Energy & Buildings 183 (2019) 659–671

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild

Development of a ranking procedure for energy performance evaluation of buildings based on occupant behavior

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

[a] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Quebec, H3G 1M8 Canada
[b] School of Information Studies, McGill University, Montreal, Quebec, H3A 1X1 Canada
[c] Department of Architecture and Building Science, Tohoku University, Japan

ARTICLE INFO

Article history:
Received 2 October 2018
Revised 25 November 2018
Accepted 27 November 2018
Available online 1 December 2018

Keywords:
Building energy management
Data mining
Occupant behavior

ABSTRACT

Identifying the impacts of occupants on building energy consumption has become an important issue in recent years. This is due to the interrelationship of influencing factors such as urban climate, building characteristics, occupant behavior, and building services and operation, which makes it challenging to identify the role of occupants in energy consumption. The research problem in this study lies in the fact that the occupants of a building may not be cautious regarding energy savings, and there exists no ground to assess their energy consumption behavior. One solution is the development of a systematic comparison procedure between similar buildings. This paper introduces a new procedure for comparison between occupants of several buildings to show the rank of each building among others and suggest occupants on reducing their energy consumption and improving their rank. The proposed framework is developed based on multiple data-mining methods, including clustering, association rules mining, and neural networks. The proposed methodology is composed of two levels. The first considers the amount of energy usage by occupants after filtering effects unrelated to the occupant behavior. The second ranks the buildings in terms of achieved and potential savings during the time under investigation. To demonstrate the application, the methodology was applied on a set of monitored residential buildings in Japan. Results suggest that the proposed method enhances the evaluation of buildings' energy-saving potential by revealing the occupants' contribution. It also provides diverse and prioritized strategies to help occupants manage their energy consumption by revealing the building energy end-use patterns.

1. Introduction

1.1. Background

According to reports published by the Natural Resources Canada [1], 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 [2]. Recently, there have been many improvements in technological solutions, such as design and operation of building services [3]. Among these, recent research highlights occupant behavior as an important contributor that can increase the energy efficiency of buildings, similar to technological solutions [4].

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

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 [5]. 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

List of abbreviations: BA, building area; E&I, entertainment and information; ELA, equivalent leakage area; EUI, energy use intensity; FRIDGE, refrigerator; GRA, grey relational analysis; H&S, housework and sanitary; HC, space heating and cooling; HLC, heat loss coefficient; HT, house type; HVAC, heating, ventilation, and air conditioning; HWS, hot water supply; KE, kitchen equipment; KITCH, kitchen; LIGHT, lighting; NO, number of occupants; OTHER, other end-use loads; PI, performance index; RA, annual mean global solar radiation; RH, annual mean relative humidity; RM, rules for modification; RS, rules for savings; T, annual mean air temperature; WS, annual mean wind speed.

[*] Corresponding author.
E-mail address: haghi@bcee.concordia.ca (F. Haghighat).

https://doi.org/10.1016/j.enbuild.2018.11.050
0378-7788/© 2018 Elsevier B.V. All rights reserved.

Table 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)

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 can 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?”

1.2. Literature review

Data-mining processes usually involve data preprocessing, knowledge discovery, and interpretation and selection. Data preprocessing involves data cleaning, outlier detection, attribute selection, dimensionality reduction, and transformation. Knowledge discovery aims to extract raw information and discover patterns in data using either supervised or unsupervised data analytics. Interpretation and selection constitute the final step, referred to as post-mining. It basically means to interpret the results and screen for potentially useful knowledge. Fan et al. [3] and Miller et al. [7] carried out a review of the application of data analytics in building engineering. Different data-mining techniques have been used by researchers to extract information from building-related data. Techniques of supervised and unsupervised learning were used in the framework development [8–13]. Generally, works are divided into identifying operational patterns, predictive modeling, and fault detection and diagnosis. Since these concepts are highly tied together, works often involve some or all of them.

Operational patterns are found using data analytics tools. One of them is conventional clustering tools, such as K-means, K-medoids, DB-scan, etc., which have been used extensively [14,15]. The observations assigned to the same group are referred to as having similar operational conditions. As an example, clustering analysis has been used to study the patterns of HVAC operation [11,13] and windows opening and closing [10]. Carmo et al. [16] clustered hourly heat load data of 139 single-family detached houses to find daily routines of thermal load demand and the effect of household and building characteristics on energy use. Usually, typical operation patterns are investigated at the building, system, or component level. At the building level, whole-building energy consumption and environment have been considered [15,17–19]. At the system and component levels, the operation of HVAC has been investigated as a whole system or as a component (pumps, fans, etc.) [20–22].

Patterns of operation have also been discovered in the form of if–then rules. They are useful for identifying the typical working conditions of the system and discovering faulty conditions. Xue et al. [23] applied clustering analysis and association rule mining to find typical seasonal operating patterns in a district heating system. Any deviations from normal patterns show a fault in the system. Similar works in variable refrigerant flow systems and chiller systems have been carried out [13,24]. If the analysis of building-related data is performed in a time-series manner such as a daily, seasonal, or yearly basis to capture any underlying patterns in time, the knowledge discovered is referred to as temporal knowledge [6]. Lavin and Klabjan [14] clustered 24-h time–series energy data of commercial and industrial buildings to find patterns of operation over time. The results revealed different patterns of usage in different months of the year. Fan et al. [25] applied motif discovery (frequent sequential patterns) and temporal association rule mining (conclusions are made after a time interval) to discover temporal patterns in chiller and air handling unit energy consumption. The knowledge discovered was successfully used to identify anomalies in an HVAC system and capture building dynamics. Patnaik et al. [26] performed techniques of motif discovery to assess chiller operation in data centers. A similar approach was applied by Miller et al. [27] for whole-building energy data and energy-saving opportunities.

Predictive modeling is another topic for data-mining application in building engineering. Neural networks are extensively used in the field. Examples are Belman-Flores and Ledesma [28] in the case of a multilayer perception (MLP) artificial neural network (ANN); Bechtler et al. [29] in the case of a dynamic neural network; and Swider [30], who used an MLP and radial basis (RBF) ANN.

As pointed out in [4], there is still a need to develop systematic frameworks for evaluation of occupant behavior and its impact on building energy consumption. Problems such as identifying the behavior of occupants accurately by analyzing their data are still at the fundamental levels and have not been solved thoroughly by current data analysis methods. On the one hand, this is due to the lack of clear and consistent definitions of occupant behavior and, on the other, to the lack of sufficient data, the high cost of establishing large databases and storage systems, or privacy issues. Novel integration of data-mining techniques can yield new insights on occupant behavior analysis and provide new pathways for energy use management. This study presents a data-mining approach for analyzing occupants’ energy usage based on a comparison among them.

The challenges that need to be dealt with are:

- How can one monitor the performance of occupants practically and accurately without the interference of other factors and based on their capabilities, so that one is able to give them applicable prioritized recommendations?
- How can one develop a process to assess the performance of a set of buildings based on the energy-saving awareness of their occupants?

1.3. Statement of novelty

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

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

Influencing factor in EUI	Attribute	Category-unit	Abbreviation
City climate	a) Annual mean air temperature	Numerical-°C	T
	b) Annual mean relative humidity	Numerical	RH
	c) Annual mean wind speed	Numerical-m/s	WS
	d) Annual mean global solar radiation	Numerical-MJ/m^2	RA
Building-related characteristics	a) House type[a]	Categorical	HT
	b) Building area	Numerical-m^2	BA
	c) Equivalent leakage area[b]	Numerical-cm^2/m^2	ELA
	d) Heat loss coefficient[c]	Numerical-W/m^{2}K	HLC
Occupant-related characteristics	a) Number of occupants	Numerical	NO
Building services system and operation[d]	a) Space heating and cooling	Categorical	HC
	b) Hot water supply	Categorical	HWS
	c) Kitchen equipment	Categorical	KE

[a] The houses are either detached or apartments and are transformed to [0, 1].
[b] Measured by fan pressurization method.
[c] Calculated based on building design plans.
[d] 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.

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.

2. Methodology

The proposed method is composed of multiple steps, shown in Fig. 1. The data from 80 buildings in Japan are collected, 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 [31]. 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 Fig. 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.

2.1. Grey relational analysis

The influencing factors that affect building energy consumption are listed in Table 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 impacts 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 Fig. 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 [32]. Considering an $n \times m$ data set, y_j denotes the objective sequence (in this case, building energy consumption), and y_i denotes the influencing factors (listed in Table 2). Therefore, i: 1, 2, ...m, and $j: 1,\ 2, \ldots P$ (in this case, $P = 1$), and $k: 1,\ 2, \ldots n$ is the index of data point.

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

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:

$$\xi_j(k) = \frac{\min\limits_i \min\limits_k |y(k) - y_i(k)| + \alpha \max\limits_i \max\limits_k |y(k) - y_i(k)|}{|y(k) - y_i(k)| + \alpha \max\limits_i \max\limits_k |y(k) - y_i(k)|} \tag{2}$$

where α is the “distinguishing coefficient” and is generally set to 0.5 [31]. 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.

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \xi_i(k) \tag{3}$$

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 [9,32].

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 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 2. Therefore, buildings in the same cluster share similar characteristics in terms of weather conditions, building structure, number of occupants, and building services.

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:

$$\sum_{i=1}^{n} \min_{\mu_j} \left\| x_i - \mu_j \right\| \tag{4}$$

where x_i is the ith observation and μ_j is the cluster center. Other similar algorithms are the K-medoids and Manhattan distance,

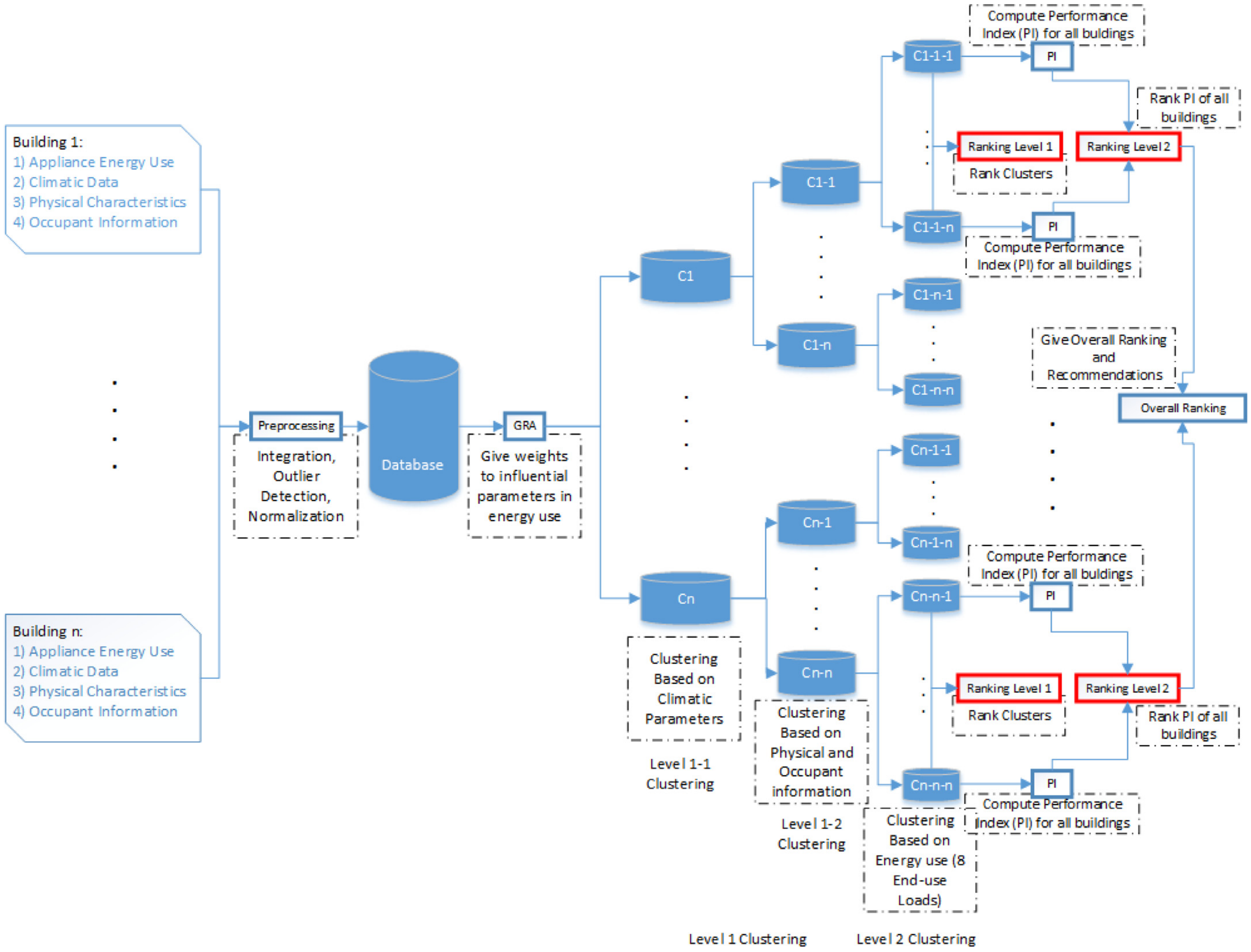

Fig. 1. Framework development for building performance comparison based on occupant behavior.

Pearson correlation, and cosine similarity algorithms [6]. 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 [33].

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 [8,34].

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 2 (*level* 1-1 and *level* 1-2 clustering).

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 Fig. 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 [33]. 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

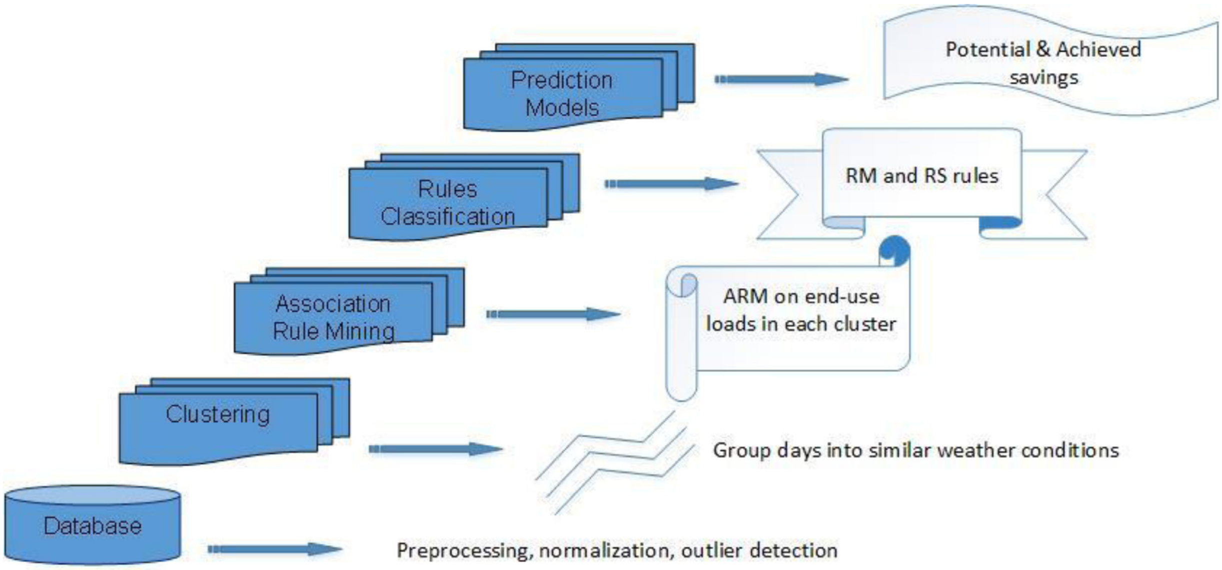
Potential & Achieved savings
Prediction Models
RM and RS rules
Rules Classification
ARM on end-use loads in each cluster
Association Rule Mining
Clustering
Group days into similar weather conditions
Database
Preprocessing, normalization, outlier detection

Fig. 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

Fig. 3. Rule categorization process [34].

described above, which give the occupants information about how to reduce their overall energy consumption to enter to the lower EUI cluster. This constitutes the *level* 1 ranking.

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 Fig. 1 and is applied on each building. PI is defined as:

$$PI = AS - PS \tag{5}$$

AS is the achieved savings and PS is the potential savings. Achieved savings mean that the occupants are lowering 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 Fig. 2. In this process, the eight end-use loads described above are broken down into more detailed data. For example, KITCH data include washing machine, dryer, rice cooker, oven, and so on, depending on the available home appliances [35]. 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 Fig. 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. The flowchart of the process is shown in Fig. 2. More details are provided in [34]. Based on this definition, if a building has low potential for improvement and high savings achievements, it is considered a very good 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.

3. Results and discussion

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 3, in which temperature, relative humidity, wind speed, and solar radiation are functions of time and location 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.

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 contri-

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

Factors\Region	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			

Table 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

bution). 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.

3.2. Reducing effects unrelated to occupant behavior

3.2.1. Clustering based on weather parameters (level 1-1)

As described in Section 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 Fig. 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.

3.2.2. Clustering based on physical 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 Fig. 5. Five clusters were obtained.

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 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.

Fig. 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 Fig. 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 Fig. 5) and apparently occupants tend to use them more than kerosene heaters.

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 influencing parameters listed in Table 2.

3.3. 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 2.3. The optimum number of clusters according to the Silhouette index [9] was two in all clusters. Thus, the cluster cen-

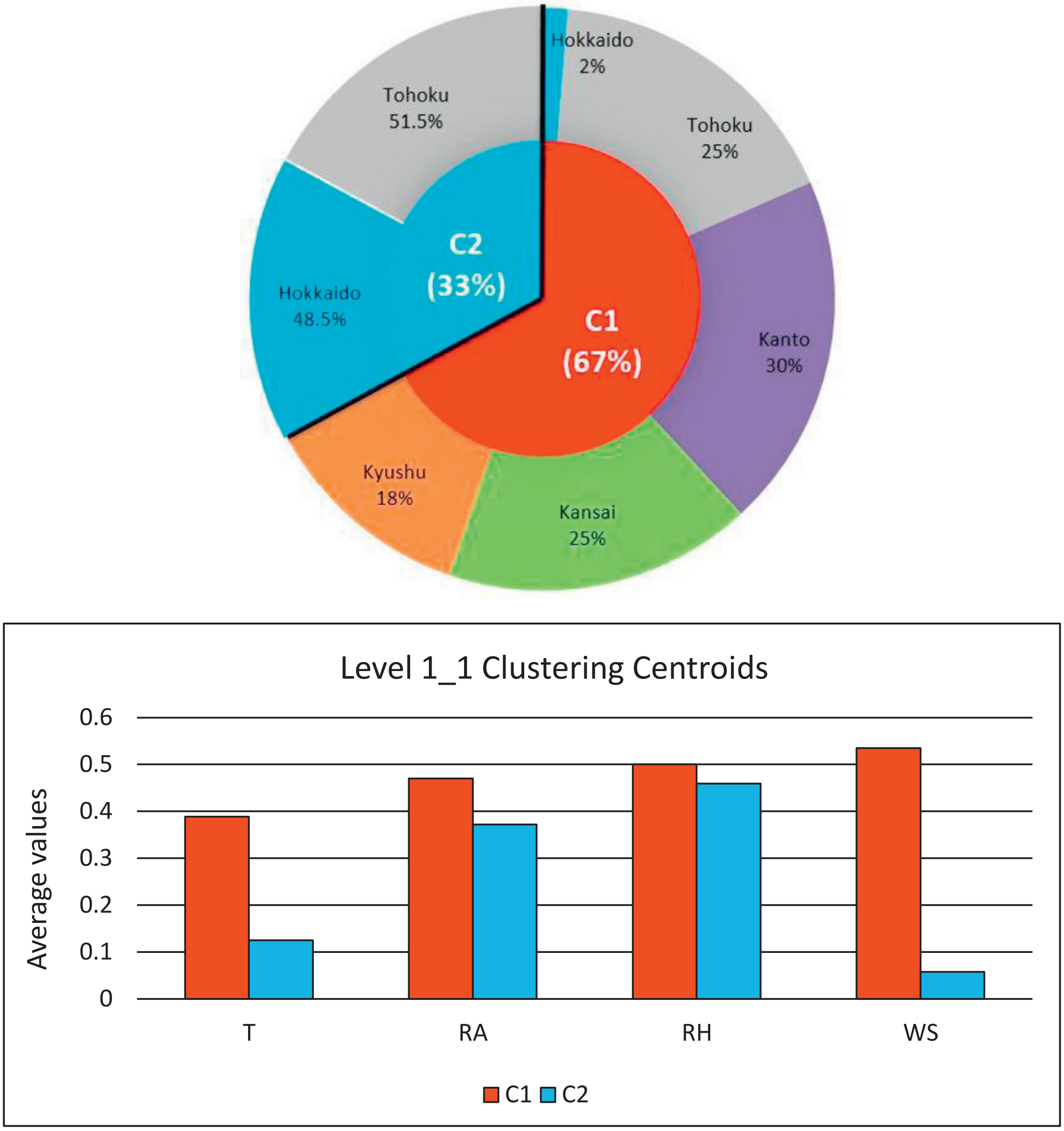

Fig. 4. Clustering *level* 1-1 results on statistics and percentages of instances assigned to each cluster.

troid 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.

Fig. 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 Fig. 7.

Fig. 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 Fig. 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

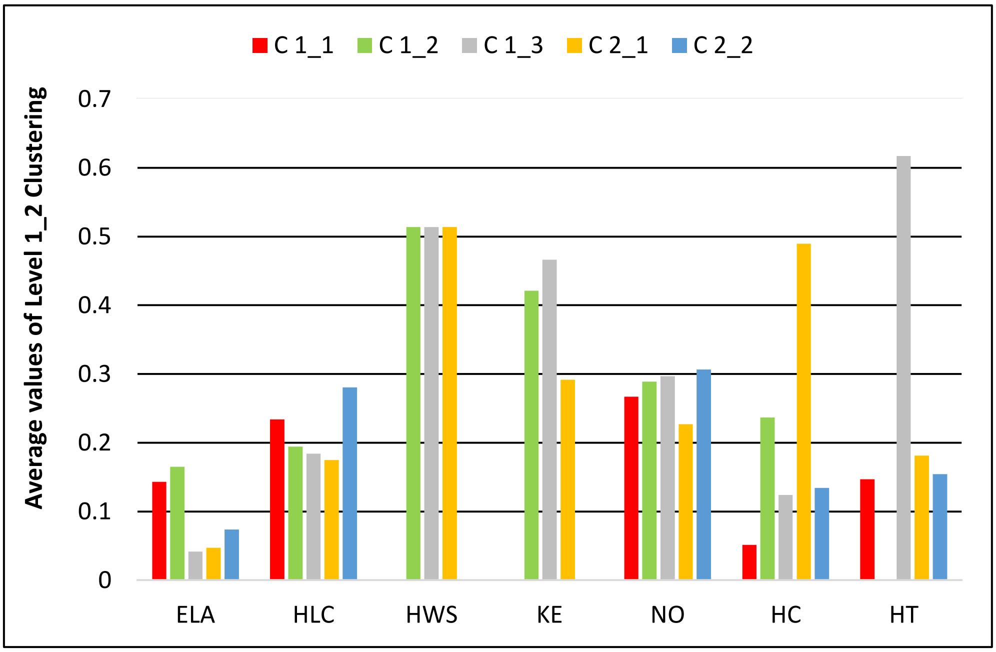

Fig. 5. Distribution of seven physical and occupant characteristics in *level* 1-2 clustering.

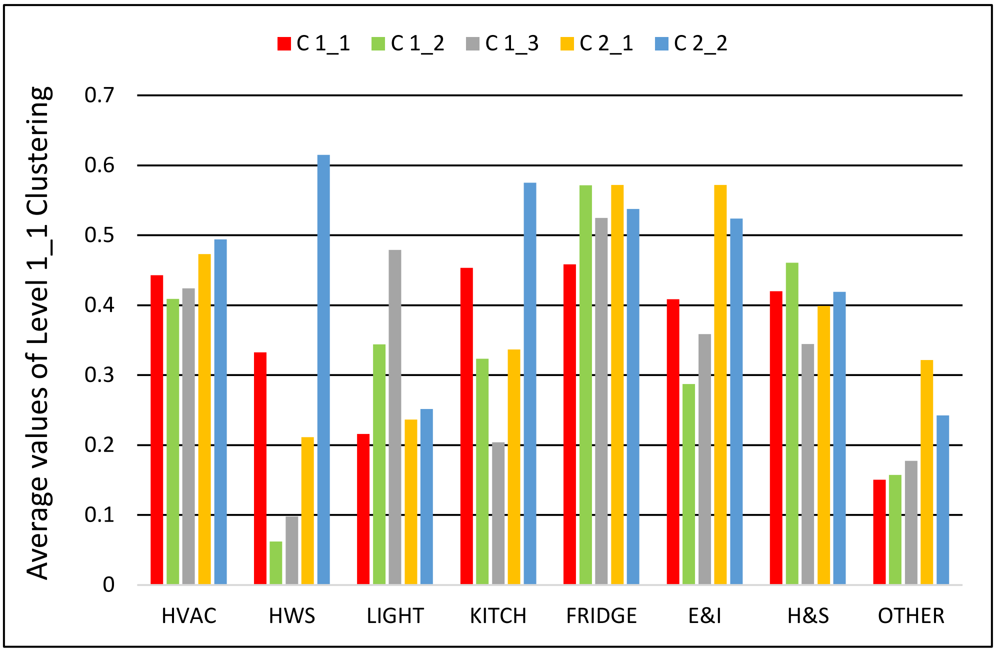

Fig. 6. Distribution of eight end-use loads in different clusters.

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.

3.4. 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 Fig. 3). Two sample rules are indicated in Table 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 histor-

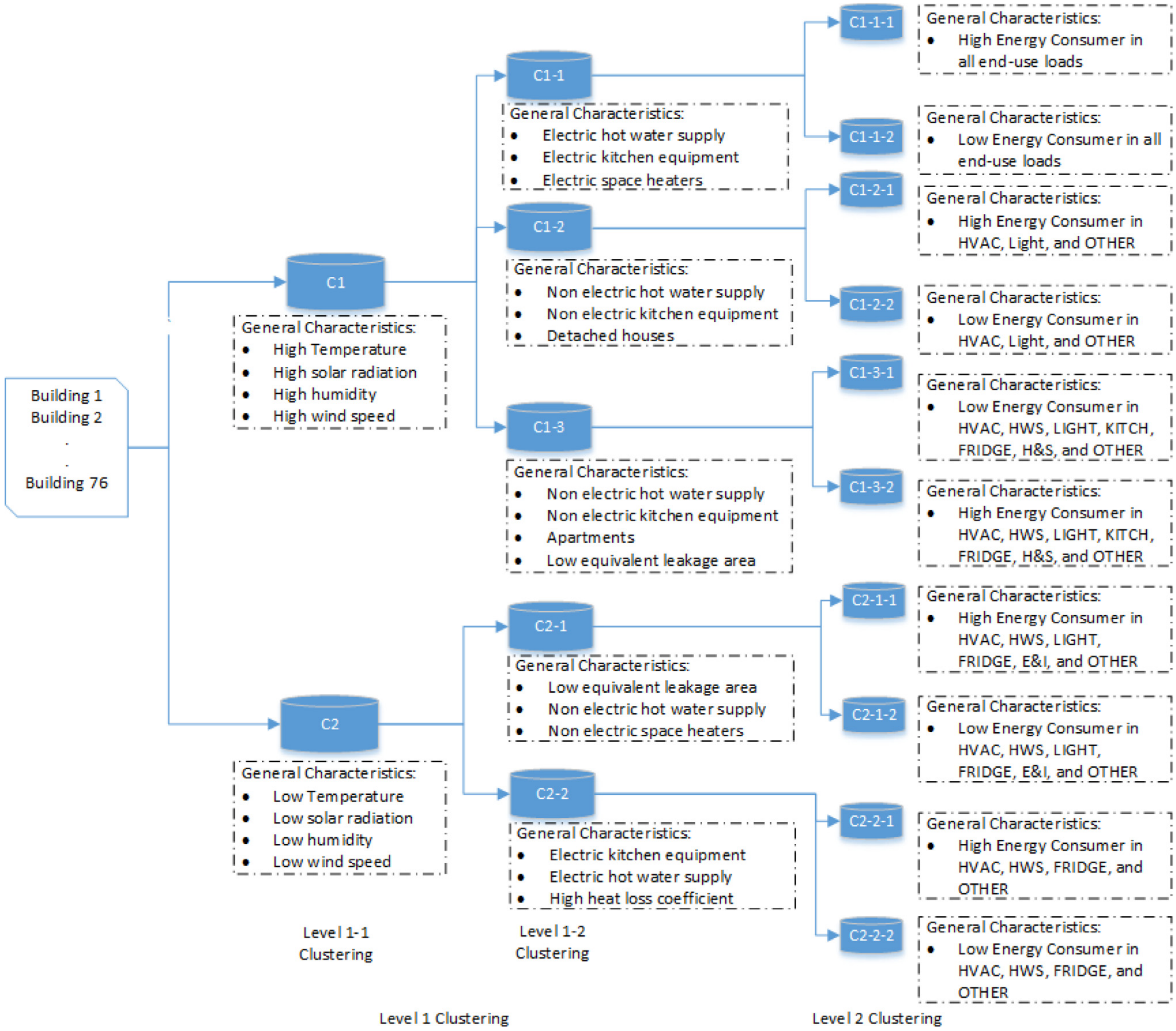
Building 1
Building 2
.
.
Building 76
C1
General Characteristics:
• High Temperature
• High solar radiation
• High humidity
• High wind speed
C2
General Characteristics:
• Low Temperature
• Low solar radiation
• Low humidity
• Low wind speed
C1-1
General Characteristics:
• Electric hot water supply
• 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

Fig. 7. Results of level ranking. Details of clusters and their general characteristics.

ical 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). Figs. 9 and 10 show the outputs of the models based on the rules represented in Table 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. Fig. 9 corresponds to the sample RS rule, and Fig. 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 5). The achieved savings is calculated in a similar manner for RS rules.

It is important to mention that in Eq. (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 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 Fig. 8) and *low energy consumers* (blue bars in Fig. 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 5, building 3 in cluster C 1_1_1 is a *high energy consumer* according to clustering level 2 (red bars in Fig. 8, top left). Therefore, the tenants are advised to try to modify their behavior (especially on the use of HVAC, KITCH, and FRIDGE based on Fig. 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 re-

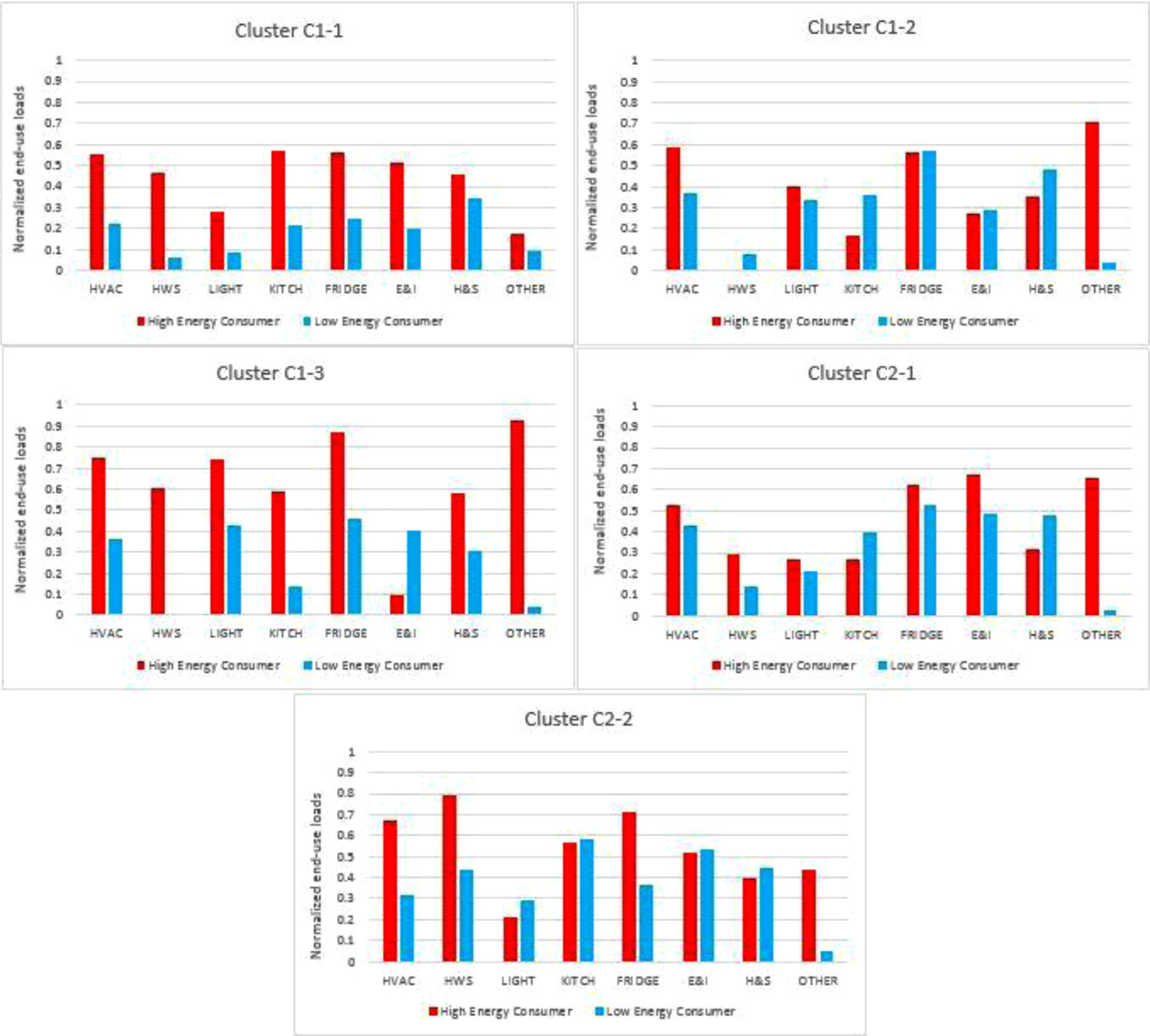

Fig. 8. Clustering *level* 2 results. Data in all clusters were clustered again in terms of EUI. Centroids are shown in blue and red bars and are categorized as buildings with either low or high energy consumption, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 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%

sults are obtained and can be reported to the occupants of other buildings.

4. Limitations and future insights

Some limitations of the proposed methodology are as follows. By resolving these shortcomings, it would be possible to increase the accuracy of the proposed methodology.

Having no hourly data for end-use loads is an important challenge in *level* 2 ranking that makes association rule mining less realistic given that the services might not be operating simultaneously. Consequently, estimations may not be as accurate as would be the case if hourly data were known. Having hourly data would improve the accuracy and reliability of the process beyond that given by daily data.

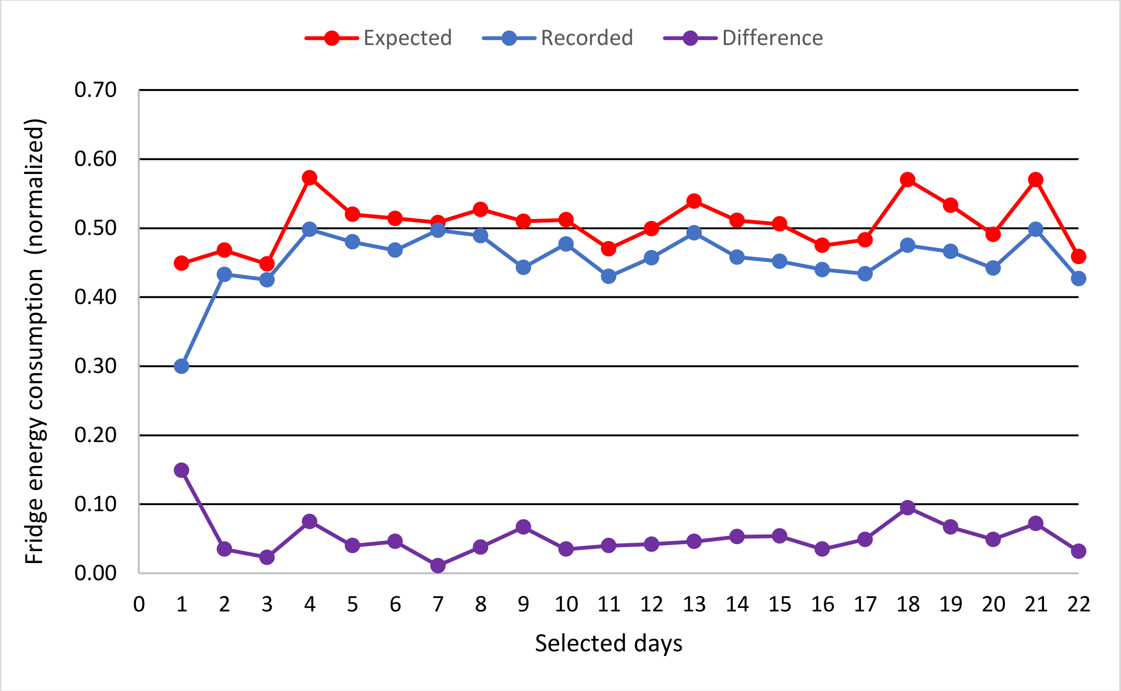

Fig. 9. Calculation of achieved savings based on a sample RS rule.

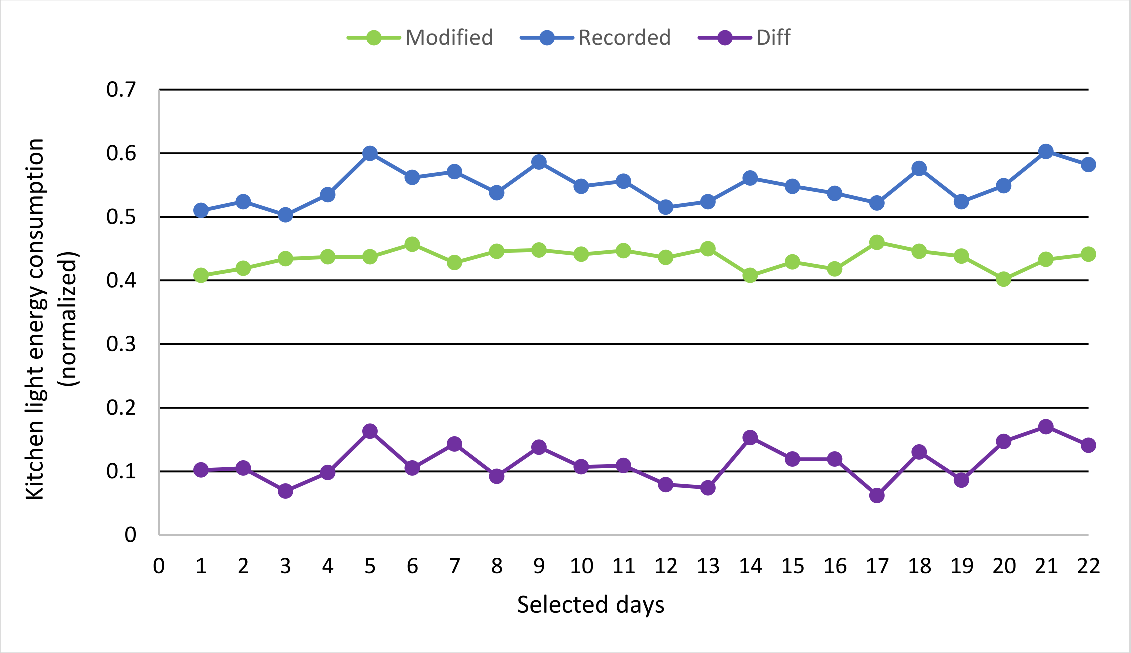

Fig. 10. Calculation of potential savings based on a sample RM rule.

Another challenge originates from data deficiency. For instance, the number of occupants is present in the data set; however, more information—age, level of activity, number of adults and children, time of arrival and departure, and so on—can be incorporated into Table 2. Also, knowing this information would 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, knowing occupants' daily schedules, preferences (e.g., lighting level, room temperature), and holidays.

The size of the data set is also an important factor for most data-mining techniques. Usually, increasing the number of data points will improve the accuracy of machine learning algorithms. Eighty buildings may not be enough to do a perfect clustering. The data set available for this study covered 2003 and 2004 and part of 2002 and 2005 on a daily, monthly, and yearly basis (depending on different parts of the process). Higher-resolution data would increase the size of the data set, making data-mining results more reliable.

One problem in clustering *level* 1 may be that it is quite possible that buildings in the same cluster have similar characteristics on some non–occupant-related features but are dissimilar in other non-occupant-related features. To reduce the effect of this problem, each characteristic was multiplied by its GRA (grey relational analysis). Therefore, those characteristics with higher GRA values are dominant in determining the cluster to which the building belongs. Also, clustering *level* 1 was divided into two subsections, *levels* 1-1 and 1-2, to prioritize the weather parameters in clustering. If one tries to increase the accuracy of the clustering in terms of any characteristics, it is possible to increase the subsections of *level* 1 clustering to even more than two (for example, one more level for 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 solve this issue.

Having a 24-hour temperature profile of each building throughout the year can further improve clustering *level* 1-1 to capture fluctuations in weather data. However, we have used the annual average values due to the data deficiency.

When feedback is given to the *high energy consumer*, they are expected to take suitable measures to improve their rank. This can be achieved by a real case evaluation to see the effectiveness of the proposed system, especially if the ranking is performed online so the occupants can see the effect of their energy saving measures within a short period of time.

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 which is not identifiable through current methodology. In other words, the buildings are not comparing themselves with the best cases. This limitation lays the foundation for the future work, which is creating a role model building for energy usage evaluation.

5. 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.

Conflict of interest

This is to confirm that there is not any conflict of interest.

Acknowledgments

The authors express their gratitude to Concordia University for supporting this research.

References

[1] N.R. Canada, Energy Efficiency Trends in Canada 1990 to 2013, 2013, pp. 11–18 https://www.nrcan.gc.ca/energy/publications/19030.
[2] T. Hong, H.-W. Lin, Occupant behavior: impact on energy use of private offices, in: ASim 2012 - 1st Asia Conf. Int. Build. Perform. Simul. Assoc., 2013.
[3] M. Sameti, F. Haghighat, Optimization of 4th generation distributed district heating system: Design and planning of combined heat and power, Renew. Energy 130 (2019) 371–387, doi:10.1016/J.RENENE.2018.06.068.
[4] Y. Zhang, X. Bai, F.P. Mills, J.C.V. Pezzey, Rethinking the role of occupant behavior in building energy performance: A review, Energy Build. 172 (2018) 279–294, doi:10.1016/J.ENBUILD.2018.05.017.
[5] S. Bhattacharjee, G. Reichard, Socio-Economic Factors Affecting Individual Household Energy Consumption: A Systematic Review, ASME 2011 5th Int. Conf. Energy Sustain, 2011, pp. 891–901 http://dx.doi.org/10.1115/ES2011-54615.
[6] C. Fan, F. Xiao, Z. Li, J. Wang, Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review, Energy Build 159 (2018) 296–308, doi:10.1016/J.ENBUILD.2017.11.008.
[7] C. Miller, Z. Nagy, A. Schlueter, A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings, Renew. Sustain. Energy Rev. 81 (2018) 1365–1377, doi:10.1016/J.RSER.2017.05.124.
[8] Z. Yu, B.C.M. Fung, F. Haghighat, Extracting knowledge from building-related data—A data mining framework, Build. Simul. 6 (2013) 207–222.
[9] Z. Yu, B.C.M. Fung, F. Haghighat, H. Yoshino, E. Morofsky, A systematic procedure to study the influence of occupant behavior on building energy consumption, Energy Build. 43 (2011) 1409–1417. https://doi.org/10.1016/j.enbuild.2011.02.002.
[10] S. D'Oca, T. Hong, A data-mining approach to discover patterns of window opening and closing behavior in offices, Build. Environ. 82 (2014) 726–739.
[11] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy Build. 75 (2014) 109–118.
[12] A. Capozzoli, D. Grassi, M.S. Piscitelli, G. Serale, Discovering knowledge from a residential building stock through data mining analysis for engineering sustainability, Energy Procedia. 83 (2015) 370–379.
[13] C. Fan, F. Xiao, C. Yan, A framework for knowledge discovery in massive building automation data and its application in building diagnostics, Autom. Constr. 50 (2015) 81–90, doi:10.1016/j.autcon.2014.12.006.
[14] A. Lavin, D. Klabjan, Clustering time-series energy data from smart meters, Energy Effic. 8 (2015) 681–689, doi:10.1007/s12053-014-9316-0.
[15] P. Howard, G. Runger, T.A. Reddy, S. Katipamula, Automated data mining methods for identifying energy efficiency opportunities using whole-building electricity data, ASHRAE Trans. (2016) 422–433.
[16] C.M.R. do Carmo, T.H. Christensen, Cluster analysis of residential heat load profiles and the role of technical and household characteristics, Energy Build. 125 (2016) 171–180. http://doi.org/10.1016/j.enbuild.2016.04.079.
[17] J. Yang, C. Ning, C. Deb, F. Zhang, D. Cheong, S.E. Lee, C. Sekhar, K.W. Tham, k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement, Energy Build. 146 (2017) 27–37, doi:10.1016/J.ENBUILD.2017.03.071.
[18] N. Gaitani, C. Lehmann, M. Santamouris, G. Mihalakakou, P. Patargias, Using principal component and cluster analysis in the heating evaluation of the school building sector, Appl. Energy. 87 (2010) 2079–2086, doi:10.1016/J.APENERGY.2009.12.007.
[19] Z. Yu, F. Haghighat, B.C.M. Fung, H. Yoshino, A decision tree method for building energy demand modeling, Energy Build. 42 (2010) 1637–1646.
[20] F.W. Yu, K.T. Chan, Using cluster and multivariate analyses to appraise the operating performance of a chiller system serving an institutional building, Energy Build. 44 (2012) 104–113, doi:10.1016/J.ENBUILD.2011.10.026.
[21] H. Bechtler, M.W. Browne, P.K. Bansal, V. Kecman, New approach to dynamic modelling of vapour-compression liquid chillers: Artificial neural networks, Appl. Therm. Eng. 21 (2001) 941–953.
[22] F.W. Yu, K.T. Chan, Assessment of operating performance of chiller systems using cluster analysis, Int. J. Therm. Sci. 53 (2012) 148–155, doi:10.1016/J.IJTHERMALSCI.2011.10.009.
[23] P. Xue, Z. Zhou, X. Fang, X. Chen, L. Liu, Y. Liu, J. Liu, Fault detection and operation optimization in district heating substations based on data mining techniques, Appl. Energy. 205 (2017) 926–940, doi:10.1016/J.APENERGY.2017.08.035.
[24] G. Li, Y. Hu, H. Chen, H. Li, M. Hu, Y. Guo, J. Liu, S. Sun, M. Sun, Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions, Appl. Energy. 185 (2017) 846–861, doi:10.1016/J.APENERGY.2016.10.091.
[25] C. Fan, F. Xiao, H. Madsen, D. Wang, Temporal knowledge discovery in big BAS data for building energy management, Energy Build. 109 (2015) 75–89, doi:10.1016/J.ENBUILD.2015.09.060.
[26] D. Patnaik, M. Marwah, R.K. Sharma, N. Ramakrishnan, Temporal data mining approaches for sustainable chiller management in data centers, ACM Trans. Intell. Syst. Technol. 2 (2011) 34:1–34:29, doi:10.1145/1989734.1989738.
[27] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, Autom. Constr. 49 (2015) 1–17, doi:10.1016/J.AUTCON.2014.09.004.
[28] J.M. Belman-Flores, S. Ledesma, Statistical analysis of the energy performance of a refrigeration system working with R1234yf using artificial neural networks, Appl. Therm. Eng. 82 (2015) 8–17, doi:10.1016/J.APPLTHERMALENG.2015.02.061.
[29] H. Bechtler, M.W. Browne, P.K. Bansal, V. Kecman, New approach to dynamic modelling of vapour-compression liquid chillers: Artificial neural networks, Appl. Therm. Eng. 21 (2001) 941–953, doi:10.1016/S1359-4311(00)00093-4.
[30] D.J. Swider, A comparison of empirically based steady-state models for vapor-compression liquid chillers, Appl. Therm. Eng. 23 (2003) 539–556, doi:10.1016/S1359-4311(02)00242-9.
[31] S. Walfish, A review of statistical outlier methods title, Pharm. Technol. 11 (2006) 82–88.

[32] C. Fu, J. Zheng, J. Zhao, W. Xu, Application of grey relational analysis for corrosion failure of oil tubes, Corros. Sci. 43 (2001) 881–889. http://dx.doi.org/10.1016/S0010-938X(00)00089-5.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[34] M. Ashouri, F. Haghighat, B.C.M. Fung, A. Lazrak, H. Yoshino, Development of building energy saving advisory: A data mining approach, Energy Build. 172 (2018) 139–151, doi:10.1016/j.enbuild.2018.04.052.

[35] Z.(Jerry) Yu, F. Haghighat, B.C.M. Fung, E. Morofsky, H. Yoshino, A methodology for identifying and improving occupant behavior in residential buildings, Energy 36 (2011) 6596–6608, doi:10.1016/J.ENERGY.2011.09.002.