Detecting Anomalous Energy Consumption from Profiles

Detecting Anomalous Energy Consumption from Profiles

Hiroto Abe [1]*, Kazuaki BOGAKI[1], H.B. RIJAL[1] & Mahito SUGIYAMA[2]

[1]Tokyo City University

[2]National Institute of Informatics

*g1793101@tcu.ac.jp

Abstract. Controlling and reducing electric energy consumption is a critical issue across all over the countries for human wellbeing. However, approaches to achieve energy consumption reduction of individual occupants have not been established yet as both problems of collecting a large amount data of energy consumption and constructing prediction models are challenging. Here we show a case-study of energy consumption analysis, in which households with anomalous energy consumption can be completely detected using a questionnaire about their profiles *without seeing actual energy consumption*. Our approach is based on simple data mining techniques of outlier detection and decision trees, hence it can be easily implemented in the condominium housing market.

1. Introduction

Recently, a massive increase in electric energy consumption is a serious problem across all over the countries. A number of countries set the target of reduction of energy consumption, for example, the goal is set to 39% energy reduction in Japan [1].

To achieve energy consumption reduction, it is fundamental to approach the behavior of each individual occupant. Typical recommended behavior includes turning off lights and setting lower (resp. higher) temperature for a heating system (resp. air conditioning). However, the approach of forcing such behavior to occupants is not sustainable as it requires lots of patience for occupants in their daily life. Another approach is automatic control of home electronics, such as an air conditioner and a washing machine, through the internet to reduce energy consumption while keeping the comfort of a room [2-7]. For example,[8]builds a prediction model of energy consumption using machine learning techniques, and [9-11] used data mining approaches such as pattern mining and decision trees to find association between energy consumption and occupant behavior. However, a huge amount of data of energy consumption is needed to learn accurate control rules which simultaneously achieve energy reduction and indoor comfort. Since such rules are different across occupants, a long training term is required to achieve an effective control system for every individual occupant. This is particularly a serious problem for households with high energy consumption as they should be detected and notified as soon as possible to avoid waste of energy.

In this paper, we focus on anomalous energy consumption, with which households typically consume much higher energy than other households but may also consume much lower energy that can be a good example for other households. Using data mining techniques, we show that households with anomalous

energy consumption can be detected using a limited amount of their profiles without using any information of actual energy consumption. More precisely, the anomalousness of energy consumption is automatically determined by an outlier detection technique [12] from a time series of 24 hours monitoring of energy consumption. We then construct decision trees [13]from a questionnaire about profiles that classifies households into anomalies or not and show that they can accurately discriminate households with anomalous energy consumption from the others. To date, although there are some attempts to predict energy consumption[8], none of studies have succeeded to predict energy consumption from a profile questionnaire.

Our approach can immediately help daily life of occupants who have just moved to a new apartment as whether or not they will show anomalous energy consumption can be judged by their profiles without observing actual energy consumption. Furthermore, our study leads to an impact in housing market as industries can prepare for preventing anomalous energy consumption from profiles of households before they start to live. This paper shows the effectiveness of data mining approach for energy control from individual occupants to home builders.

2. Our Approach

Our approach consists of two components: First we identify occupants with anomalous energy consumption from records of energy consumption by unsupervised outlier detection, which is used as ground-truth labels in decision tree classification. Second, we construct decision trees from a questionnaire about profiles that classifies anomalous and non-anomalous occupants. We analyze the learned decision trees and identify which questions are effective in anomaly classification. We summarize our approach in Figure 1. The remarkable feature of our approach is that we do not use energy consumption data in classification of anomalous and non-anomalous energy consumption, and the classification phase is achieved using only questionnaire data of occupant profiles without seeing energy consumption records.

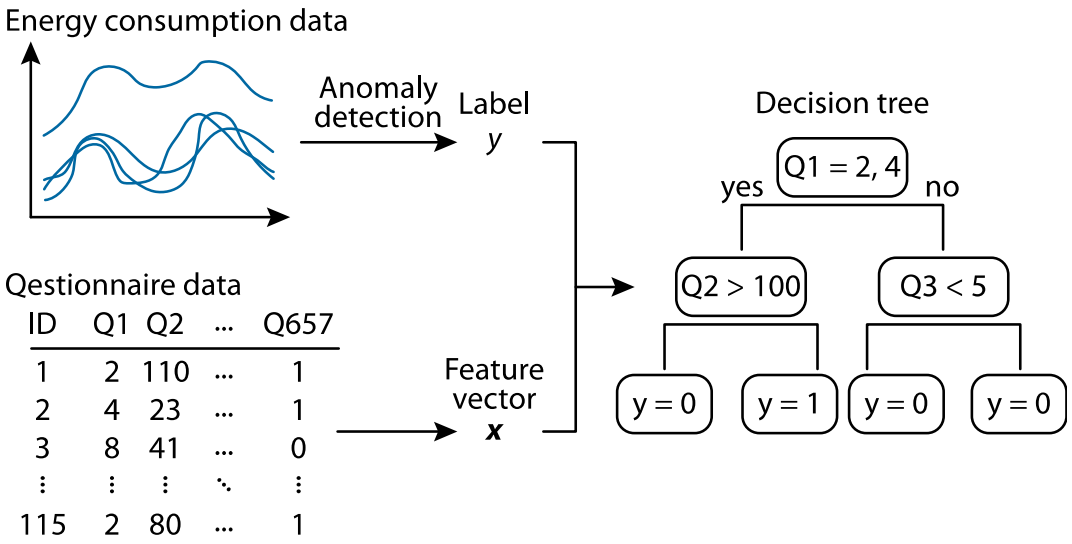

Figure 1: Overview of our approach.

2.1. Datasets

We use two types of datasets: an electric energy consumption dataset and a questionnaire dataset.

The energy consumption dataset was collected from a newly build condominium in Shinagawa-ku, Tokyo in Japan, which includes 358 households. In each household, electric consumption is recorded every 30 minutes for 24 hours from January 2017 to December 2017. We have cleaned the dataset by removing errors of recording devises. Since electric consumption largely differs depending on season

and days (weekdays or weekends), we divided energy consumption into weekdays or weekends, where national holidays are included in weekends, and also divided into each month (Figure 2). As a result, we have 24 datasets in total, each of which consists of 48m-dimensional 358 data points, where m is the number of days in each month.

The questionnaire dataset has been obtained from the same condominium and room IDs match with those of the energy consumption data. The questionnaire consists of 657 questions, which includes a wide range of questions from basic profiles such as sex, age, family form, working style, and more detailed information such as previous residences, owned home electronics, energy saving consciousness, and daily life style. This questionnaire was done after more than three years of the beginning of living at the condominium for every household and obtained from 115 households. As a result, we have a dataset with 115 data points, each of which is a 657-dimensional vector.

2.2. Anomaly Detection

To identify households with anomalous energy consumption, we applied the state-of-the-art outlier detection method proposed by [12] to the energy consumption dataset, which was shown to be effective and efficient in unsupervised outlier detection. Given a dataset $T = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n\}$, $\mathbf{t}_i \in \mathbb{R}^d$, the *outlierness score* for each data point $\mathbf{t} \in T$ is defined as

$$q(\mathbf{t}) = \min_{\mathbf{t}' \in S(T)} D(\mathbf{t}, \mathbf{t}')$$

where $S(T) \subset T$ is randomly and independently sampled from T and $D(\mathbf{t}, \mathbf{t}')$ is distance between $\mathbf{t}$ and $\mathbf{t}'$. In this paper, we consistently use Euclidean distance defined as

$$D(\mathbf{t}, \mathbf{t}') = \sqrt{\sum_{i=1}^{d} (t_i - {t'}_i)^2}$$

for $\mathbf{t} = (t_1, t_2, \dots, t_d)$ and $\mathbf{t}' = ({t_1}', {t_2}', \dots, {t_d}')$. We set the sample size $|S(T)| = 20$, which is recommended in [12]. Since a larger score corresponds to higher anomalousness, data points in T are sorted in decreasing order according to their scores, and top-k data points are detected as anomalies. We set $k = 10$ throughout the paper. We have used the R package `spoutlier`.

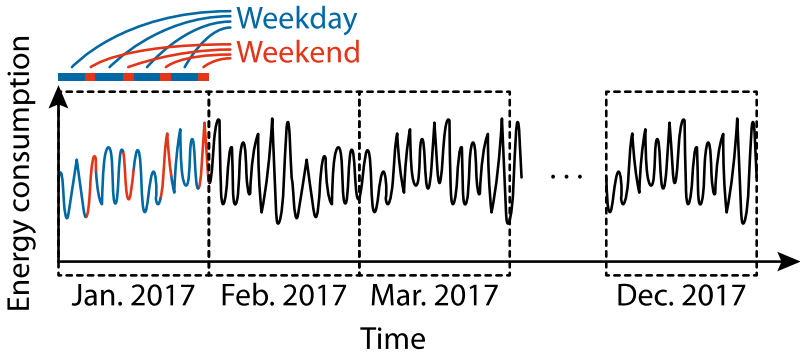

Figure 2: Structure of energy consumption data.

2.3. Decision Tree Classification

We formulate the problem of finding anomalies as classification and use the decision tree method CART[13], which is known to be the standard decision tree method, as the resulting decision tree is interpretable. Given the questionnaire dataset $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ with $n = 115$, where each feature corresponds to a respective question, we first obtain their labels by the above outlier detection method.

More precisely, for each data point $\mathbf{x}_i$ its label $y_i = 1$ if it is included in the top-k outliers ($k = 10$) and $y_i = 0$ otherwise. Then a decision tree is constructed from the labeled dataset. We have used the R package `rpart`.

2.4. Evaluation

We use precision and recall to evaluate the performance of decision trees, which can be used for evaluation even if datasets are largely imbalanced [14]. Let A={x_i∈X|y_i=1}⊆X be the set of anomalies and B⊆X be the set of data points that are classified into anomalies by the constructed decision tree. Then the precision is defined as "precision"=|A∩B|/|B| and the recall is defined as "recall"=|A∩B|/|A|. Both scores take from 0 to 1 and higher is better.

3. Experiments

We apply our approach (Figure 1) to the energy consumption and questionnaire datasets. We used macOS 10.36.6 and ran all experiments on 4.0 GHz Intel Core i7 and 32 GB of memory. All analysis are performed in R version 3.5.1 [15].

3.1. Detecting Anomalous Energy Consumption from Questionnaire of Profiles

First we identified households with anomalous energy consumption from each of the 24 energy consumption datasets. Since we have questionnaire data for 115 out of 358 households, we used only such 115 households in anomaly detection and identified top-10 anomalies. We plot energy consumption of 24 hours for August and February with weekdays and weekends in Figure 3, where anomalous and non-anomalous energy consumption are averaged, respectively. In these datasets, all anomalies consume much higher energy than non-anomalies. We then applied CART and observed that *we can always perfectly classify anomalies* (both precision and recall are 1) regardless of months and weekdays or weekends by a decision tree.

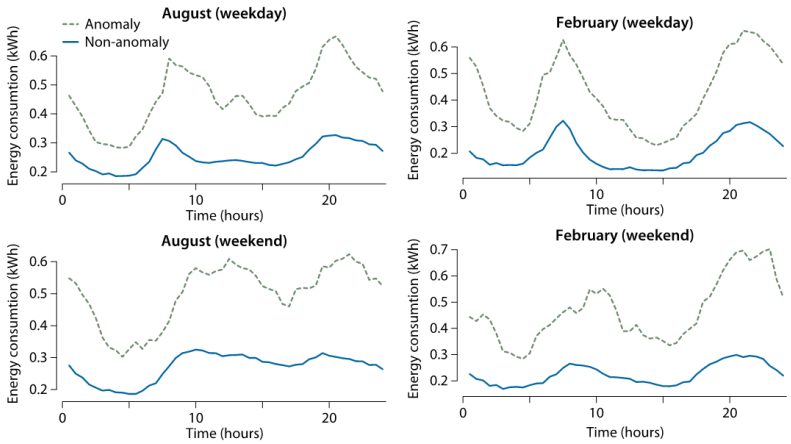

Figure 3: Anomalous and non-anomalous energy consumption for 24 hours.

Here an interesting problem to solve is: which question is important in classification of anomalies out of our 657 questions? To solve the problem, first we calculated the frequency of used questions across 24 datasets (12 months × weekday or weekend). The resulting histogram of top-20 frequently used questions is plotted in Figure 4 and each question is shown in Table [table:questions]. The top-20 questions out of 657 questions are categorized as follows:

- About previous residence: 6

- About energy saving consciousness and behavior: 7
- About home electronics: 5
- About age of the person: 1
- About duration of stay on weekdays: 1

This means that three topics: previous residence, energy saving consciousness and behavior, and home electronics, might have a stronger association with anomalous energy consumption than basic profiles of individuals such as sex, age range, and number of family members. It should be also noted that basic profiles, which are used in a various types of analysis, are not informative in prediction of anomalous energy consumption.

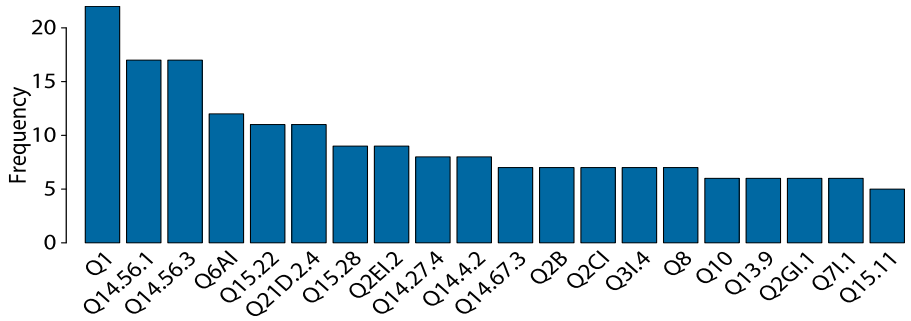

Figure 4: Frequency of top-20 questions in 24 decision trees.

Table 1: Top-20 frequently used questions.

Questions about previous residence		Type	Description
Q1	Housing type	Nominal	1 (owned house), 2 (owned apartment), 3 (rented house), 4 (rented apartment), 5 (UR), 6 (public house), 7 (rented small apartment), 8 (company housing), 9 (dormitory), 10 (parents' house), 11 (others)
Q2B	Construction type	Nominal	1 (wood), 2 (steel), 3 (reinforced concrete), 4 (unknown)
Q2CI	Size of the room	Numerical	Value
Q2EI.2	Number of air conditioners (AC)	Numerical	Value
Q2GI.1	Age of a building	Numerical	Value
Q3I.4	Annual gasoline consumption	Numerical	Value
Questions about energy saving consciousness and behavior		Type	Description
Q8	Have daily energy saving consciousness	Ordinal	1 (yes) – 4 (no)
Q10	Reason not to act on energy saving	Nominal	1 (no time), 2 (bothering), 3 (costly), 4 (no merit), 5 (others)
Q13.9	Energy saving information source	Binary	1 (do not obtain), 0 (obtain)
Q15.11	Use ventilation by opening of windows	Ordinal	1 (yes) – 4 (no)
Q15.22	Use a gas stove	Ordinal	1 (yes) – 4 (no)
Q15.28	Use a heating toilet seat	Ordinal	1 (yes) – 4 (no)
Q21.D.2.4	Opportunities to see the HEMS (home energy management system) screen	Binary	1 (before bath), 0 (not before bath)
Questions about home electronics		Type	Description
Q14.4.2	Use air conditioners from previous residence	Binary	1 (yes), 0 (no)
Q14.27.4	Have a second energy saving desktop PC	Binary	1 (yes), 0 (no)
Q14.56.1	Do not have a table top dishwasher	Binary	1 (no), 0 (yes)
Q14.56.3	Newly buy a table top dishwasher	Binary	1 (yes), 0 (no)
Q14.67.3	Newly buy a private room lighting	Binary	1 (yes), 0 (no)
Other questions		Type	Description
Q6AI	Age	Numerical	Value
Q7I.1	Absence hours of weekday	Numerical	Value

3.2. Selecting Informative Questions in Detecting Anomalous Energy Consumption

Next we tried to detect anomalies using only these top-20 frequently used questions, that is, we constructed a decision tree from not the original 657-dimensional but k-dimensional feature vectors, where we took top-k frequent questions for the features and we varied k as $k = 1,2, \ldots ,20$. Hence only the question Q1 is used when $k = 1$ and all questions in Table [table:questions] are used if $k = 20$. Figure 5 shows precision and recall when we increase the number of questions from 1 to 20. These scores are averaged over 24 energy consumption datasets. This result shows that precision and recall increase as the number of features (questions) increases and both of them reach to 1.0 if more than 12 questions are used in construction of decision trees. Hence our analysis shows that anomalous energy consumption can be characterized by the only 13 questions.

These 13 questions include:

- About previous residence: 4
- About energy saving consciousness and behavior: 3
- About home electronics: 5
- About age: 1

In particular, three questions about energy saving consciousness and behavior are as follows:

- Q15.22: Use a gas stove
- Q15.28: Use a heating toilet seat
- Q21.D.2.4: Opportunities to see the HEM screen

These questions ask energy saving behavior, while they do not directly ask the consciousness of occupants. This result indicates that questions about life style facts are important for energy saving, while direct questions about the consciousness of occupants are not informative in anomaly prediction. The reason might be that people tend to give a positive answer to such questions about the consciousness of energy saving, and such answers do not match to the actual life styles. Moreover, it is clear that the question of age of occupants (Q6AI), which is the fourth frequent question, affects to the performance as the scores (both precision and recall) increase a lot if this question is added (see the third and the fourth points in each plot in Figure 5). Hence this is the most important feature in the basic profiles.

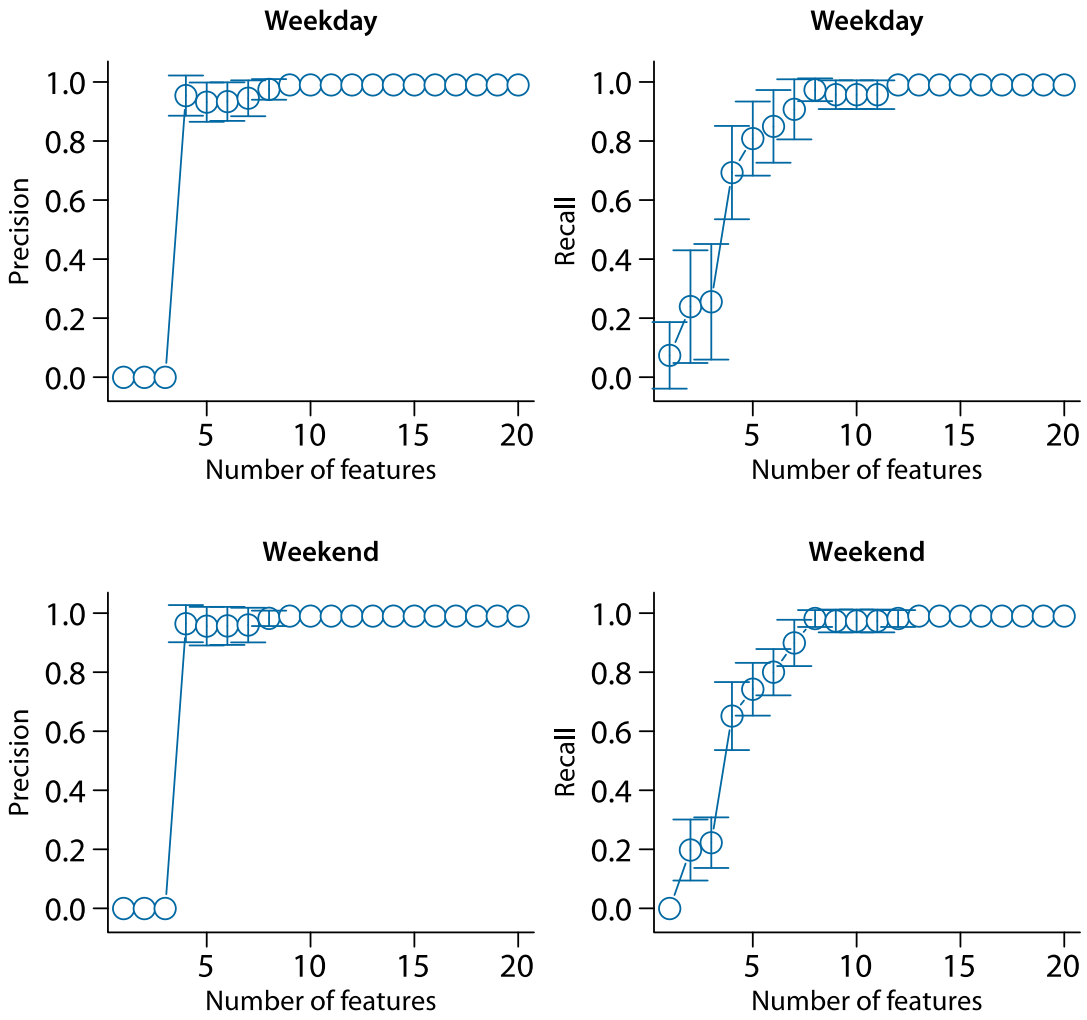

Figure 5: Precision and recall in classification of anomalous energy consumption by the decision tree. Each point shows mean ± SD. Missing error bar means that SD is zero.

3.3. Analysis of Decision Trees

We have shown that all anomalies can be discriminated from non-anomalies by 13 questions. Since one of the most dominant issues of energy consumption is air conditioning, in the following we perform a detailed analysis of decision trees for August and February with the highest and the lowest average temperatures.

Decision trees for August and February with weekdays and weekends are shown in Figure [figure:rpart]. We provide the distribution of nodes in the decision trees according to their question types in Table [table:type]. Interestingly, in all cases, more than 50% of nodes are questions about previous residences. Despite the fact that most of energy consumption is dominated by air conditioning and this questionnaire was done after more than three years of the beginning of living at the condominium, this result means that life style of the previous residents still strongly affects to the current life style.

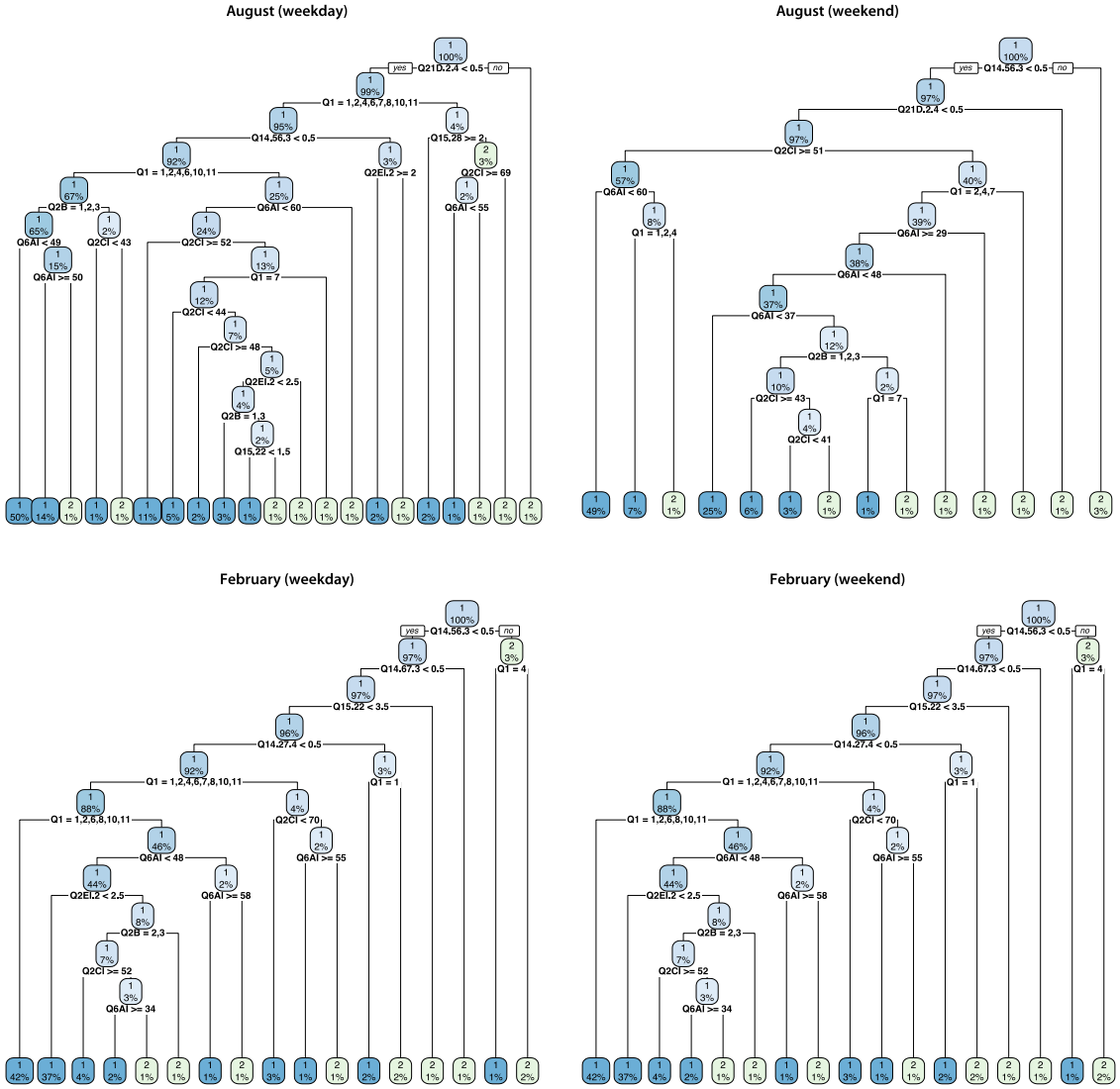

Figure 6: Decision trees for August and February with weekdays and weekends using top-13 questions of profiles. The list of question and their answers are presented in Table 2.

Table 2: Node types in decision trees

August (weekday)	
The number of nodes	20
Previous residence	12
Energy saving consciousness and behavior	3
Home electronics	1
Age	4
February (weekday)	
The number of nodes	16
Previous residence	8
Energy saving consciousness and behavior	1
Home electronics	3
Age	4

August (weekend)	
The number of nodes	13
Previous residence	7
Energy saving consciousness and behavior	1
Home electronics	1
Age	4
February (weekend)	
The number of nodes	16
Previous residence	8
Energy saving consciousness and behavior	1
Home electronics	3
Age	4

4. Conclusions

This paper presents a case-study of energy consumption analysis at a condominium. We have shown that anomalous energy consumption can be completely characterized by only 13 questions about occupant profiles without using actual energy consumption data. Moreover, the majority of the used 13 questions is about previous residences, which means that the past life style could still dominate the way of energy consumption of occupants.

Since our approach is based on simple data mining methods, our approach can be easily implemented in the condominium housing market and used in communication between real estate developers or energy suppliers and home purchasers. More precisely, the implementation scenario of our approach can be divided into the following five steps.

1. The user who wants to purchase a condominium contacts a housing supplier.
2. A housing supplier obtains the profile of the user using the questions shown in this paper.
3. Using the obtained profile and the learned decision tree, the housing supplier can classify whether or not the user uses electric power abnormally.
4. Information about energy consumption saving can be provided to the potential user who may abnormally use electric power.
5. The user can obtain the information about energy saving before moving to the condominium, hence the user can avoid abnormal energy power usage, which leads to a smooth and satisfying life.

The information about energy saving contains more efficient usage of air conditioners, a favorable effect on health condition by suppressing excessive use of them, and the comfortableness of natural wind. Furthermore, the user tend to newly buy or replace home electronics when he/she moves to a new house, thereby it is effective to provide such information about energy saving to address replacement of energy consuming home appliances.

To summarize, our approach can effectively reduce the overall energy consumption at a condominium by detecting few anomalous users and informing them useful information about energy saving. Therefore social implementation of our approach may leads to the significant impact on energy saving.

Moreover, analyzing non-anomalous energy consumption is also an interesting study. We have already obtained data about electric energy consumption at each branch, gas consumption, and water consumption at the same condominium. Using these datasets to find detailed energy consumption patterns and discover associations with such consumption patterns and occupant profiles will lead to more advanced energy saving technologies.

References

[1] Nagura, Y. 2015. “On Future Global Warming Countermeasures of Japan That Received the COP 21 Agreement.” In *Global Warming Forum*, 1–13.

[2] Yu, Zhun, Benjamin CM Fung, Fariborz Haghighat, Hiroshi Yoshino, and Edward Morofsky. 2011. “A Systematic Procedure to Study the Influence of Occupant Behavior on Building Energy Consumption.” *Energy and Buildings* 43 (6): 1409–17.

[3] Hargreaves, Tom, Michael Nye, and Jacquelin Burgess. 2010. “Making Energy Visible: A Qualitative Field Study of How Householders Interact with Feedback from Smart Energy Monitors.” *Energy Policy* 38 (10): 6111–9.

[4] Monacchi, Andrea, Dominik Egarter, Wilfried Elmenreich, Salvatore D’Alessandro, and Andrea M Tonello. 2014. “GREEND: An Energy Consumption Dataset of Households in Italy and Austria.” In *IEEE International Conference on Smart Grid Communications*, 511–16. IEEE.

[5] Hong, Taehoon, Choongwan Koo, and Sungki Park. 2012. “A Decision Support Model for Improving a Multi-Family Housing Complex Based on CO2 Emission from Gas Energy Consumption.” *Building and Environment* 52: 142–51.

[6] Gouveia, J. P., J. Seixas, and G. Long. 2018. “Mining Households’ Energy Data to Disclose Fuel

Poverty: Lessons for Southern Europe." *Journal of Cleaner Production* 178: 534–50.
[7] Ebrahim, A., and O. Mohammed. 2018. "Pre-Processing of Energy Demand Disaggregation Based Data Mining Techniques for Household Load Demand Forecasting." *Inventions* 3 (3): 45.
[8] Tso, G. K. F., and K. K. W. Yau. 2007. "Predicting Electricity Energy Consumption: A Comparison of Regression Analysis, Decision Tree and Neural Networks." *Energy* 32 (9): 1761–8.
[9] Figueiredo, V., F. Rodrigues, Z. Vale, and J. B. Gouveia. 2005. "An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques." *IEEE Transactions on Power Systems* 20 (2): 596–602.
[10] Ashouri, M., F. Haghighat, B. C. M. Fung, A. Lazrak, and H. Yoshino. 2018. "Development of Building Energy Saving Advisory: A Data Mining Approach." *Energy and Buildings* 172: 139–51.
[11] Wang, F., K. Li, N. Duić, Z. Mi, B.-M. Hodge, M. Shafie-khah, and J. P. S. Catalão. 2018. "Association Rule Mining Based Quantitative Analysis Approach of Household Characteristics Impacts on Residential Electricity Consumption Patterns." *Energy Conversion and Management* 171: 839–54.
[12] Sugiyama, M., and K. M. Borgwardt. 2013. "Rapid Distance-Based Outlier Detection via Sampling." In *Advances in Neural Information Processing Systems 26*, 467–75.
[13] Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. *Classification and Regression Trees*. CRC press.
[14] Aggarwal, C. C. 2013. *Outlier Analysis*. Springer.
[15] R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.