Development of Building Energy Saving Advisory: A Data Mining Approach

Milad Ashouri[a], Fariborz Haghighat[a*], Benjamin C. M. Fung[b],

Amine Lazrak[c], Hiroshi Yoshino[d]

[a] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Quebec, Canada H3G 1M8

[b] School of Information Studies, McGill University, Montreal, Quebec, Canada H3A 1X1

[c] Fraunhofer Center for Sustainable Energy Systems, Boston, U.S.A.

[d] Department of Architecture and Building Science, Tohoku University, Japan

Abstract

Occupants' behavior and their interaction with home appliances are crucial for assessing building energy consumption. This study proposes a new methodology for monitoring the energy consumed in building end-use loads to build an advisory system. The built system alerts occupants to take certain measures (prioritized recommendations) to reduce energy consumption of end-use loads. The quantification of potential savings is also provided upon following said measures. The proposed methodology is also capable of evaluating the energy savings performed by the occupants. The system works based on the analysis of historical data generated by occupants using data mining techniques to output highly feasible recommendations. For demonstration purposes, the methodology was tested on the real dataset of a building in Japan. The dataset includes detailed energy consumption of end-use loads, categorized as hot water supply, lighting, kitchen, refrigerator, entertainment & information, housework & sanitary, and others. Results suggest that the developed models are accurate, and that it is possible to save up to 21% of total energy consumption by only changing occupants' energy use habits.

Key words: Occupant behavior, data mining, building energy

Corresponding author: haghi@bcee.concordia.ca

1. Introduction

1.1 Background

Buildings' impacts on global energy consumption have been increasing steadily, reaching up to 20-40% in developed countries [1]. According to Natural Resources Canada [2], 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 [3], 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 [4]. 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 [5]. It is the only way to reduce energy consumption without costly fundamental changes such as upgrading building systems, reconstruction, envelope renovation etc.

1.2 Previous Work

Some reviews of unsupervised data analytics exist in literature [6, 7]. The key studies lie in finding the patterns of operation. Cluster analysis is an efficient tool for finding patterns by grouping data into *k* separate subsets. Fan et al. [8] found the typical operation patterns of a building cooling system by clustering the annual energy consumption data. The results suggested two power consumption patterns: weekends and weekdays. A similar project revealed three patterns using the same technique [9]. D'Oca et al. [10] reported patterns in window opening and closing as an

important factor in the energy consumption of sixteen office buildings using regression analysis and clustering.

These patterns are interesting to incorporate into simulation packages to produce more accurate load calculations. For example, patterns of window opening duration during the day (long, medium or short intervals) affect infiltration and consequently energy consumption. In addition to finding patterns, clustering helps isolate the effect of influential factors in energy consumption. For instance, Do Carmo et al. [11] clustered the hourly heat load data of 139 single family detached houses into three separate groups. The goal was to eliminate the effect of weather conditions to show the effect of household and building characteristics on thermal load demand. Pattern discovery also determines daily routines (consumption patterns) for each household, which can be compared to the best, worst or "normal" scenarios to change habits. Abreu et al. [12] developed a framework to extract the daily households routines and annual patterns of energy consumption. They identified recurrent behaviors during the day (daily routines of households) by applying the principal component analysis (PCA) to the daily energy consumption data. They 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 [13], energy consumption of domestic appliances [14], and indoor air quality [15].

Clustering and other techniques have also been used to rate building energy performance. Wang et al. [16] applied a multi-criteria benchmarking method to rank 324 single family dwellings according to their energy performance indicator. TOPSIS[1] was used to rank building energy

[1] Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is a multi-criteria decision analysis method.

performance. Over three years, more than half of the buildings had a TOPSIS score beyond 0.50 showing an efficient performance. Clustering (K-Means) was applied to TOPSIS results to rate buildings from “unsatisfactory” to “excellent”. This provided a framework for studying energy performance of buildings, and offered a rating system to provide instructions to tenants and building owners.

Data mining (DM) is also helpful for finding associations between building variables. Sometimes these associations may reveal hidden and useful knowledge that cannot be discovered through simulation programs. For instance, Yu et al. [14] applied association rule mining to measure energy consumption of domestic appliances. Association rule mining measured highly correlated activities (such as TV [High]→ Ventilator [High]), which may be unexpected and thus useful to give tenants energy saving recommendations. However, the extracted rules are qualitative (low and high). Fan et al. [8] applied quantitative association rule mining on a building cooling system to reveal associations between its components. D'Oca et al. [10] found associations between window opening and closing behaviors. Clustering and association rule mining can be used together to develop a step-by-step framework [8-10, 14, 17, 18], a process that usually involves data pre-processing, pattern discovery, knowledge discovery, knowledge interpretation and selection.

Currently, building simulation software can only incorporate pre-defined occupant behavior such as scheduled occupancy presence, appliance use, etc. Data mining can address this problem. Yu et al. [14] proposed a framework to isolate the effect of occupants from other factors of building energy consumption to reduce overall energy consumption. D’Oca et al. [10, 19] measured window opening and closing patterns in sixteen office buildings, identifying the most influential factors, which produce these patterns using a logistic regression. This information can be used in

simulation tools as well. Occupancy presence and scheduling (occupants/hour) is also modeled using data mining methods. Predicting occupant presence could affect building energy modeling (Liang et al. [13], D'oca et al. [19, 20], and Sun et al. [21]). Cluster analysis and decision tree were used by Liang et al. [13] to improve occupancy schedule accuracy. They showed that although ASHRAE standard of 90.1 [22] provides an acceptable general estimation of occupancy pattern, it is not applicable to any specific building. Sun et al. [21] reported that working overtime (presence of occupants in an office after regular working hours) can directly influence the state of HVAC equipment and therefore cooling energy consumption. They presented a stochastic model for overtime during weekdays based on the measured occupancy data from an office building. Results indicated that occupants' number and presence duration followed binomial and exponential distributions, respectively.

1.3 Statement of Novelty

The contribution of this study lies in developing a new framework for data analytics in building engineering by focusing specifically on occupants' role in energy consumption. The research problem outlined in this paper originates from the fact that occupants may not be cautious while using the appliances (such as forgetting to switch off the lights, letting the windows open, and similar examples) which cause waste of energy. To be more specific, the correlation between occupant activities and their influence on building energy consumption are determined and recommendations are given to tenants to keep the consumption patterns in normal level. To the author's best knowledge, by this time there has been no tool to monitor the occupant's energy awareness and alert them in cases of extra energy use. The proposed framework can help building owners or managers monitor building energy performance, and estimate potential savings to equip them with profitable measures. The proposed data-mining framework involves a series of

consecutive steps applied to the data. Each methodology has been applied to various datasets in building engineering, but their new integration can yield pioneer insights into data. The main objectives of the present work are:

- Developing a methodology to advise occupants how to reduce their end-use loads energy consumption and prioritize the recommendations.
- Providing an accurate quantitative report for tenants regarding potential energy savings if recommendations are respected.
- Giving occupants a precise report of achieved savings to persuade them to take more measures.

2. Methodology

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 1 shows the overall approach. Three data mining tasks (clustering, 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.

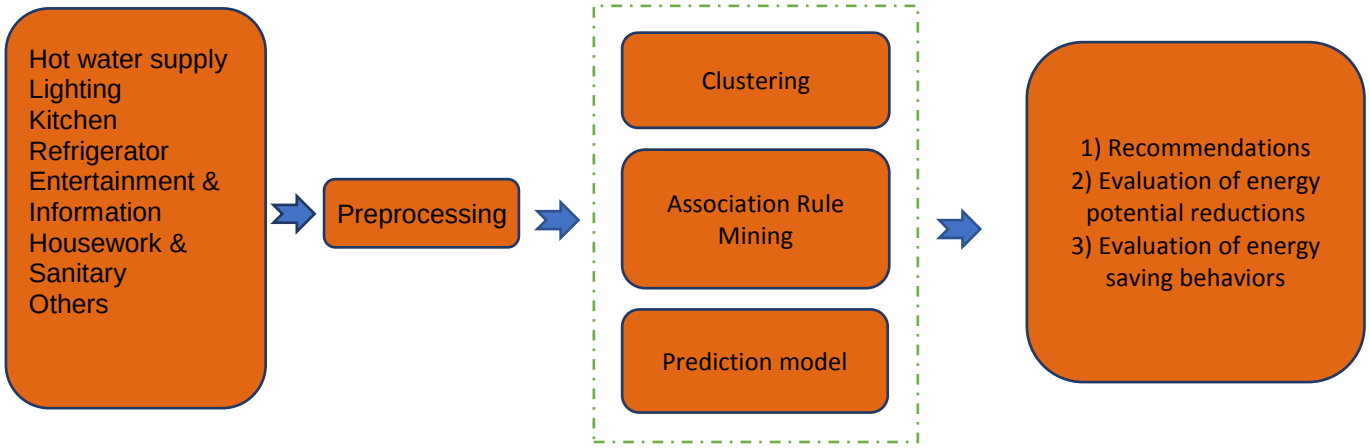

Figure 1. The proposed methodology for the recommender system. Inputs are detailed energy consumption of end-use loads and weather data. The outputs are prioritized recommendations and quantification of achieved energy savings along with the potential savings upon following the recommendations.

Reducing the effect of weather, energy consumption data is grouped by weather conditions and each cluster is analyzed separately. Figure 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 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). 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.

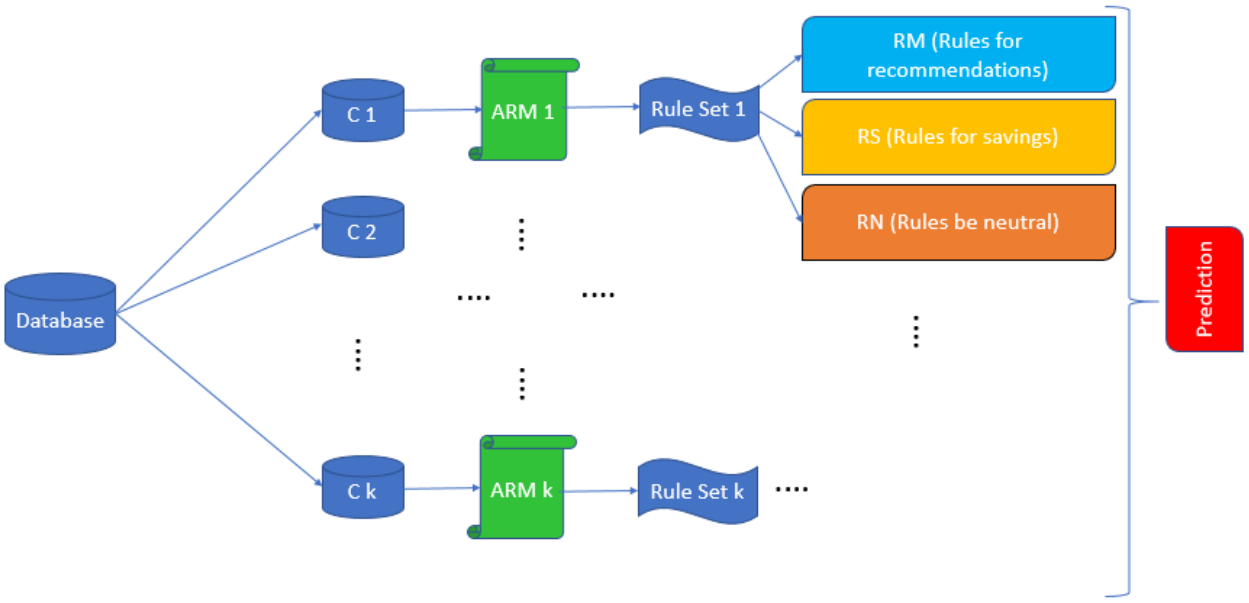

Figure 2. The data mining framework for knowledge discovery of occupants. The framework includes clustering, association rule mining, and neural networks to give the occupants useful information about their behavior in energy consumption of end-use services.

Figure 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).

2.2 Data Preprocessing

The data was extracted from a project entitled "Investigation on Energy Consumption of Residents All over Japan" [23]. 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 [23]. 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 [14, 24]. Table 1 shows the survey items corresponding to investigation methods, and time intervals. More information on measurement methods are found in [14].

Table 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 [8], lower and upper quantile (Q) [25], complete case analysis [26], simple moving average method [8, 9, 27] 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 [28], Min-max normalization was performed on the data as follows:

$$x' = \frac{\mathrm{x} - x_{min}}{x_{max} - x_{min}} (x'_{max} - x'_{min}) + x'_{min} \quad (1)$$

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

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 [29, 30]:

$$obj = \sum_{i=0}^{n} \min_{\mu_j} \|x_i - \mu_j\| \quad (2)$$

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.) [31]. 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 [30]. 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:

$$s = \frac{b - a}{\max(a, b)} \tag{3}$$

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.

2.4 Association Rule Mining (ARM)

Association rule mining is the next step of process and is applied on each cluster separately (see Figure 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. 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 [25].

Mathematically, support and threshold are defined as follows:

$$Support\ (X \rightarrow Y) = P(X \cup Y) \tag{4}$$

$$Confidence\ (X \rightarrow Y) = P(Y|X) \tag{5}$$

$$Lift\ (X \rightarrow Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = \frac{P(Y|X)}{P(Y)} \tag{6}$$

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 [25]. 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 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). Figure 4 shows the association rule extraction process.

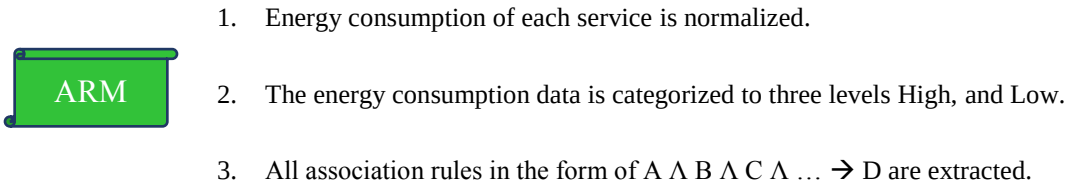

Figure 4. Association rule mining on the energy consumption data of end-use loads to find the correlations between occupants' various activities.Table 2 shows a selection of the result of this process. All attribute names are listed. Further discussion will be presented in section 3.2.

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.

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 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.

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:

$$\%\, Savings = \frac{sum(b) - sum(a)}{sum(b)} \tag{7}$$

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. Results and Discussion

As mentioned in Figure 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.

3.1 Clustering Results

The open source software Python [32] 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 5 and Figure 6 for results).

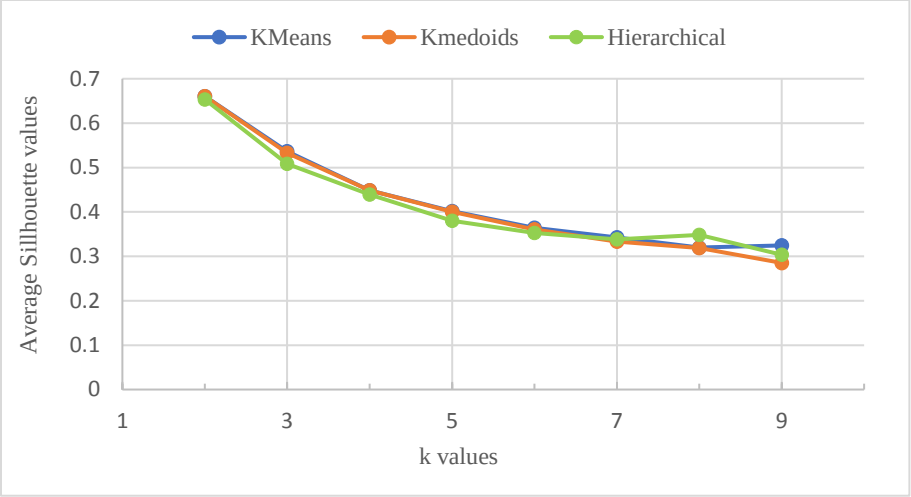

Figure 5. Clustering validation results with three different clustering algorithms and silhouette index as the evaluator. Three clusters roughly give the same silhouette index. The optimum number of clusters is two.

Figure 5 and Figure 6 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.

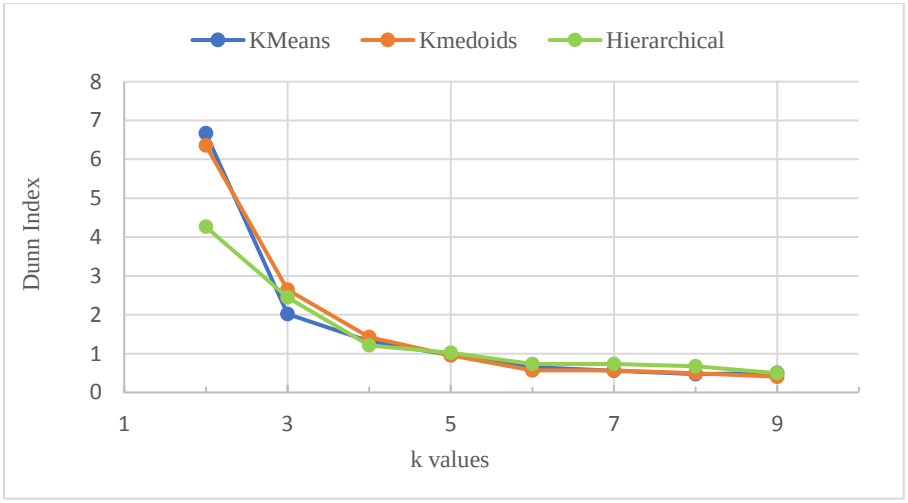

Figure 6. Clustering validation results with three different clustering algorithms and Dunn index as the evaluator. K-Means works best when number of clusters is two (optimum), but with increasing k the three algorithms show similar performance.

Figure 7 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 8 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 9. 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.

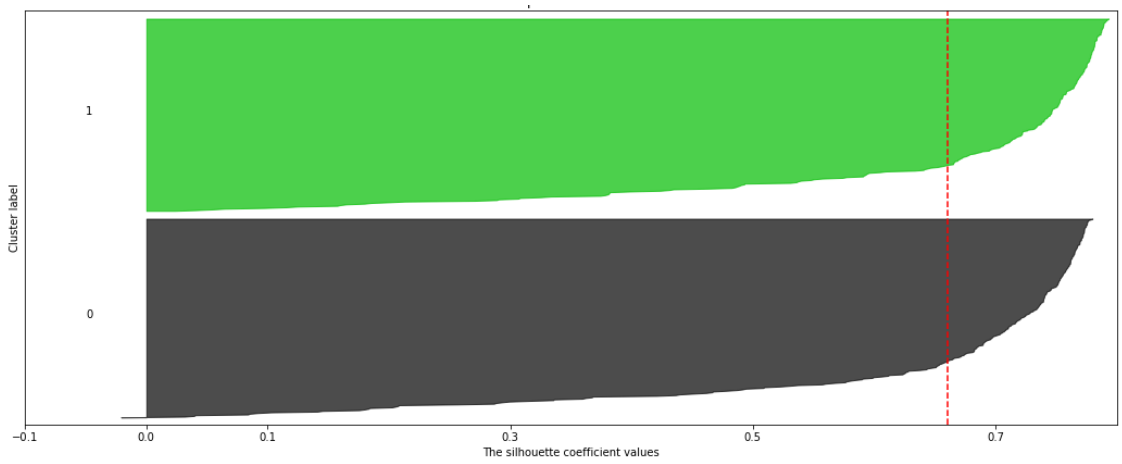

Figure 7. The silhouette plot for two clusters. Number of clusters is 2. The number of data points (thickness) is roughly the same in each cluster. The average silhouette value is 0.533 (the dotted red line).

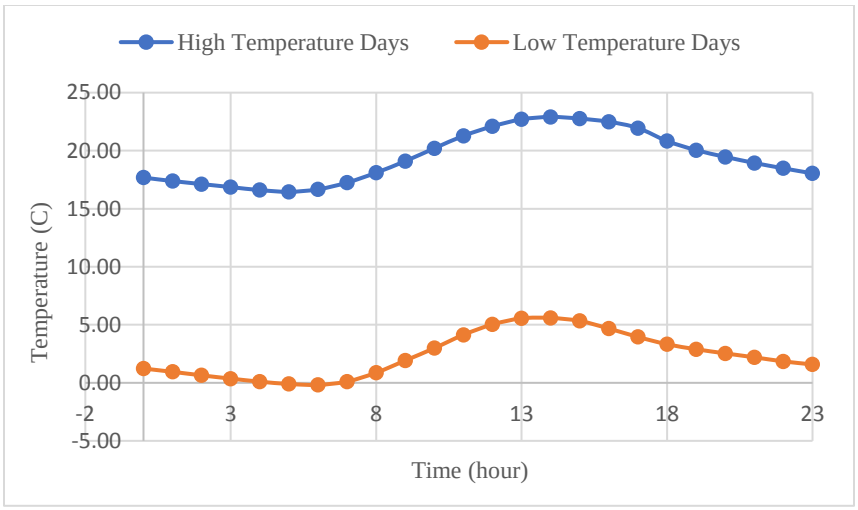

Figure 8. Clustering results (centroids) using K-Means algorithm and k=2. Attributes are the 24 hours outdoor temperatures. The results clearly show that the clusters are externally separated (High temperature and low temperature profiles). The study of occupant behavior can be performed on each cluster separately, so the days are similar in terms of outside temperature which affects the occupant behavior.

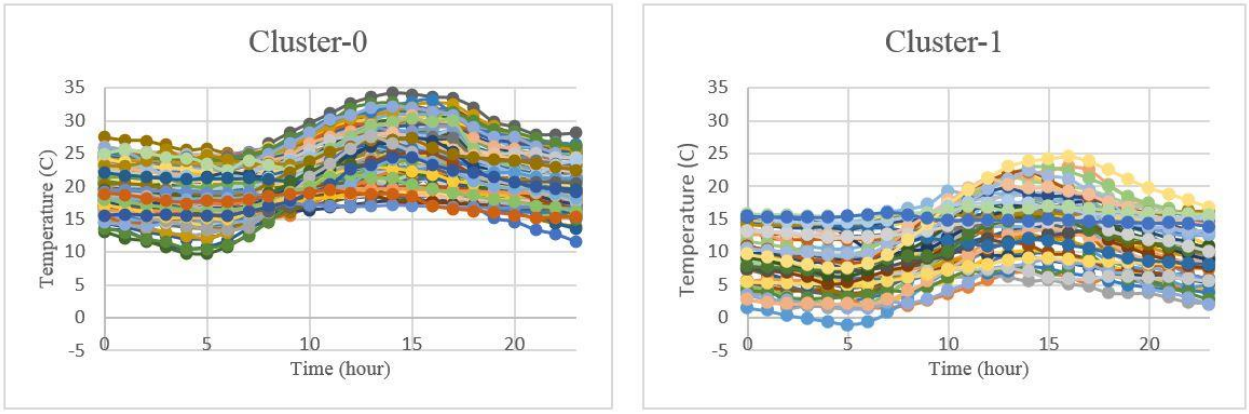

Figure 9. Clustering results for a sample data set using K-means algorithm and k=2. Attributes are the 24 hours outdoor temperatures. The results clearly show that the clusters are externally separated.

3.2 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 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 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 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 [33]), the rule was rejected. To demonstrate this process, two rules (one from RM and one from RS) were selected and shown in Table 3:

Table 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 10 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.

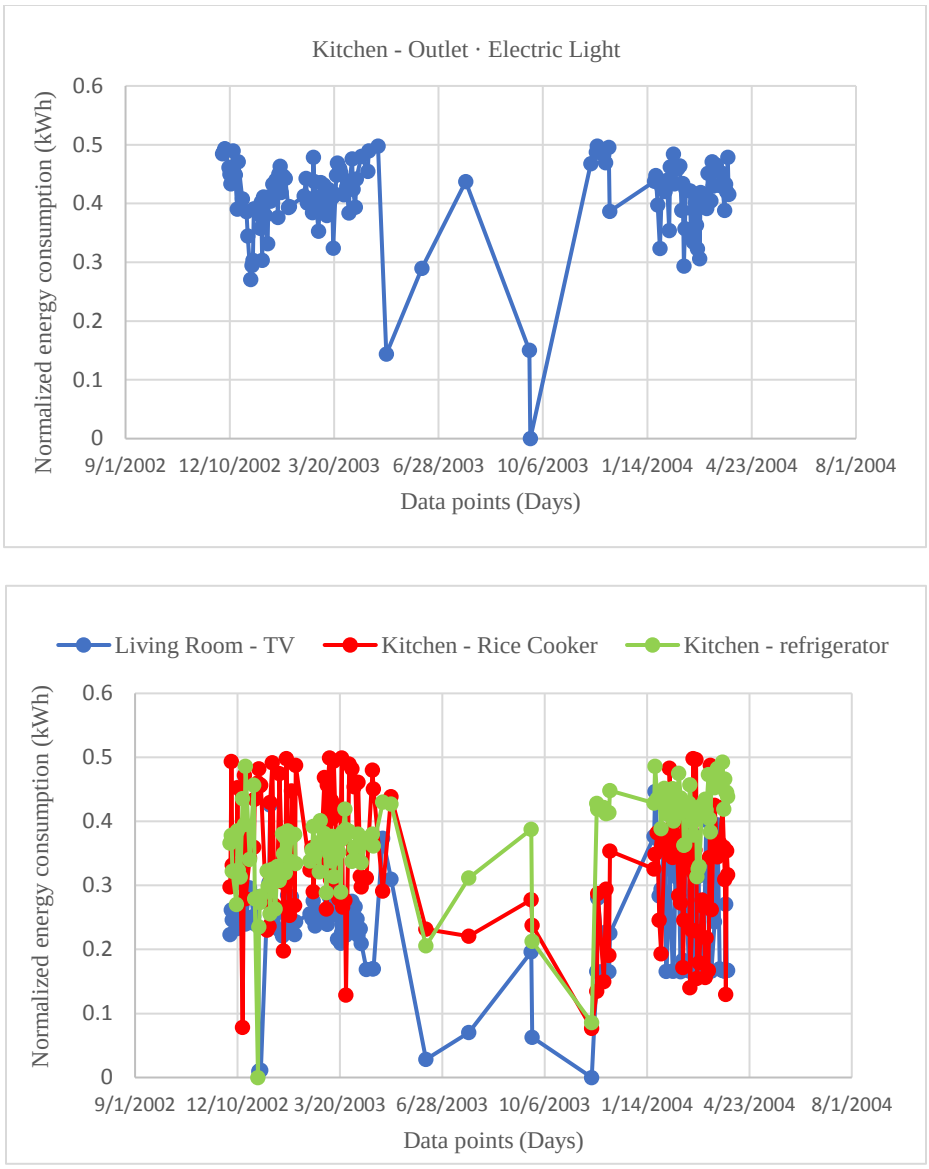

Figure 10. Normalized energy consumption of end-use loads based on rule 1. The upper and lower figures show the inputs and output of the ANN model. The data points are extracted from the dataset and are not successive days. Also, the days come from the same cluster.

Figure 11 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 12 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.

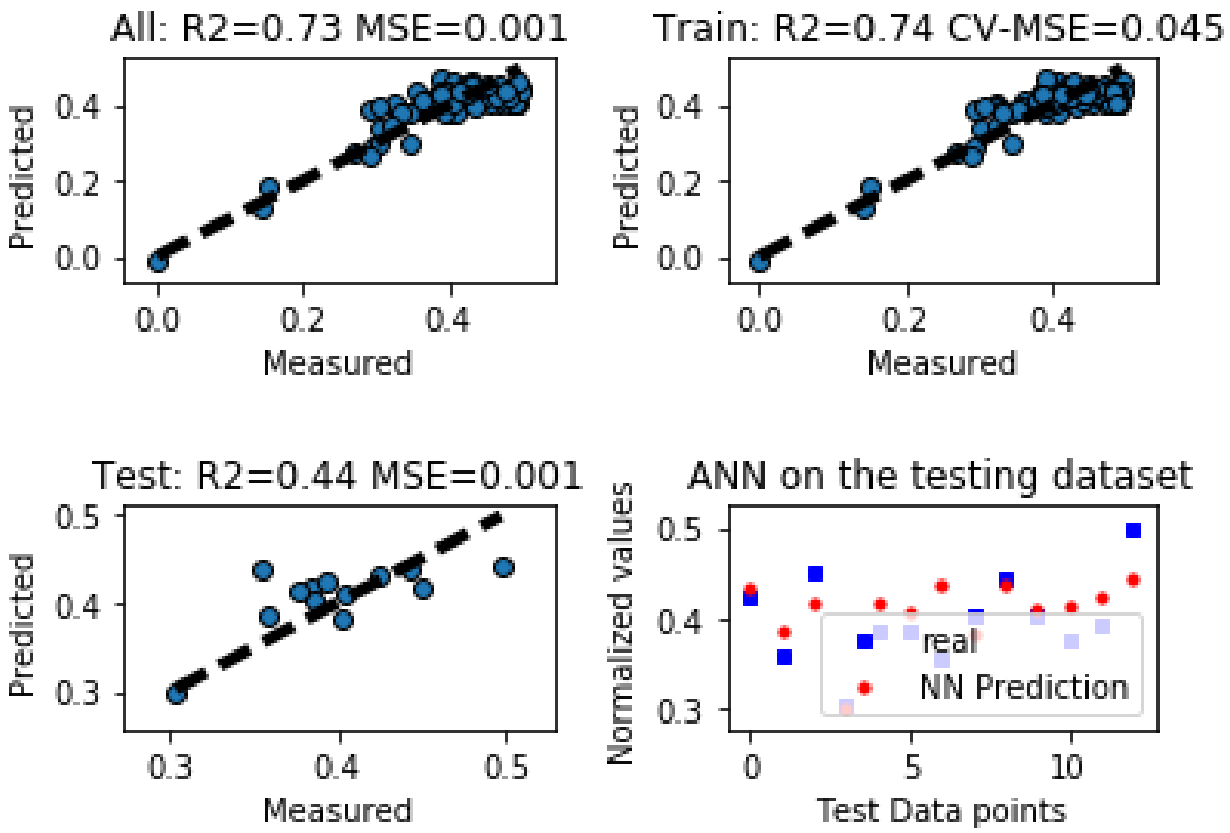

Figure 11. Regression fit of the real and modeled output. It is seen that the network is able to model the energy consumption of output (kitchen lamp in this case) using other end-use loads as inputs (living room TV, Rice cooker and refrigerator in this case). The cross-validation error and test error are 0.04 and 0.001 which are in acceptable range.

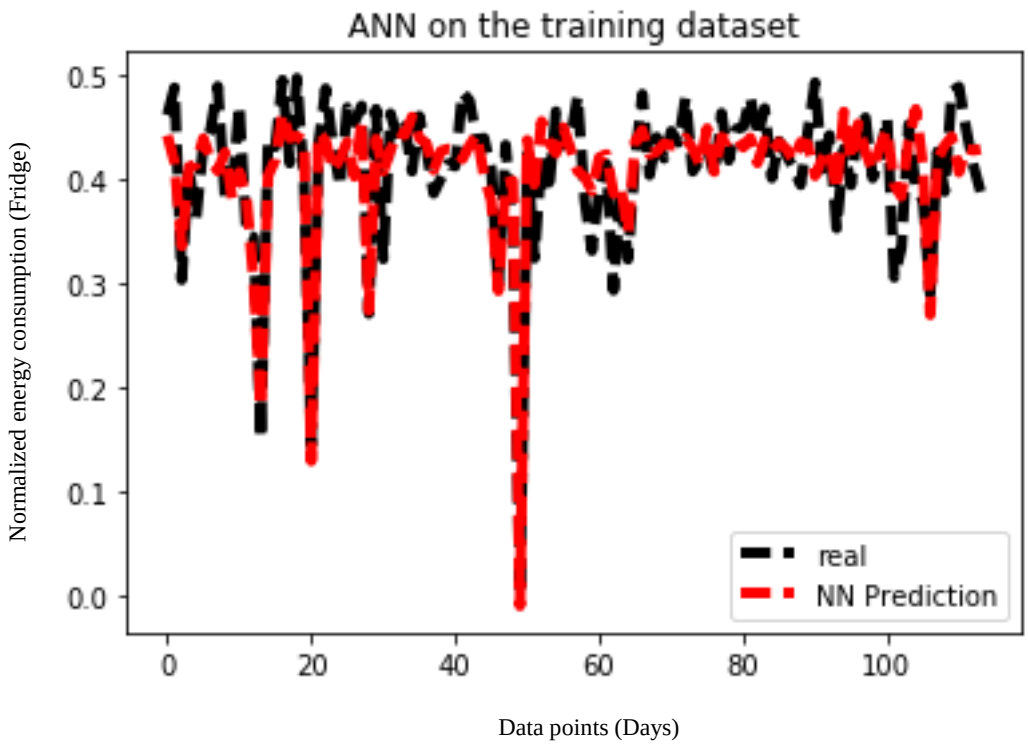

Figure 12. Comparison of real and network generated values. The model can follow the trend of the real dataset. Mean squared error is 0.001.

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 13). Using equation 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

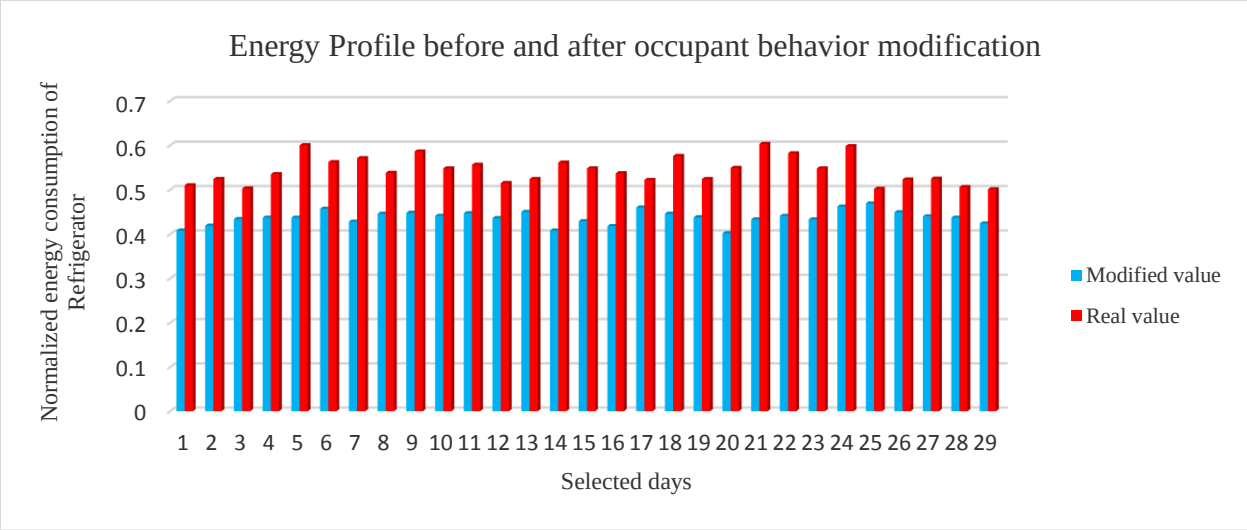

Figure 13. Applying the model on the new data set to estimate the potential savings upon following the recommendations. The cumulative sum of difference shows the potential savings. The cumulative savings are up to 19% which shows a good opportunity to save energy.

online (current data is fed to the system and inspected), when abnormal behavior 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 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 14 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.

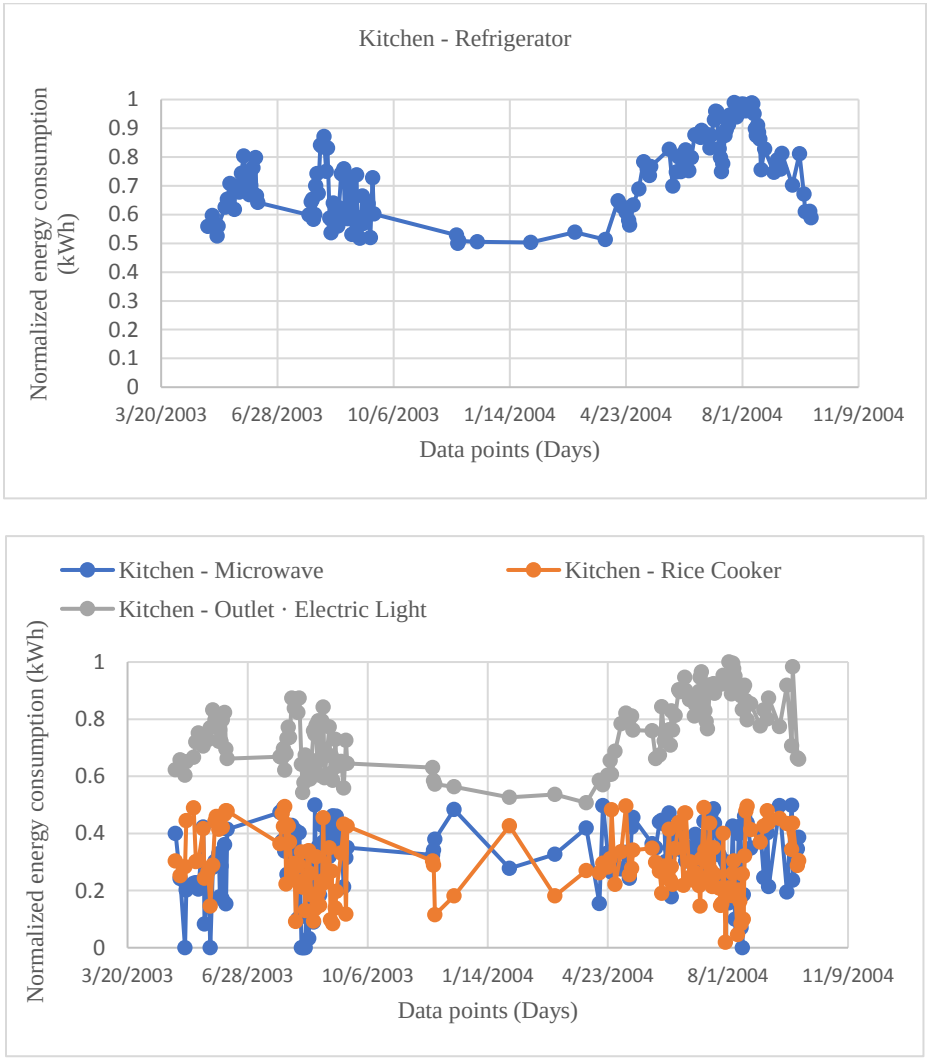

Figure 14. Relationships between inputs and the output based on associations. The inputs are shown in the upper figure and the output in the lower part. There is a complicated relation between various end-use loads which are modeled using a neural network model.

Figure 15 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).

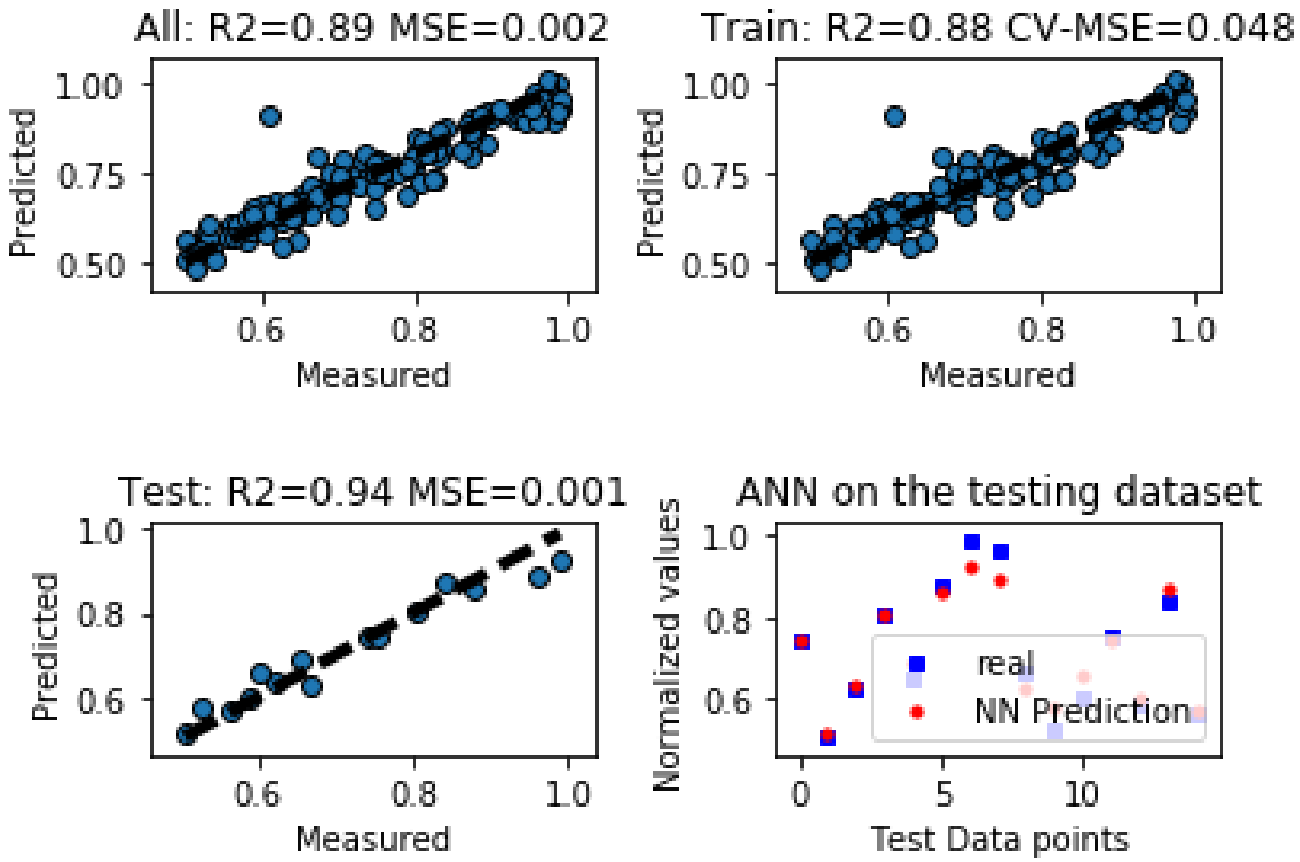

Figure 15. Regression fit of the real modeled output. It is seen that the network is able to model the energy consumption of refrigerator using other end-use loads Microwave, kitchen lamp, and Rice cooker.

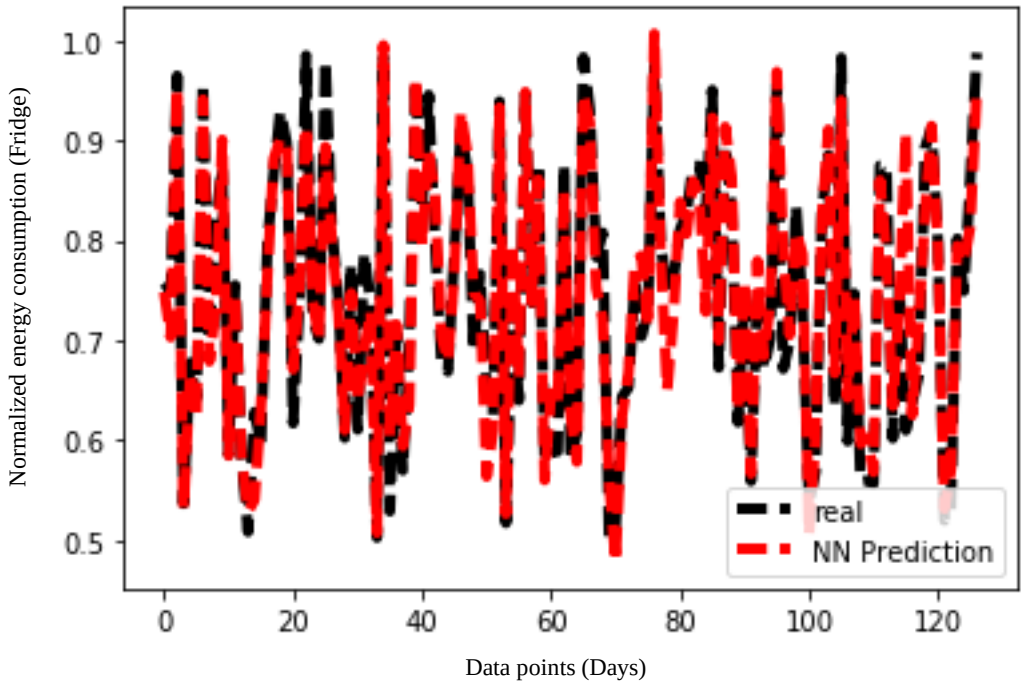

Figure 16. Result of training an MLP neural network on the obtained rule. The cross-validation error is 0.04.

Figure 16 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%.

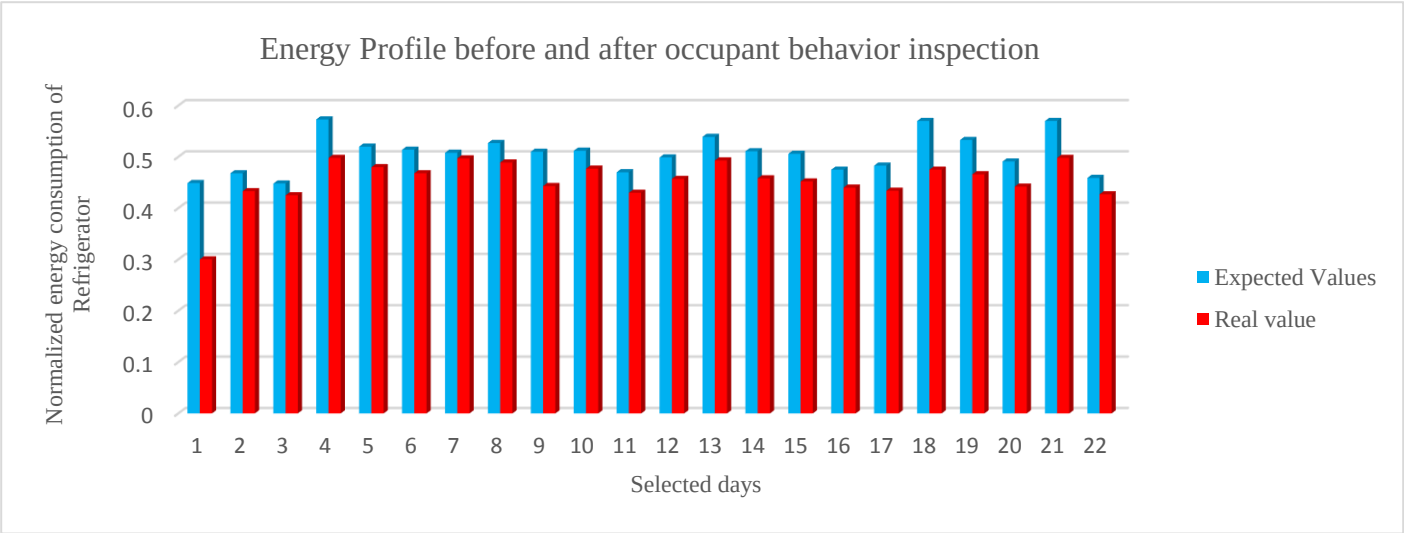

Figure 17. Applying the model on the new dataset to estimate the achieved savings by the occupants. The cumulative sum of difference shows the improvements. The cumulative savings are up to 10%.

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 17 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 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 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%

4. Research Limitations and Future Work

Although the developed methodology can be applied to any building and in any climatic region, there are some difficulties to overcome. Having no hourly data for end-use loads is an important challenge that makes associating rule mining processes less realistic given the services might not be operating simultaneously. Consequently, recommendations are less accurate. Having 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 true.

Another challenge originates from data deficiency. For instance, the number of occupants is present in the dataset, but their age is not. Knowing their age would help us estimate the level of their activity or provide more detailed recommendations based on which activities or behaviors children or adults waste most in building energy. The most useful information would be knowing occupants' daily schedule, preferences (e.g. lighting level, room temperature), and holidays.

The size of the dataset is also a problem for most data mining techniques. Usually, increasing the number of data points will improve the accuracy of machine learning algorithms. Some, such as

deep learning (which cannot be used in this case), do not work efficiently. The data set available for this study covered 2003 and 2004, and part of 2002 and 2005. Higher resolution data would make data mining results more reliable.

The obtained rules in this study originate from the measured energy consumption of building occupants. Energy reduction recommendations are therefore based on them. However, the occupants of a single building may not be aware of energy savings. If residents use high energy rates almost all the time, good behaviors are not detected in historical data, thus no RM rule is found. By domain knowledge, it is possible to introduce some artificial rules (similar to real rules) which may help find energy inefficient behaviors. For example, at noon (while tenants are at work) and the lights are off, any appliance such as TV, Boom Box, kitchen appliance and PC should be turn off. The idea could be extended to windows state (open/closed) as well.

Another approach would be developing a procedure to rank buildings in terms of occupants' energetic performance to demonstrate each building's rank relative to others and tell tenants which end-use loads require more attention. This way, they will be aware of their actual rank and be motivated to take energy saving measures. However, because several factors impact energy consumption patterns (such as the number of occupants, floor area, house type, etc.), simply comparing the amount of energy consumption of several buildings is not sufficient. The measures occupants have taken to save energy and their level of awareness are not accounted for. Thus, such comparisons are not yet achieved. One must consider as many influencing factors as possible to make a fair comparison. If the procedure is properly designed, it can reveal potential saving opportunities for tenants. This idea lays the foundation for our future work which is developing a methodology for performance comparison of several buildings via a data mining framework.

Conclusion

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.

Acknowledgements

Authors would like to express their gratitude to the Concordia University for supporting this research through *Concordia Research Chair in Energy and Environment*.

References

[1] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy and Buildings,* vol. 40, no. 3, pp. 394-398, 2008/01/01/ 2008.

[2] N. R. Canada, "Energy Efficiency Trends in Canada 1990 to 2009, Cat. No. M141–1/2009E-PDF (Online)," 2011, Available: HTTP://OEE.NRCAN.

[3] D. Bourgeois, "Detailed occupancy prediction, occupancy-sensing control and advanced behavioral modeling within whole-building energy simulation," Ph.D. Thesis, l'Universite Laval, Quebec, 2005.

[4] V. Dhar, "Data Science and Prediction," *Communications of the ACM,* vol. Vol. 56, no. No. 12, pp. 64-73, December 2013.

[5] T. Hong and H.-W. Lin, "Occupant behavior: impact on energy use of private offices," Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US)2013.

[6] 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 and Buildings,* 2017/11/09/ 2017.

[7] C. Miller, Z. Nagy, and A. Schlueter, "A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings," *Renewable and Sustainable Energy Reviews,* vol. 81, no. Part 1, pp. 1365-1377, 2018/01/01/ 2018.

[8] C. Fan, F. Xiao, and C. Yan, "A framework for knowledge discovery in massive building automation data and its application in building diagnostics," *Automation in Construction,* vol. 50, pp. 81-90, 2015/02/01/ 2015.

[9] F. Xiao and C. Fan, "Data mining in building automation system for improving building operational performance," *Energy and Buildings,* vol. 75, pp. 109-118, 2014/06/01/ 2014.

[10] S. D'Oca and T. Hong, "A data-mining approach to discover patterns of window opening and closing behavior in offices," *Building and Environment,* vol. 82, pp. 726-739, 2014/12/01/ 2014.

[11] 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 and Buildings,* vol. 125, pp. 171-180, 8/1/ 2016.

[12] J. M. Abreu, F. Câmara Pereira, and P. Ferrão, "Using pattern recognition to identify habitual behavior in residential electricity consumption," *Energy and Buildings,* vol. 49, pp. 479-487, 6// 2012.

[13] X. Liang, T. Hong, and G. Q. Shen, "Occupancy data analytics and prediction: A case study," *Building and Environment,* vol. 102, pp. 179-192, 6// 2016.

[14] 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 and Buildings,* vol. 43, no. 6, pp. 1409-1417, 2011/06/01/ 2011.

[15] M. Saarikoski, "A data mining approach to indoor environment quality assessment, a study on five detached houses in Finland," MSc Thesis, Environmental Science,

Department of Environmental and Biological Sciences, University of Eastern Finland, March 2016.

[16] 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," *Applied Energy,* vol. 146, pp. 92-103, 5/15/ 2015.

[17] 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, 2015/12/01 2015.

[18] Z. Yu, B. C. M. Fung, and F. Haghighat, "Extracting knowledge from building-related data — A data mining framework," *Building Simulation,* journal article vol. 6, no. 2, pp. 207-222, 2013.

[19] S. D'Oca, S. Corgnati, and T. Hong, "Data Mining of Occupant Behavior in Office Buildings," *Energy Procedia,* vol. 78, pp. 585-590, 11// 2015.

[20] S. D'Oca and T. Hong, "Occupancy schedules learning process through a data mining framework," *Energy and Buildings,* vol. 88, pp. 395-408, 2/1/ 2015.

[21] K. Sun, D. Yan, T. Hong, and S. Guo, "Stochastic modeling of overtime occupancy and its application in building energy simulation and calibration," *Building and Environment,* vol. 79, pp. 1-12, 9// 2014.

[22] ASHRAE, "Energy Standard for Buildings except Low-RiseResidential Buildings, 90.1," 2004.

[23] S. Murakami, Akabayashi, S., Inoue, T., Yoshino, H., Hasegawa, K., Yuasa, K., Ikaga, "Energy consumption for residential buildings in Japan, Architectural Institute of Japan, Maruzen Corp," 2006.

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

[25] M. K. Jiawei Han and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed ed. Elsevier, 2012.

[26] J. H. Andrew Gelman, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 1st ed. Cambridge University Press.

[27] R. T. T Hastie , J Friedman *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer Series in Statistics, Springer, New York, USA, 2008.

[28] M. K. Jiawei Han, Jian Pei, *Data Mining Concepts and Techniques*, 3rd ed. Elsevier, 2012.

[29] D. Arthur, and Sergei Vassilvitskii, "k-means++: The advantages of careful seeding," *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics* 2007.

[30] S.-l. documentation. Available: http://scikit-learn.org/stable/modules/clustering.html

[31] L. R. Oded Maimon, *The Data Mining and Knowledge Discovery Handbook* Springer 2005.

[32] *Python (3.5)*. Available: https://www.python.org/

[33] 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 and Buildings,* vol. 151, no. Supplement C, pp. 332-343, 2017/09/15/ 2017.