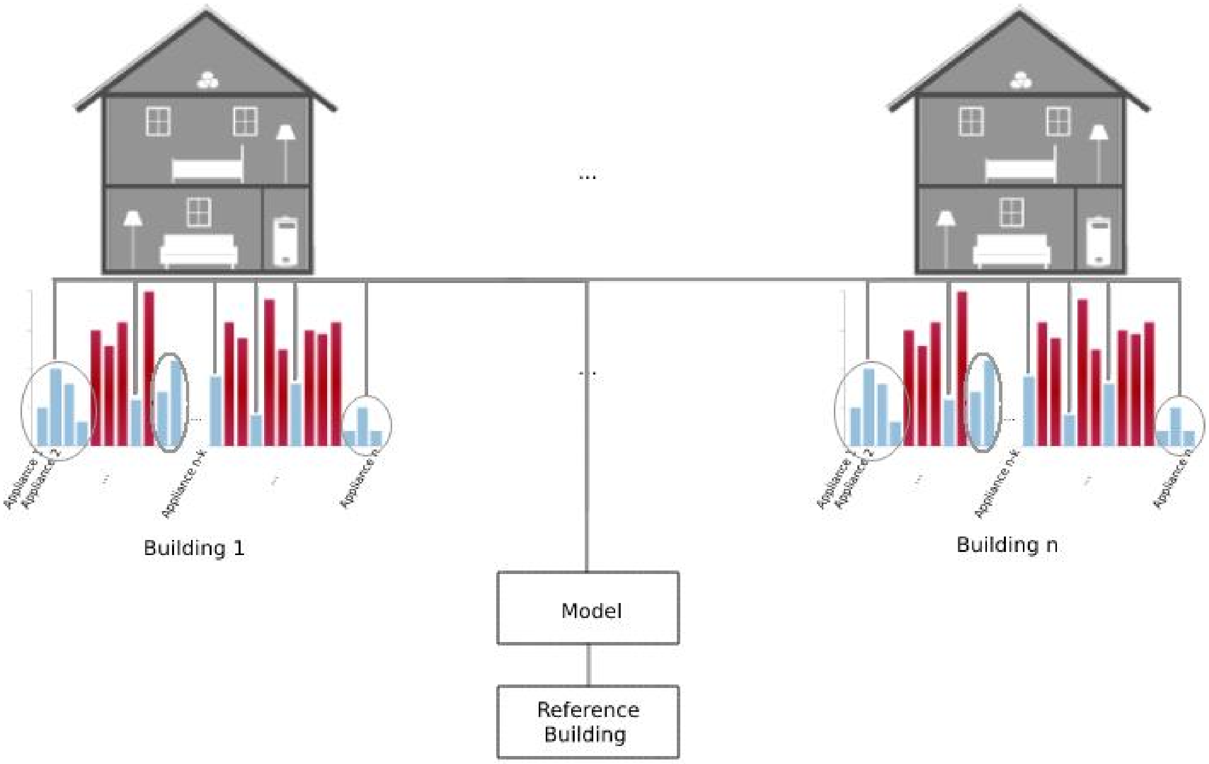

Figure 5-1 Creation of reference building through analysis of appliances energy consumption.

In this study, a new approach for evaluation of energy consumption in a single dwelling is presented, using clustering analysis and ANN models. The following is a summary of the contributions of the current work:

1. The current paper is the first featuring a proposal to create a nonexistent RB to assess the energy consumption of a given building. Knowing that each building may show a low energy consumption pattern regarding one specific end-use load such as HVAC or lighting, it is possible to make an RB that contains all energy-saving characteristics of different buildings. Therefore, for a given building, its RB is created by the proposed methodology. Potential savings are revealed through comparison of the given building with the RB.

2. The creation of an RB through this methodology is an accurate approach to assess the real performance of a building. (A comparison of existing work with another authors' publication is described in section 5.3.4)
3. The methodology introduced here is a generic and scalable approach, meaning that by increasing the number of buildings under investigation, we can achieve even better and more robust models as RBs to assess the performance of a given building.

5.2 Methodology

Figure 5-2 shows the overview of the framework. The tasks are outlined in the framework: 1) data selection; 2), data aggregation; outlier detection and diagnosis; and normalization; 3) database creation; 4) grey relational analysis (GRA); 5) level 1-1 clustering; 6) level 1-2 clustering; 7) level 2 clustering; 8) cluster ranking and combination; and 9) model development. Each task is meant for a purpose and is applied to a portion of the dataset. The designed process is as follows:

1) Data on energy consumption from each building are selected, summarized, cleaned, and integrated together. The data consist of weather parameters, building characteristics, and energy consumption of all end-use loads. Energy consumption data are averaged to annual values and normalized and stored in the database. This methodology uses the annual building energy consumption, but it can be further improved to show seasonal energy consumption.
2) GRA is performed on the database to give weight to input parameters. The contribution of each parameter on a specific end-use load may be different. GRA is performed for each specific load separately for use in generating the RB.

3) A two-level clustering is performed on the dataset to group buildings with similar characteristics in terms of climatic and physical and occupant information, as shown in Figure 5-2. This step is called level one clustering.

4) A second clustering is performed on each resultant dataset to place them into groups of low- and high-energy consumers.

5) Clusters containing buildings with low-energy consumer tags are combined to develop the dataset of low-energy consumers. Note that steps 3–5 are performed for each end-use load separately.

The following sections describe each step in more detail.

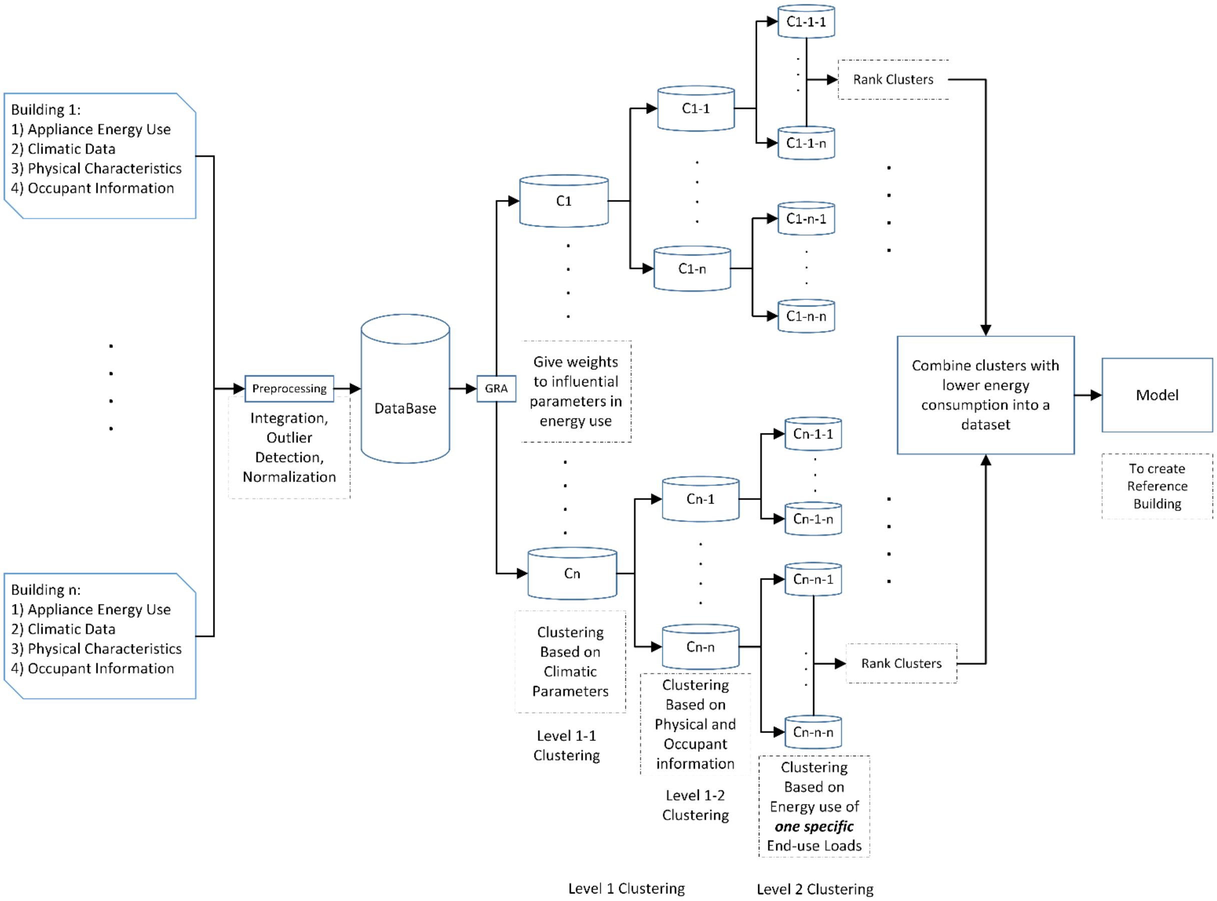
Building 1:
1) Appliance Energy Use
2) Climatic Data
3) Physical Characteristics
4) Occupant Information
Building n:
1) Appliance Energy Use
2) Climatic Data
3) Physical Characteristics
4) Occupant Information
Preprocessing
Integration, Outlier Detection, Normalization
DataBase
GRA
Give weights to influential parameters in energy use
C1
Cn
C1-1
C1-n
Cn-1
Cn-n
C1-1-1
C1-1-n
C1-n-1
C1-n-n
Cn-1-1
Cn-1-n
Cn-n-1
Cn-n-n
Rank Clusters
Rank Clusters
Combine clusters with lower energy consumption into a dataset
Model
To create Reference Building
Clustering Based on Climatic Parameters
Level 1-1 Clustering
Clustering Based on Physical and Occupant information
Level 1-2 Clustering
Clustering Based on Energy use of one specific End-use Loads
Level 1 Clustering
Level 2 Clustering

Figure 5-2. The proposed data mining framework. The inputs are home appliance energy use, weather, physical data, and occupant information. The output is the model to create an RB.

5.2.1 Data Selection

The data containing the information about home appliance energy use, weather data, building characteristics, and occupants' information were input. Home appliances' energy use were summed up in eight categories as follows:

1) HVAC
2) Hot water supply (HWS)
3) Lighting (LIGHT)
4) Kitchen (KITCH)
5) Refrigerator (FRIDGE)
6) Entertainment and information (E&I)
7) Housework and sanitary (H&S)
8) Other end-use loads (OTHER)

The energy consumption of each of these end-use loads was monitored and measured every day for each building. There are 76 buildings in total, and the study was performed between 2002 and 2004. Therefore, approximately $3 \times 365 \times 76 = 83220$ datapoints were collected. Climate data included:

1) Annual average outside air temperature (T)
2) Annual average relative humidity (RH)
3) Annual mean wind speed (WS)
4) Annual mean global solar radiation (IR)

In some cases, where the values were missing, the nearest weather station's data were used. Building physical parameters were:

5) House types (HT) (detached or apartment)

6) Building area (A) (m^2)

7) Equivalent leakage area (ELA) (W/m^2K)

8) Type of space heating and cooling (HC) (electric or nonelectric)

9) Type of HWS (electric or non-electric)

10) Kitchen equipment (KE) (electric or nonelectric)

Occupant information included only:

11) Number of occupants (NO)

5.2.2 Data Aggregation, Outlier Detection, and Normalization

The data from all buildings were aggregated, and outlier detection was performed using the lower and upper quantile method. This means that data lower than $\mathrm{Q1} - 1.5(\mathrm{Q3} - \mathrm{Q1})$ or greater than $\mathrm{Q3} + 1.5(\mathrm{Q3} - \mathrm{Q1})$ were considered as outliers, where $\mathrm{Q1}$ and $\mathrm{Q3}$ were the first and third quantiles, respectively. After outlier removal, they were filled using a regression model based on other available attributes. Data were then normalized using min–max normalization to obtain the values between 0 and 1. Furthermore, the categorical values were converted to 0, 1, or in between.

5.2.3 Database Development

Once pre-processing was completed, the database was developed for the DM tasks. For each task, a subset of the attributes was selected. For example, in level 1-1 clustering, climate data were used to cluster the buildings into similar groups, and other attributes were kept, while in level 1-2 clustering, the building's physical information was used. Finally, in level 2 clustering, the energy consumption data of a specific end-use load were used.

5.2.4 Grey Relational Analysis

GRA identifies the causative factors of an objective (energy use, here) and weighs them according to their contribution[60]. The contribution of each of these factors in the overall energy consumption of buildings varies greatly. For example, a variation in the number of occupants may have larger effects on the KITCH energy consumption than would variation in the house type. In this study, GRA was applied to the following variables: temperature, relative humidity, solar radiation, wind speed, house type, equivalent leakage area, heat loss coefficient, number of occupants, heating/cooling type, hot water supply, and kitchen equipment. The GRA results were then multiplied by their corresponding variable. The process was repeated for each end-use load. Results were stored in the database for the next steps.

5.2.5 Level 1-1 Clustering

This step filters out the effects of weather parameters by clustering buildings with similar weather conditions together. The attributes used for this clustering are annual mean air temperature, annual mean relative humidity, annual mean wind speed, and annual mean global solar radiation. The output of this clustering is groups of buildings with similar weather parameters. It is worth mentioning that it is possible to prioritize the attributes used in this clustering, for example, making the outside temperature the main attribute by giving it a weight. However, in this study, all four climatic attributes were given the same importance (weight) and used in clustering.

5.2.6 Level 1-2 Clustering

At this stage, the buildings are grouped by occupant information (number of occupants) and building physical characteristics (house type, equivalent leakage area, heat loss coefficient, heating/cooling type, hot water supply, and kitchen equipment). The result is the grouping of buildings with similar occupants and physical parameters.

5.2.7 Level 2 Clustering

The differences between energy consumption of buildings in level 2 clustering are due to different occupant behaviour. At this level, the buildings are clustered in terms of energy use intensity (EUI, energy use per unit of area), which is an indicator of the energy consumption of the occupants. The steps in sections 2.4–2.7 are repeated for all eight end-use loads (mentioned in section 2.1) separately and stored in the database.

5.2.8 Cluster Ranking and Combination

The result of level 2 clustering is the grouping of buildings with similar energy use regarding one specific end-use load (for example, HVAC energy consumption). The centroids of clusters (average energy use of each cluster) are then sorted from the highest to the lowest: the cluster with the lowest centroid represents building occupants who are cautious regarding energy use of the end-use load under investigation. By extracting all low consumer buildings and their corresponding information (listed in level 1-1 and 1-2 clustering), we were able to make a database, which was a subset of the original database but contained only low-consumption buildings, as shown in Figure 5-2. This process was performed for all eight end-use loads listed in section 2.1. The resulting data lay the foundation to create the RB.

5.2.9 Model Development

The aim of this step is to use the developed databases as the base to generate the RB. As mentioned before, all buildings in the created datasets were low-consumption buildings. The goal was to develop a model that could estimate the energy consumption of an end-use load. The model inputs were number of occupants, climatic conditions, and building physical information. The model took the inputs and estimated how much energy the building would approximately consume if it belonged to the set of low-energy-consumption buildings. The model used an artificial neural network to map inputs (characteristics) to outputs (energy consumption). This process was performed for all eight end-use loads, and a model was created for each end-use load. Therefore, the RB consisted of eight ANN models. The outputs of the RB were in the range of low-consumption buildings because the databases were created using such buildings. After comparing the model outputs and real values from a given building, we could evaluate said building's thermal performance. This is shown in Figure 5-3. It is worth mentioning that the RB was customized to each given building, considering that the inputs of the models came from the given building and those developing the RB estimated the desired values from the low-consumption buildings using an ANN. The ANN model used in this study was a multilayer perception model with regularization parameters. The Scikit-learn package of Python 3.7.3 was used for the analysis, and the optimal parameters were chosen according to the grid-search method available with the package scikit-learn [62]. A three-layer neural network was used in all cases, and the activation function, regularization parameter, and learning rate method were chosen by the grid search.

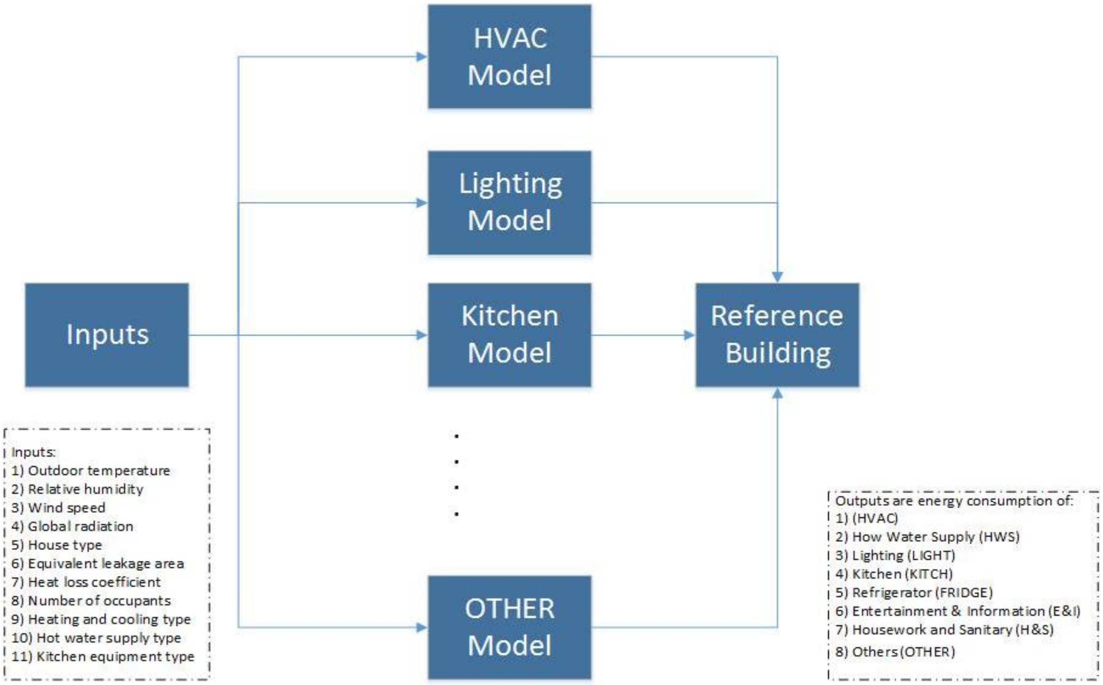

Figure 5-3. Creation of reference building using eight ANN models.

5.3 Results and Discussion

5.3.1 Grey Relational Analysis

Table 5-1 shows the result of GRA for every end-use load. As can be seen, for every load, the contributions of characteristics are different. The colors in each column sort the values from highest to lowest. Pure red shows the maximum, pure green shows the minimum, and others are in between. For example, in HVAC energy consumption, NO, followed by T, are the key variables, while HT is the least important characteristic. This shows that the type of dwelling (detached or apartment) has the least effect on HVAC energy consumption by occupants. Regarding HWS, RH and NO are the dominant factors, which seems reasonable, while KE is the least important characteristic. In general, the important characteristics are NO, HLC, T, and IR, while HT and KE are considered less important characteristics. Weather factors and NO play a greater role than building characteristics. The contributions of each characteristic were multiplied by the normalized energy consumption data for clustering purposes (next section).

Table 5-1 Results of GRA for each end-use load

	HVAC	HWS	LIGHT	KITCH	FRIDGE	E&I	H&S	OTHER
HT	0.625	0.653	0.637	0.562	0.614	0.621	0.588	0.707
NO	0.778	0.726	0.732	0.765	0.812	0.777	0.763	0.698
HLC	0.77	0.708	0.697	0.739	0.802	0.751	0.779	0.679
ELA	0.734	0.696	0.694	0.706	0.723	0.696	0.74	0.702
HC	0.688	0.683	0.645	0.641	0.699	0.704	0.673	0.73
hws	0.744	0.639	0.707	0.649	0.751	0.712	0.725	0.696
KE	0.667	0.605	0.645	0.546	0.651	0.658	0.642	0.691
T	0.773	0.688	0.721	0.709	0.783	0.747	0.753	0.671
RH	0.722	0.734	0.668	0.66	0.693	0.703	0.704	0.767
WS	0.753	0.668	0.707	0.702	0.766	0.699	0.768	0.677
IR	0.768	0.684	0.707	0.738	0.806	0.766	0.784	0.66

5.3.2 Clustering Analysis

Figure 5-4 shows the result of data clustering regarding HVAC energy consumption; Df represents the original dataset after applying GRA. The number of clusters was determined using the silhouette index [53] which was two in this case. Df1 and Df2 are the output of level 1-1 clustering regarding climatic conditions of the buildings. This means that all buildings in Df1 and Df2 have similar weather parameters. Cluster Df1 contains buildings in warmer regions considering that in their centroid, temperature and solar irradiation have higher average values. Clustering level 1-2 divides each cluster further into groups of buildings with similar physical characteristics and NO (Df1-1 to Df1-2), as shown in Figure 5-4. Df1-1 contains buildings with a nonelectric HWS and KE, whereas Df1-2 is all electric. The details of clustering are shown in Figure 5-4. The last clustering step is level 2, which partitions data based on target end-use load (HVAC energy consumption here). This last clustering step is performed based on the end-use load. As seen, the high-energy-consumption buildings have normalized HVAC consumption of more than 0.60 in all clusters, while their low-consumption counterparts use less than 0.30 in all clusters. Through comparing cluster centroids of level 2 clustering, the clusters with lowest energy consumption were selected and combined. This dataset made the "low-consumption buildings," which are shown in Figure 5-4 in the case of HVAC energy consumption. It is important to note that clustering level 2 may end up with more than two clusters (here, we ended up with two clusters, so we named them low- and high-energy-consumption buildings). The buildings belonging to lower consumption groups are chosen and combined regardless of the number of clusters.

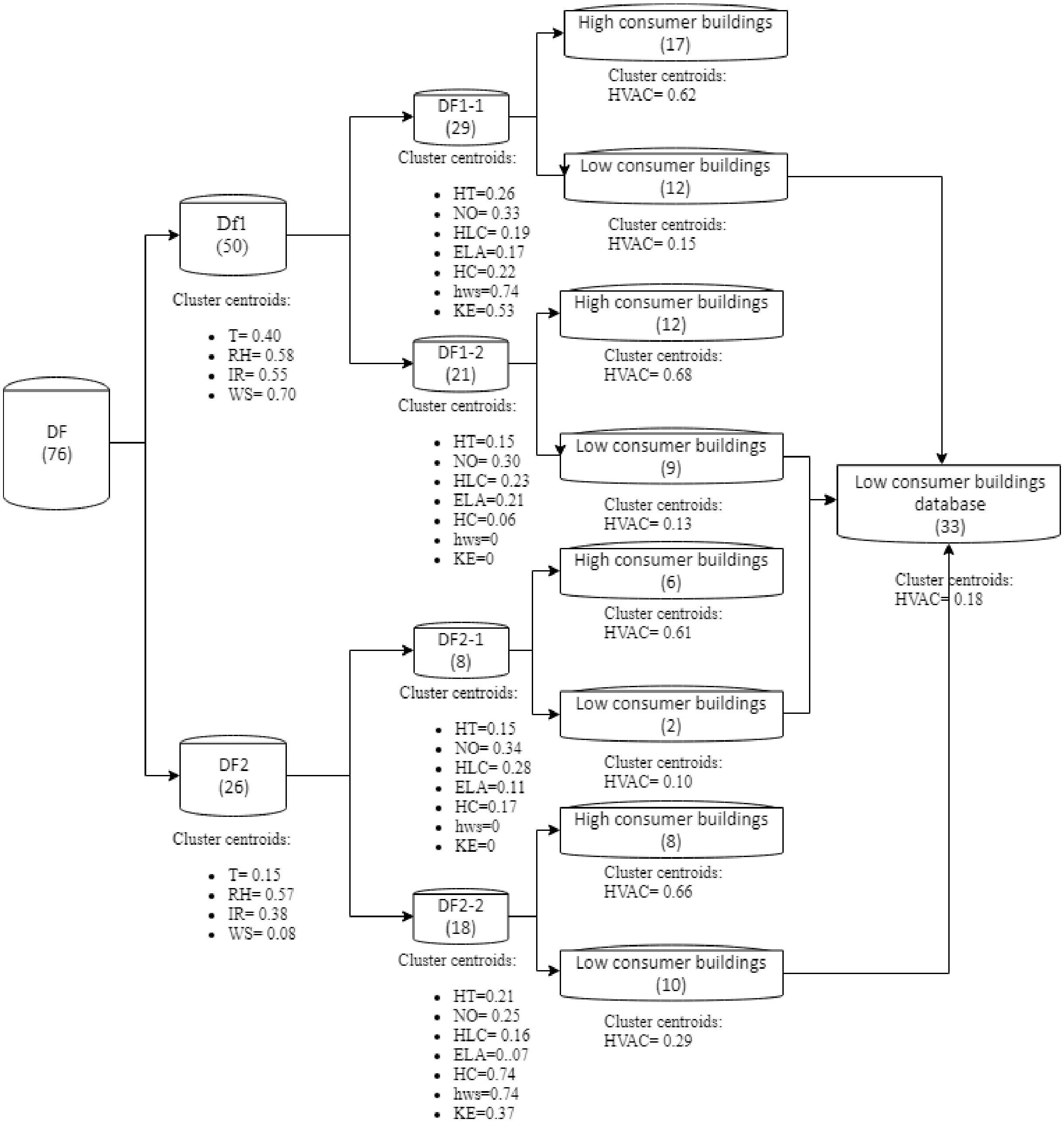

Figure 5-4. Clustering schematic considering HVAC energy consumption as the target.

5.3.3 Reference Building Generation and Building Performance Evaluation

Figure 5-5 shows the result of clustering regarding HVAC energy consumption. There are two sets of buildings: one with higher energy consumption (shown in orange) and the other with lower energy consumption (shown in blue). The high-consumption buildings are the combination of clusters with high energy consumption, while low-consumption buildings are the collection of clusters with low energy consumption. The ANN model used the low-consumption buildings as the training dataset. Therefore, by plugging in the characteristics of any given building (shown in red in Figure 5-5), the model output is an estimation of how much the HVAC energy consumption of the given building would have been had it belonged to low-consumption buildings (RB shown in red in Figure 5-5). The difference between the actual and estimated value of the given building and the RB shows the possible energy consumption savings. It is important to mention that because the training dataset comes from a combination of clusters, the input of the model is sufficiently diverse to show that the model is a good estimator.

The number of low consumer buildings in Figure 5-5 (shown in blue) is a function of three steps clustering. It means that the higher number of clusters divides the dataset into smaller groups in size. Therefore, it is up to the expert to choose the groups with lowest possible energy consumptions or choose a wider range of clusters as low consumer buildings. This in fact affect the reference building from a stricter (the lowest consumer) to a less strict behavior (higher energy consumer).

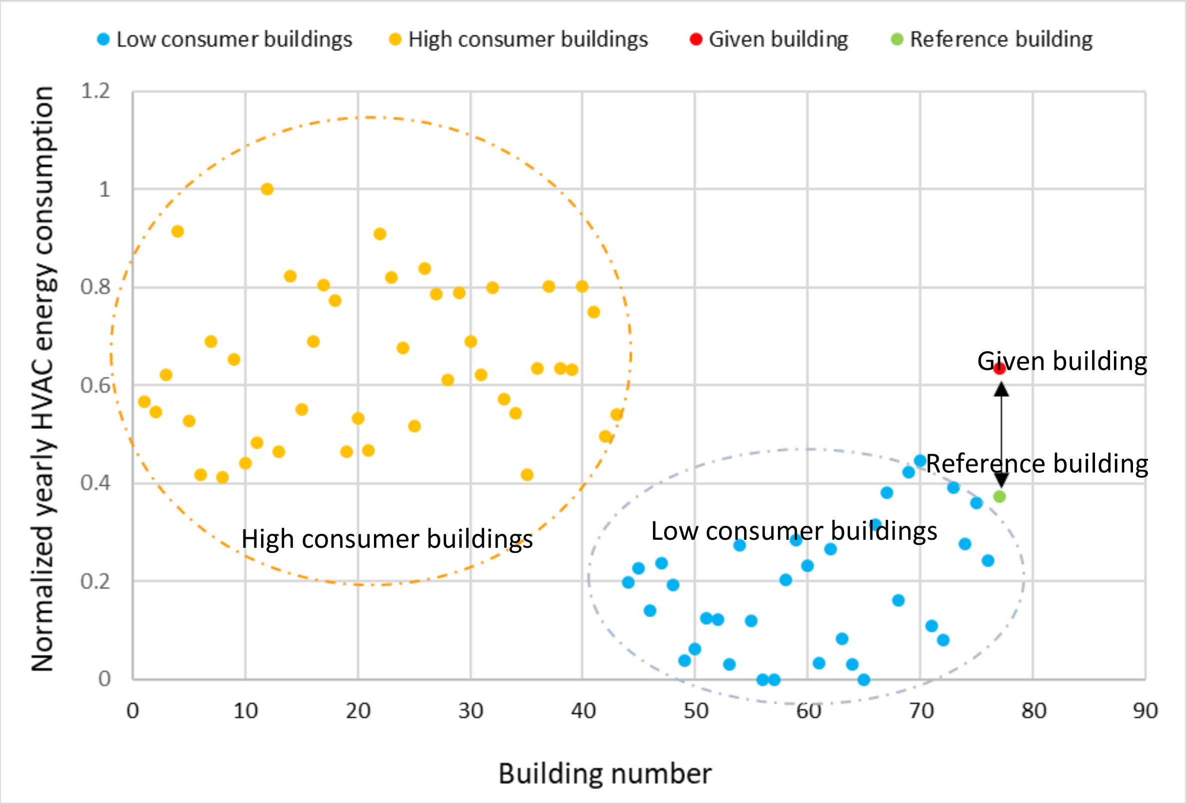

Figure 5-5. Distribution of buildings regarding HVAC energy consumption, showing clusters of high- and low-consumption buildings. Building number in x-axis is representative of buildings only and has no other meaning (no ordering of buildings).

Table 5-2 shows the results of the eight models, along with the real consumption of eight end-use loads. The difference between them is the possible energy savings. For example, HVAC energy consumption is 0.634 (normalized values) in the given building and 0.256 in the RB; the difference (0.378) is the possible energy-saving potential. The same logic applies regarding HWS, FRIDGE, E&I, H&S, and OTHER. However, it is evident that LIGHT and KITCH have negative energy savings. This shows that the given building is performing better than the RB. In other words, the occupants of this building are cautious regarding LIGHT and KITCH appliances and may be focusing on other end-use loads to reduce their bills even further.

Table 5-2 Comparison of given and reference building for calculation of possible energy savings

	HVAC	HWS	LIGHT	KITCH	FRIDGE	E&I	H&S	OTHER
Given building	0.634	0.189	0.102	0.092	0.639	0.754	0.366	0.605
Reference Building	0.256	0.062	0.144	0.601	0.527	0.507	0.342	0.033
Energy savings	0.378	0.127	-0.042	-0.509	0.112	0.247	0.024	0.572

5.3.4 Assessment of the Methodology

Table 5-3 shows the result of the given building assessment through the developed methodology and the previous work [64] (a different approach, based on mere comparison of similar buildings with each other). It is observed that for all end-use loads except KITCH, the RB implies that there are more energy-saving opportunities, while according to the previous authors' work, there are limited or no such opportunities. This is because the RB assumes the advantage over all low-consumption buildings, while the previous work took similar buildings with lower average energy consumption than the given building.

Table 5-3 Comparison of given building through reference building and comparison with existing buildings [64].

	HVAC	HWS	LIGHT	KITCH	FRIDGE	E&I	H&S	OTHER
Given building	0.634	0.189	0.102	0.092	0.639	0.754	0.366	0.605
Energy savings according to reference building	0.378	0.127	-0.042	-0.509	0.112	0.247	0.024	0.572

Energy savings according to comparison with existing buildings [64]	0.161	-0.02	-0.135	-0.245	0.067	0.182	-0.033	0.284

5.4 Conclusions

A general methodology was developed to generate an energy-efficient building—RB—using field-measured data. The generated RB could be used to evaluate the thermal energy performance of any given building and provide information to building occupants about their building energy performance with respect to any specific end use (such as HVAC, kitchen appliances, entertainment, etc.). A dataset of 76 buildings was studied to extract buildings with low energy consumption regarding the specific end use. Considering eight end-use loads, a dataset of eight buildings with low energy consumption was extracted from the original dataset and named as a set of low-consumption buildings. By building neural network models using these low-consumption buildings as the training dataset, mapping of building characteristics to an estimation of energy use was performed. Therefore, any given building could be compared with its RB (which was generated using ANN models). The results show that by comparing the two, it is possible to spot the end-use loads that are consuming more than the RB; therefore, the occupants may focus on those energy uses and take measures (i.e., turning off unnecessary lights or HVAC, etc.) to improve building energy performance. The methodology introduced here can be applied to larger building datasets in any climate.

6. Summary, Limitations and Future Work

6.1 Summary

Buildings are complicated systems with numerous interactions between components. As a result, it is typical for there to be energy loss in buildings. Moreover, the literature has underestimated or neglected the role of a building's occupants in energy consumption and energy savings. This thesis leveraged data analysis as a powerful technology to extract hidden information in data generated by occupants of buildings. The overall goal of this study was to find energy saving opportunities for building occupants that would motivate them to follow recommendations for applying energy saving measures. This thesis's basic ideas include finding correlations between home appliance energy use, ranking similar buildings in terms of energy use, and estimating possible energy savings using data analysis tools. Three tasks were defined in this thesis. The first task focuses on a single building, examining all the correlations between the home appliances it contains, which allow us to spot inefficient energy consumption patterns. The second task is based on a clear comparison method between occupants of several buildings that enables them to find their place among other buildings and learn how to modify their performance. The last task focuses on developing a methodology to create a RB to serve as a model for a given building to estimate possible energy savings by comparing a given building with its reference building. Data analysis tools used in this thesis included clustering, association rule mining, and neural networks.

6.2 Limitations

Although developed methodologies can be applied to any building in any climatic region, some difficulties must be overcome. Most of this study's limitations come from data deficiency, which affects the results obtained here; however, in this thesis an effort was made to develop general methodologies that can be applied to similar data. The following is a list of limitations and challenges that arose during this study.

- Lack of hourly data for end-use loads is an important challenge that makes associating rule mining processes less realistic given that services might not be operating simultaneously. Consequently, recommendations are less accurate. The availability of hourly data would improve the accuracy and reliability of the process beyond that given by daily data. Despite these shortcomings, the methodology introduced in this study remains valid.
- The obtained rules in this study originate from the measured energy consumption of building occupants and therefore provide the basis for energy reduction recommendations. However, the occupants of a single building may not be aware of possible energy savings. If the rate of resident energy usage is almost always high, good behaviors are not detected in historical data; thus, no RM rule is found. Domain knowledge enables us to introduce some artificial rules (similar to real rules) that may help find energy-inefficient behaviors. For example, at noon (while tenants are at work) and the lights are off, appliances such as televisions, stereos, kitchen appliances, or computers should be turned off. This rule could also be applied to closing and opening windows.
- The energy consumption data in this study were collected annually. However, a seasonal analysis would enhance the accuracy of detection of low-consumption buildings. For example, a building whose HVAC energy consumption is high in winter and low in summer could be categorized among low-consumption buildings when considering yearly energy consumption. However, the same building may be a high energy consumer if one considers only the winter season or a low consumer if one considers only summer. In addition, we have used annual weather data in clustering level 1-1, which does not capture

underlying fluctuations in weather parameters. One solution would be to include standard deviations of the weather parameters or to use 24-hour temperature profiles of each building throughout the year.

- More information about individual buildings would enhance the process. For instance, the number of occupants is present in the dataset; however, more information (e.g., age, level of activity, number of adults and children, time of arrival and departure) could be used in clustering analysis when putting similar buildings together. Knowing this information would also help us provide more detailed recommendations based on which activities or behaviors of children or adults consume the most energy. The most useful information would be occupants' daily schedules, preferences (e.g., lighting level, room temperature), and holidays.
- In clustering level 1 (Section 4.2.2 in chapter 4), it is quite possible that buildings in the same cluster might have similar non-occupant-related characteristics but may be dissimilar in other respects. To reduce the effect of this problem, we multiplied each characteristic by its contribution to energy consumption by means of GRA. Therefore, those characteristics with higher GRA values dominated in determining the cluster to which the building belonged. In addition, clustering level 1 was divided into two subsections, levels 1-1 and 1-2, to prioritize the weather parameters in clustering. If one attempts to increase the clustering accuracy in terms of any characteristics, one can increase the subsections of level 1 clustering to even more than two (e.g., level 1-3 as number of occupants). However, this makes the cluster sizes smaller, and we were limited by the database size in this study. Increasing the size of the database can resolve this issue, as the next point discusses.

- The number of buildings in this study was 76. Including more buildings in the clustering analysis would enhance it by breaking it into more levels (e.g., levels 1-1, 1-2, 1-3, etc.) in order of importance of variables. Furthermore, sufficient data points would improve creation of the neural network models. In this case, a few thousand buildings would work well.
- High energy consumers who receive feedback are expected to take suitable measures to improve their ranking. A real case evaluation to see the effectiveness of the proposed system can achieve this, especially if the ranking is performed online so that occupants can more quickly see the effect of their energy-saving measures.

  In chapter 4, the first part of the methodology (level 1 ranking) is based on comparison between several buildings. If the low energy consumers are wasting some energy, there is still room for improvement that is not identifiable through this methodology. In other words, the buildings are not comparing themselves with the best cases. Chapter 5 addresses this limitation by introducing the reference building

6.3 Future Work

The current study aims to provide a complete DM framework for knowledge discovery and information retrieval in the building engineering domain. However, buildings are complex systems, and the application of DM in building energy analysis still yields many opportunities. Here are some suggested directions for future research:

- Combining simulation and data analysis for modelling occupant behavior. Using mathematical simulations can be helpful in representing a typical building occupant. Data analysis can help customize the occupant for any specific building by analyzing data and learning the behavior of actual occupants.

- Applying data analysis to data coming from building control systems. Huge savings are possible by using real data to build predictive models to estimate the heat/cool load of a building and adjust the control set-points/scheduling of the building.
- Automated fault detection and diagnosis. Buildings are equipped with hundreds of sensors and controls. The analysis of such massive amounts of data can reveal insights for building owners to optimize the building infrastructure. Leveraging powerful big data analytics to automatically detect and diagnose faults in the HVAC system, reducing operating costs and utility bills, increasing equipment life, and improving tenant comfort are worthy areas of research that still are in the initial phase.

  Some DM methods have been successfully used to address the problems within the building engineering domain in this research. However, other advanced methods are useful for analyzing measured building-related data and extracting useful knowledge. For example, the deep learning method could be used to predict numerical variables in building energy demand modeling. Using deep learning was not justifiable in this study because of the relatively small size of the available datasets and the low number of features under study

6.4 Guidelines for application of the research

It is important to mention that all methods established in this thesis are generic and scalable; therefore, it is applicable to any building(s) to obtain the results. The higher resolution data (e.g. hourly versus daily) can help increase the accuracy of the methodologies. Also, the methodologies introduced in this thesis can be applied on specific seasons (hot season vs. cold seasons) to obtain a better understanding of occupant role in energy consumption regardless of climatic conditions.

There are some parameters needed to be set before applying algorithms. Cluster analysis, association rule mining, and neural networks are the algorithms used in this thesis. For cluster analysis, the number of clusters is an important parameter because all subsequent analysis and results may change depending on the number of clusters set by the user. In order to address this issue, there are some internal and external measures to evaluate number of clusters (such as those introduced in Chapter 1). In this thesis, the number of clusters was found to be two in nearly all cases, but it is important to try different values in different climatic conditions (hot seasons vs. cold seasons) to make sure the applicability. Also, some more recent methods such as XMeans algorithm can use some splitting information (Bayesian Information Criterion) to automatically find the number of clusters.

Regarding association rule mining, the two parameters support and confidence were set at 40% and 85%, respectively. These value are considered to be reliable regarding building application and were also used in the literature [6], [11].

Regarding the ANN model, the details of the model can be found in Chapter 1. However, the objective of the methodologies developed in this thesis was to provide a framework for occupant behavior analysis, not optimizing the performance of ANNs. It is likely that some better performed ANN algorithms are introduced in recent years which perform better than MLP models used in this thesis. However, the size of the dataset is an important factor and should be considered while using the estimator models. It is recommended that different regressor models be applied and tested for the best results.

REFERENCES

[1] “International Energy Agency, Key World Energy Statistics,” 2006.

[2] “A review on buildings energy consumption information,” *Energy Build.*, vol. 40, no. 3, pp. 394–398, Jan. 2008.

[3] N. R. Canada, “Energy Efficiency Trends in Canada 1990 to 2013,” pp. 11–18, 2013.

[4] D. Bourgeois, “Detailed occupancy prediction, occupancy-sensing control and advanced behavioral modeling within whole-building energy simulation,” l’Universite Laval, Quebec, 2005.

[5] V. Dhar, “Data Science and Prediction,” *Commun. ACM*, vol. 56, no. 12, pp. 64–73, Dec. 2013.

[6] C. Fan, F. Xiao, and C. Yan, “A framework for knowledge discovery in massive building automation data and its application in building diagnostics,” *Autom. Constr.*, vol. 50, no. C, pp. 81–90, Feb. 2015.

[7] P. Waide, J. Ure, N. Karagianni, G. Smith, and B. Bordass, “The scope for energy and CO2 savings in the EU through the use of building automation technology,” *Final Rep. Eur. Copp. Inst.*, 2013.

[8] A. Capozzoli, D. Grassi, M. S. Piscitelli, and G. Serale, “Discovering knowledge from a residential building stock through data mining analysis for engineering sustainability,” *Energy Procedia*, vol. 83, pp. 370–379, 2015.

[9] I. Khan, A. Capozzoli, S. P. Corgnati, and T. Cerquitelli, “Fault Detection Analysis of

Building Energy Consumption Using Data Mining Techniques," *Energy Procedia*, vol. 42, pp. 557–566, Jan. 2013.

[10] A. Capozzoli, F. Lauro, and I. Khan, "Fault detection analysis using data mining techniques for a cluster of smart office buildings," *Expert Syst. Appl.*, vol. 42, no. 9, pp. 4324–4338, Jun. 2015.

[11] Z. Yu, B. C. M. Fung, F. Haghighat, H. Yoshino, and E. Morofsky, "A systematic procedure to study the influence of occupant behavior on building energy consumption," *Energy Build.*, vol. 43, no. 6, pp. 1409–1417, 2011.

[12] and J. P. J. Han, M. Kamber, *Data mining, concepts and techniques 3rd ed*, 3rd ed. Elsevier, 2012.

[13] D. J. Hand, P. Smyth, and H. Mannila, *Principles of Data Mining*. Cambridge, MA, USA: MIT Press, 2001.

[14] Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling," *Energy Build.*, vol. 42, no. 10, pp. 1637–1646, 2010.

[15] Z. Du and X. Jin, "Detection and diagnosis for sensor fault in HVAC systems," *Energy Convers. Manag.*, vol. 48, no. 3, pp. 693–702, Mar. 2007.

[16] Y. Hu, H. Chen, G. Li, H. Li, R. Xu, and J. Li, "A statistical training data cleaning strategy for the PCA-based chiller sensor fault detection, diagnosis and data reconstruction method," *Energy Build.*, vol. 112, pp. 270–278, Jan. 2016.

[17] S. Wang and F. Xiao, "AHU sensor fault diagnosis using principal component analysis method," *Energy Build.*, vol. 36, no. 2, pp. 147–160, Feb. 2004.

[18] X. Cipriano, A. Vellido, J. Cipriano, J. Martí-Herrero, and S. Danov, “Influencing factors in energy use of housing blocks: a new methodology, based on clustering and energy simulations, for decision making in energy refurbishment projects,” *Energy Effic.*, vol. 10, no. 2, pp. 359–382, Apr. 2017.

[19] E. Wang, “Benchmarking whole-building energy performance with multi-criteria technique for order preference by similarity to ideal solution using a selective objective-weighting approach,” *Appl. Energy*, vol. 146, pp. 92–103, 2015.

[20] A. Rakotomamonjy, “Variable Selection Using SVM-based Criteria,” *J. Mach. Learn. Res.*, vol. 3, pp. 1357–1370, 2003.

[21] J.-S. Chou, Y.-C. Hsu, and L.-T. Lin, “Smart meter monitoring and data mining techniques for predicting refrigeration system performance,” *Expert Syst. Appl.*, vol. 41, no. 5, pp. 2144–2156, 2014.

[22] S. Wang and F. Xiao, “Detection and diagnosis of AHU sensor faults using principal component analysis method,” *Energy Convers. Manag.*, vol. 45, no. 17, pp. 2667–2686, Oct. 2004.

[23] M. Peña, F. Biscarri, J. I. Guerrero, I. Monedero, and C. León, “Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach,” *Expert Syst. Appl.*, vol. 56, pp. 242–255, Sep. 2016.

[24] Z. Du, B. Fan, X. Jin, and J. Chi, “Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis,” *Build. Environ.*, vol. 73, pp. 1–11, Mar. 2014.

[25] J. M. Abreu, F. C. Pereira, and P. Ferrão, “Using pattern recognition to identify habitual behavior in residential electricity consumption,” *Energy Build.*, vol. 49, pp. 479–487, 2012.

[26] F. Xiao and C. Fan, “Data mining in building automation system for improving building operational performance,” *Energy Build.*, vol. 75, pp. 109–118, 2014.

[27] S. D’Oca and T. Hong, “A data-mining approach to discover patterns of window opening and closing behavior in offices,” *Build. Environ.*, vol. 82, pp. 726–739, 2014.

[28] C. M. R. do Carmo and T. H. Christensen, “Cluster analysis of residential heat load profiles and the role of technical and household characteristics,” *Energy Build.*, vol. 125, pp. 171–180, 2016.

[29] X. Liang, T. Hong, and G. Q. Shen, “Occupancy data analytics and prediction: a case study,” *Build. Environ.*, vol. 102, pp. 179–192, 2016.

[30] M. Saarikoski, “A data mining approach to indoor environment quality assessment: A study on five detached houses in Finland,” 2016.

[31] Z. Yu, B. C. M. Fung, and F. Haghighat, “Extracting knowledge from building-related data—A data mining framework,” in *Building Simulation*, 2013, vol. 6, no. 2, pp. 207–222.

[32] S. D’Oca, S. Corgnati, and T. Hong, “Data Mining of Occupant Behavior in Office Buildings,” *Energy Procedia*, vol. 78, pp. 585–590, 2015.

[33] M. Sameti and F. Haghighat, “Optimization of 4th generation distributed district heating system: Design and planning of combined heat and power,” *Renew. Energy*, vol. 130, pp. 371–387, Jan. 2019.

[34] T. Hong, D. Yan, S. D’Oca, and C. Chen, “Ten questions concerning occupant behavior in

buildings: The big picture," *Build. Environ.*, vol. 114, pp. 518–530, Mar. 2017.

[35] S. D'Oca and T. Hong, "Occupancy schedules learning process through a data mining framework," *Energy Build.*, vol. 88, pp. 395–408, 2015.

[36] K. Sun, D. Yan, T. Hong, and S. Guo, "Stochastic modeling of overtime occupancy and its application in building energy simulation and calibration," *Build. Environ.*, vol. 79, pp. 1–12, 2014.

[37] Ashrea, "'Energy Standard for Buildings except Low-RiseResidential Buildings, 90.1,'" 2004.

[38] A. Capozzoli, M. S. Piscitelli, A. Gorrino, I. Ballarini, and V. Corrado, "Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings," *Sustain. Cities Soc.*, vol. 35, pp. 191–208, Nov. 2017.

[39] Y. Wang and L. Shao, "Understanding occupancy pattern and improving building energy efficiency through Wi-Fi based indoor positioning," *Build. Environ.*, vol. 114, pp. 106–117, Mar. 2017.

[40] I. Kastner and E. Matthies, "Implementing web-based interventions to promote energy efficient behavior at organizations – a multi-level challenge," *J. Clean. Prod.*, vol. 62, pp. 89–97, Jan. 2014.

[41] A. Meinke, M. Hawighorst, A. Wagner, J. Trojan, and M. Schweiker, "Comfort-related feedforward information: occupants' choice of cooling strategy and perceived comfort," *Build. Res. Inf.*, vol. 45, no. 1–2, pp. 222–238, Feb. 2017.

[42] C. Fischer, "Feedback on household electricity consumption: a tool for saving energy?,"

Energy Effic., vol. 1, no. 1, pp. 79–104, Feb. 2008.

[43] Z. Yu, J. Li, H. Q. Li, J. Han, and G. Q. Zhang, “A Novel Methodology for Identifying Associations and Correlations Between Household Appliance Behaviour in Residential Buildings,” *Energy Procedia*, vol. 78, pp. 591–596, Nov. 2015.

[44] M. Mohanraj, S. Jayaraj, and C. Muraleedharan, “Applications of artificial neural networks for refrigeration, air-conditioning and heat pump systems—A review,” *Renew. Sustain. Energy Rev.*, vol. 16, no. 2, pp. 1340–1358, 2012.

[45] M. Y. Haller *et al.*, “Dynamic whole system testing of combined renewable heating systems–The current state of the art,” *Energy Build.*, vol. 66, pp. 667–677, 2013.

[46] J. M. Belman-Flores and S. Ledesma, “Statistical analysis of the energy performance of a refrigeration system working with R1234yf using artificial neural networks,” *Appl. Therm. Eng.*, vol. 82, pp. 8–17, May 2015.

[47] D. J. Swider, “A comparison of empirically based steady-state models for vapor-compression liquid chillers,” *Appl. Therm. Eng.*, vol. 23, no. 5, pp. 539–556, Apr. 2003.

[48] E. Mocanu, P. H. Nguyen, M. Gibescu, and W. L. Kling, “Deep learning for estimating building energy consumption,” *Sustain. Energy, Grids Networks*, vol. 6, pp. 91–99, Jun. 2016.

[49] T. Hong and H.-W. Lin, “Occupant behavior: impact on energy use of private offices,” in *ASim 2012 - 1st Asia conference of International Building Performance Simulation Association*, 2013.

[50] S. Murakami *et al.*, “Energy consumption for residential buildings in Japan,” *Archit. Inst.*

Japan, Maruz. Corp, 2006.

[51] M. Ashouri, F. Haghighat, B. C. M. Fung, A. Lazrak, and H. Yoshino, "Development of Building Energy Saving Advisory: A Data Mining Approach," *Energy Build.*, May 2018.

[52] C. Nguyen, "Computer-aided Nonlinear Analysis of Microwave and Millimeter Wave Amplifiers and Mixers," University of Central Florida, Orlando, FL, USA, 1991.

[53] L. Rokach and O. Maimon, "The Data Mining and Knowledge Discovery Handbook: A Complete Guide for Researchers and Practitioners." New York: Springer, 2005.

[54] Scikit-Learn, "Scikit-Learn Documentation." [Online]. Available: http://scikit-learn.org/stable/modules/clustering.html.

[55] "Python (3.5)." [Online]. Available: https://www.python.org/.

[56] S. M. C. Magalhães, V. M. S. Leal, and I. M. Horta, "Modelling the relationship between heating energy use and indoor temperatures in residential buildings through Artificial Neural Networks considering occupant behavior," *Energy Build.*, vol. 151, pp. 332–343, Sep. 2017.

[57] Y. Zhang, X. Bai, F. P. Mills, and J. C. V. Pezzey, "Rethinking the role of occupant behavior in building energy performance: A review," *Energy Build.*, vol. 172, pp. 279–294, Aug. 2018.

[58] S. Bhattacharjee and G. Reichard, "Socio-Economic Factors Affecting Individual Household Energy Consumption: A Systematic Review," *ASME 2011 5th Int. Conf. Energy Sustain.*, no. 54686, pp. 891–901, 2011.

[59] S. Walfish, "A Review of Statistical Outlier MethodsTitle," *Pharm. Technol.*, vol. 11, no.

30, pp. 82–88, 2006.

[60] C. Fu, J. Zheng, J. Zhao, and W. Xu, “Application of grey relational analysis for corrosion failure of oil tubes,” *Corros. Sci.*, vol. 43, no. 5, pp. 881–889, 2001.

[61] C. Fan, F. Xiao, Z. Li, and J. Wang, “Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review,” *Energy Build.*, vol. 159, pp. 296–308, Jan. 2018.

[62] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[63] Z. (Jerry) Yu, F. Haghighat, B. C. M. Fung, E. Morofsky, and H. Yoshino, “A methodology for identifying and improving occupant behavior in residential buildings,” *Energy*, vol. 36, no. 11, pp. 6596–6608, Nov. 2011.

[64] M. Ashouri, F. Haghighat, B. C. M. Fung, and H. Yoshino, “Development of a ranking procedure for energy performance evaluation of buildings based on occupant behavior,” *Energy Build.*, vol. 183, pp. 659–671, Jan. 2019.