The 6th International Conference on Construction Engineering and Project Management (ICCEPM 2015)
Oct. 11 (Sun) ~ 14 (Wed) 2015 • Paradise Hotel Busan • Busan, Korea
www.iccepm2015.org

Comparison of the Performance of Clustering Analysis using Data Reduction Techniques to Identify Energy Use Patterns

Kwonsik Song[1], Moonseo Park[2], Hyun-Soo Lee[3] and Joseph Ahn[4]

Abstract: Identification of energy use patterns in buildings has a great opportunity for energy saving. To find what energy use patterns exist, clustering analysis has been commonly used such as K-means and hierarchical clustering method. In case of high dimensional data such as energy use time-series, data reduction should be considered to avoid the curse of dimensionality. Principle Component Analysis, Autocorrelation Function, Discrete Fourier Transform and Discrete Wavelet Transform have been widely used to map the original data into the lower dimensional spaces. However, there still remains an ongoing issue since the performance of clustering analysis is dependent on data type, purpose and application. Therefore, we need to understand which data reduction techniques are suitable for energy use management. This research aims find the best clustering method using energy use data obtained from Seoul National University campus. The results of this research show that most experiments with data reduction techniques have a better performance. Also, the results obtained helps facility managers optimally control energy systems such as HVAC to reduce energy use in buildings.

Keywords: Energy Use Patterns, Energy Saving, Data Mining, Clustering Analysis, Data Reduction

I. INTRODUCTION

Understanding energy use patterns in building provides a good opportunity to reduce energy use in two aspects. The first one is related to optimizing the operation of building equipment such as HVAC system. In multi-zone buildings, there are several operational strategies to reduce energy use and satisfy occupant's thermal comfort simultaneously [9, 10]. Before controlling HVAC systems, it is necessary to identify when energy use occurs in each room and how much energy each room uses. On the basis of these observations, multi zones can be optimally operated with different system control methods. The second one is obtained by offering more comparative and individualized information about energy use. Energy feedback, one of the methods for occupant intervention, enables users to compare their energy use with that of similar users and affect their energy use behaviors [4, 8]. In case of categorizing users into similar groups based on energy use patterns, a group can receive the following characterized tip for energy saving, e.g., "Time of the day your group used the most energy". In this context, it is necessary to identify what energy use patterns exist in buildings.

In order to reflect this necessity, clustering analysis has been used to find energy use patterns due to its ability to categorize a set of objects into multiple characterized groups. In the course of clustering analysis with energy time-series, data preprocessing has a great impact on clustering performance since it suffers the curse of dimensionality. To solve this problem, it is necessary to reduce the dimensionality of the original data. However, previous efforts rarely account for the high dimensionality of energy use time-series. It remains unclear which technique can provide the best performance when considering data type, purpose and application. To address this issue, this research aims to compare the performance of clustering energy use patterns of different preprocessed datasets. In order to achieve this objective, we preprocess energy use time-series obtained from case buildings in terms of data reduction. With these energy data, we also conduct experiments to compare the performance of clustering the collected data. Clustering performance assessment is based on Silhouette Index (SI).

This paper is organized as follows. Section 2 describes research methodologies used to categorize energy use time-series. Section 3 presents energy use data and preprocesses the collected data. Section 4 apply clustering algorithms to the preprocessed datasets and discusses the clustering performance.

II. RESEARCH METHODOLOGIES

Clustering analysis is the process to find similarities between objects and categorize a set of data objects into the meaningful groups [6]. Basically, objects within a group are similar to one another and different from the objects in other groups. The well performed clustering results show high intra class similarity and low inter class

[1] Ph.D. Student, Dept. of Architecture and Architectural Engineering, Seoul national Univ., Seoul, Korea, woihj@snu.ac.kr
[2] Prof., Dept. of Architecture and Architectural Engineering, Seoul national Univ., Seoul, Korea, mspark@snu.ac.kr
[3] Prof., Dept. of Architecture and Architectural Engineering, Seoul national Univ., Seoul, Korea, hyunslee@snu.ac.kr
[4] Ph.D. Student, Dept. of Architecture and Architectural Engineering, Seoul national Univ., Seoul, Korea, woihj@snu.ac.kr (*Corresponding Author)

similarity. In general, similarity is determined by calculating the distance between two objects. A number of distance measures have been suggested such as Euclidean, Manhattan, Minkowski and Supremum distance. Of these measures, Euclidean distance is commonly used in literature due to its simplicity and applicability [6]. If $p = (p_1, p_2, p_3,..., p_n)$ and $q = (q_1, q_2, q_3,..., q_n)$ are two objects, Euclidean distance is defined as

$$dist(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

In the above equation, object p and q should have non negative values. Based on the distance value, clustering algorithm assigns each object to the group with the closest centroid or link the closest object.

A. Clustering Algorithms

Clustering algorithms most widely employed in literature include *k*-means and agglomerative hierarchical clustering [1, 3, 7, 12].

The *k*-means algorithm is one of partitioning clustering methods and categorizes a dataset with multiple objects into *k* clusters [6]. In *k*-means clustering, the number of clusters should be previously defined before clustering analysis. This clustering algorithm consists of four steps as follow. 1) Arbitrarily determine *k* objects as the initial centroids. 2) Assign each object to the cluster with the closest centroid based on the distance between an object and centroid for each cluster. 3) Replace the current centroid with the object having mean value of each cluster. 4) Iterate step 2 and 3 until there are no more new assignment.

The agglomerative hierarchical clustering hierarchically decomposes the given set of data objects and generates a tree diagram, called a dendrogram [6]. In contrast to *k*-means, this algorithm determines the number of clusters after forming a dendrogram. This algorithm is implemented in four steps as follow: 1) Construct a proximity matrix which is a symmetric matrix representing distance among all objects. 2) Consider each object as one cluster. 3) Merge the two closest clusters until all clusters are merged in a single cluster. 4) Cut the dendrogram at the proper level.

For these algorithms above, users may have a difficulty determining the best number of clusters. Practically, the number of clusters in a given dataset is unknown in real data. To overcome this difficulty, a number of validity indexes have been developed and used for clustering analysis. In this study, we use Silhouette Index (SI) and Davies-Bouldin Index (DBI) proposed in [11] and [13] to evaluate the performance of clustering analysis. For Silhouette Index, the value *s(i)* is given by combining *a(i)* and *b(i)* as follows:

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

where *a(i)*: average dissimilarity of object i to all other objects within the identical cluster and *b(i)*: minimum average dissimilarity of object i to any other clusters. Through the above formula, we obtain the silhouette value ranging from -1 to +1. A high positive value indicates that object i is well assigned to its own cluster. Therefore, if the mean of silhouette values for all objects appears positively high, it should be mentioned that the clustering results are proper.

The Davies-Bouldin Index (DBI) is the mean value of a ratio of inter-cluster and intra-cluster distances. This index can be calculated the following formula:

$$\mathrm{DBI} = \frac{1}{\mathrm{k}}\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{max}_{\mathrm{j}\neq\mathrm{i}}\left\{\frac{\bar{\mathrm{d}}_\mathrm{i} + \bar{\mathrm{d}}_\mathrm{j}}{\mathrm{d}_\mathrm{ij}}\right\}$$

where *k*: the number of clusters and d_i: the average distance between all objects in the i_{th} cluster and the centroid of the i_{th} cluster and d_j: the average distance between all objects in the j_{th} cluster and d_{ij}: the distance between the centroids of the i_{th} and j_{th} clusters. The minimum value of DBI corresponds to good clusters and is regarded as an optimal clustering solution.

B. Clustering Analysis with Time-series Data

Time series data generally changes its values over time and has long length, which means high dimension. In clustering analysis with time series data, the main problem is the sparcity of the available data in high dimensional spaces so that it is impossible to get significant results [5]. Therefore, in the most cases analyzing time-series data, it is highly recommended to reduce the dimensionality of time series. To map the original data into the lower dimensional spaces, Principle Component Analysis (PCA), Autocorrelation Function (ACF), Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) have been commonly used in literature.

PCA converts time-series data into a set of feature vectors called principal components. The extracted features are linearly uncorrelated and its dimension is less or equal to the number of original variables. ACF samples autocorrelation coefficients by measuring the correlation between values of time-series data at time *t* and time *t-n*, where *n = 0, ..., N*. DFT and DWT transform the original data from the time domain into the frequency domain and the time-frequency domain, respectively. These techniques extract a set of new features from the original data using orthogonal function such as sine and cosine for DFT and Haar wavelet for DWT. Unlike DFT and DWT, PCA selects variables of original data based on the magnitude of their coefficients.

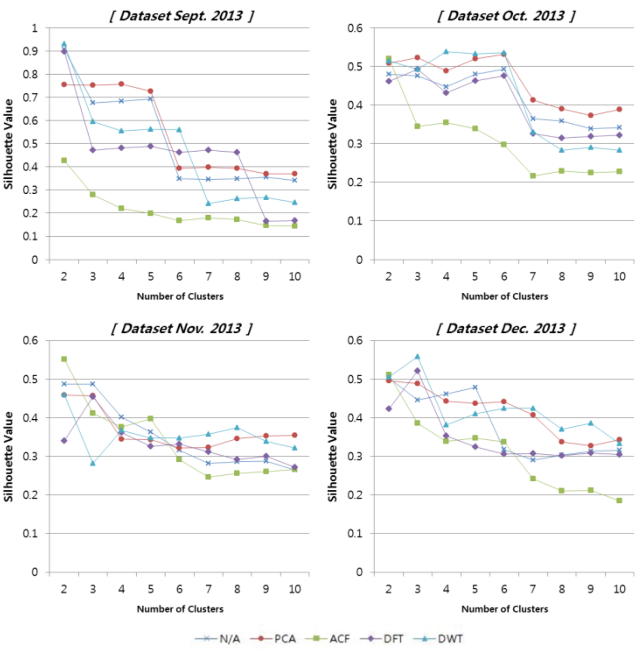


FIGURE I
CLUSTERING RESULTS BY SILHOUETTE INDEX

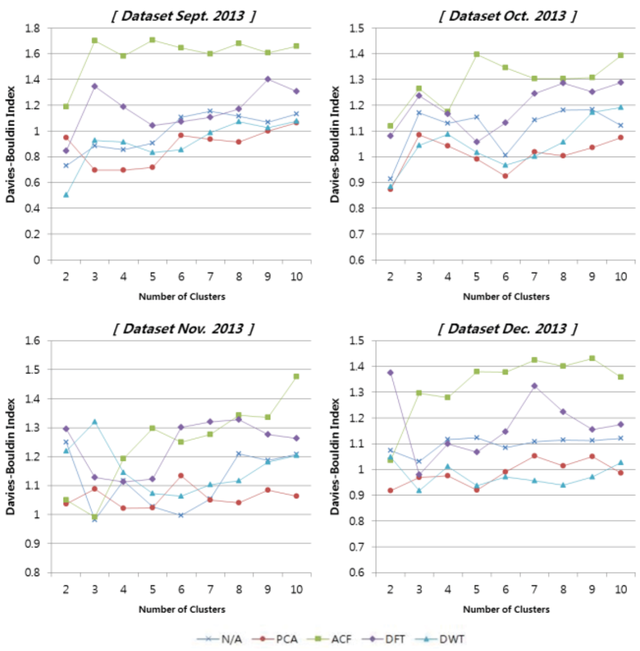


FIGURE II
CLUSTERING RESULTS BY DAVIES-BOULDIN INDEX

III. APPLICATION AND DISCUSSION

We perform experiments using energy use data collected from case buildings to evaluate the performance of clustering analysis with the different preprocessed datasets.

A. Data Collection and Preprocessing

Energy use data was collected in 1,375 dormitory rooms across seven buildings on the Seoul National University campus in Seoul, South Korea. These buildings consist of 250 single rooms and 1,175 double rooms. Each room have energy metering equipment, which facilitates to measure electricity hourly consumed by heating, cooling (only indoor unit), lighting and outlets in each room. This data was collected from March 1, 2013 through February 31, 2014. However, there was a system malfunction from April 24 to May 3 so we excluded energy use data gathered during this period.

In order to see the effect of data type on clustering results, we divided the collected data into 12 monthly datasets and generated the reference profiles for each room by averaging hourly energy use. Especially, in the course of calculating the average hourly energy use, we excluded the energy consumed on weekends since the inconsistent behavior of occupants on weekends could disrupt a generation of the reference profiles. Also, for the comparison of clustering performance, we choose *Dataset Sept. 2013*, *Dataset Oct. 2013*, *Dataset Nov. 2013* and *Dataset Dec. 2013* due to the fact that cooling energy cannot be collected using the energy meter.

Each monthly dataset has 24 columns (hours) so that it is necessity to reduce the dimensionality of the original datasets. To reflect this necessity of data reduction, we adopt PCA, ACF, DFT and DWT to the four monthly datasets. Table I describes the details of the four preprocessed datasets.

TABLE I
FEATURES EXTRACTED USING DATA REDUCTION TECHNIQUES

Data Reduction	*Dataset Sept. 2013*	*Dataset Oct. 2013*	*Dataset Nov. 2013*	*Dataset Dec. 2013*
PCA	8 features	6 features	6 features	5 features
ACF	20 features	20 features	20 features	20 features
DFT	17 features	17 features	17 features	17 features
DWT	12 features	12 features	12 features	12 features

B. Comparison of Clustering Performance

Experiments are carried out to assess the clustering performance using the agglomerative hierarchical clustering method. Although *k*-means algorithms have been widely used for clustering analysis, its results are variable according to the initial cluster centroid position for iterations. Therefore, we select the agglomerative hierarchical clustering with Euclidean distance metrics and Ward linkage. Clustering performance assessment is consequently conducted on the basis of the Silhouette Index (SI) and the Davies-Bouldin Index (DBI).

With this experimental design, we performed clustering experiments using the different preprocessed datasets. From the experiments, we intend to recognize the effect of data preprocessing on clustering analysis with energy use time-series. For each experiment, in order to find the best number of clusters, we make a variation of cluster *k* variable from 2 to 10. In practical, the higher number of clusters leads to having difficulties controlling energy system with various options so we set the limitation of its maximum number to 10 clusters. Figure 1 indicates clustering results for each monthly dataset using values of SI and DBI.

Compared to the clustering results using the original datasets without data reduction techniques, reducing the data dimensionality generally gives the better performance for each monthly dataset (Figure I and II). However, we observed that the application of data reduction doesn't exactly mean the improvement in the performance of clustering analysis. In dataset Nov. 2013, the original dataset without data reduction (N/A) produces the most accurate clustering of energy use profiles when evaluating its performance using the Davies-Bouldin index.

C. *Discussions*

The different results of the clustering analysis performed on the four monthly datasets stem from the fact that data characteristics are fundamentally different (Table II). Also, data reduction techniques have a significant effect to alleviate the curse of dimensionality.

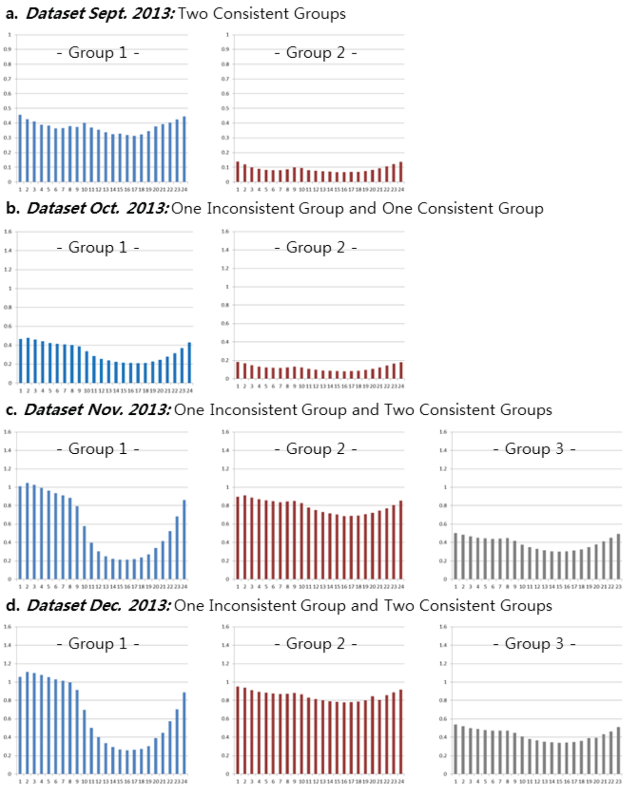
TABLE II
DETERMINATION OF NUMBER OF CLUSTERS BASED ON SI AND DBI

	Dataset Sept. 2013	*Dataset Oct. 2013*	*Dataset Nov. 2013*	*Dataset Dec. 2013*
1) The best performance in clustering analysis using SI				
Data Reduction	DWT	DWT	ACF	DWT
Number of Clusters	2 (0.93)	4 (0.54)	2 (0.55)	3 (0.56)
2) The best performance in clustering analysis using DBI				
Data Reduction	DWT	PCA	N/A	DWT
Number of Clusters	2 (0.50)	2 (0.87)	3 (0.98)	3 (0.92)

Figure III shows energy use patterns from September 2013 to December 2013 in the case buildings. For all the monthly datasets, there are only two representative trends, which are defined as *Inconsistent* and *Consistent Group* in this research. In all monthly datasets, the *Inconsistent Group* means that occupants inconsistently consume electricity by time. For this group, most of energy use occurs between night and dawn. In addition, it appears that they commonly turn off heaters, lighting and equipment when leaving out their rooms. In contrast, occupants in *Consistent Group* use the similar amount of electrical energy every hour as represented in the right profiles. Considering that the occupants living in the case buildings are mostly students, we can also infer that they rarely turn off the equipment before leaving, especially heater during this period.

IV. CONCLUSIONS

Energy use patterns are used for energy saving in buildings in two aspects: 1) optimal system operation and normative energy feedback. In order to identify energy use patterns in buildings, clustering analysis has been commonly used in literatures. From literature reviews, we found that clustering analysis with time-series data has a problem with high dimensionality and it leads to a decrease in clustering performance. To address this problem, it is necessary to preprocess the original data. However, there is a challenging issue about which dimensionality reduction techniques show best performance in clustering analysis. To validate the performance of different preprocessed datasets, we carried out several experiments using real dataset. Compared to the results of clustering analysis using the original datasets, the dimension-reduce datasets generally showed the best performance in categorizing energy use patterns. However, data reduction doesn't always lead to an improvement in clustering performrmance. Therefore, we need to search for the best clustering method by applying various data reduction techniques. From this research, we can provide clustering methods which facilitates to overcome the curse of dimensionality and find out energy use patterns in buildings. In order to improve the availability of the suggested method, we will try to apply this method to different types of buildings.



a. ***Dataset Sept. 2013:*** Two Consistent Groups

b. ***Dataset Oct. 2013:*** One Inconsistent Group and One Consistent Group

c. ***Dataset Nov. 2013:*** One Inconsistent Group and Two Consistent Groups

d. ***Dataset Dec. 2013:*** One Inconsistent Group and Two Consistent Groups

FIGURE III
REPRESENTATIVE ENERGY USE PATTERNS FOR THE FOUR MONTHLY DATASETS

ACKNOWLEDGEMENT

This research was supported by a grant (15CTAP-B080352-02) from Infrastructure and Transportation Technology Promotion Program funded by Ministry of Land, Infrastructure and Transport of Korean government.

REFERENCES

[1] C. Miller, Z. Nagy, A. Schlueter, "Automated daily pattern filtering of measured building performance data", *Automation in Construction*, vol. 49, Part A, pp. 1-17, 2015.

[2] G. Chicco, “Overview and performance assessment of the clustering methods for electrical load pattern grouping” *Energy*, vol. 42, no. 1, pp. 68-80, 2012.

[3] G. Chicco, R. Napoli, F. Piglione, “Comparisons among Clustering Techniques for Electricity Customer Classification”, *IEEE Transactions on power systems*, vol. 21, no. 2, pp. 933-940, 2006.

[4] H. Allcott, “Social norms and energy conservation”, *Journal of Public Economics*, vol. 95, no. 9-10, pp. 1082-1095, 2011.

[5] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, E. Keogh, “Querying and Mining of Time Series Data: Experimental Comparison of Representation and Distance Measures.” *Proceedings of the VLDB Endowment*, vol. 1, no. 2, pp. 1542-1522, 2008.

[6] J. Han, M. Kamber, J. Pei, “Data Mining Concepts and Techniques”, 3rd ed., Waltham: Morgan Kaufmann, pp. 83-124, 2012.

[7] J. Kwac, J. Flora, R. Rajagopal, “Household Energy Consumption Segmentation Using Hourly Data” *IEEE Transactions on smart grid*, vol. 5, no. 1, pp. 420-430, 2014.

[8] M. Iyer, W. Kempton, C. Payne, “Comparison groups on bills: Automated, personalized energy information”, *Energy and Buildings*, vol. 38, no. 8, pp. 988-996. 2006.

[9] M. Mossolly, M. Ghali, N. Ghaddar, “Optimal control strategy for a multi-zone air conditioning system using a genetic algorithm”, *Energy*, vol. 34, no. 1, pp. 58-66, 2009.

[10] N. Nassif, K. Stanislaw, R. Sabourin, “Optimization of HVAC control system strategy using two-objective genetic algorithms”, *HVAC&R Res*, vol. 11, no. 3, pp. 459-486. 2005.

[11] P. J. Rouseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis” *Journal of Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53-65, 1987.

[12] T. Rasanen, J. Ruuskanen, M. Kolehmainen, “Reducing energy consumption by using self-orgnizing maps to create more personalized electricity use information”, *Applied Energy*, vol. 85, no. 9, pp. 830-840, 2008.

[13] Y. H. Zurigat, H. Al-Hinai, B. A. Jubran, Y. S. Al-Masoudi, “Energy Efficient Building Strategies for School Buildings in Oman”, *International Journal of Energy Research*, vol. 27, no. 3, pp. 241-253, 2003.