Exploring the Effect of Data Granularity on Personalized Normative Messaging Interventions for Reducing Household Energy Consumption

K. Song[1]; K. Anderson[2]; S. Lee[3]; K. T. Raimi[4]; and P. S. Hart[5]

[1]Post-Doctoral Research Fellow, Tishman Construction Management Program, Dept. of Civil and Environmental Engineering, Univ. of Michigan, 2350 Hayward St., Ann Arbor, MI 48109. E-mail: kwonsiks@umich.edu
[2]Research Associate, Tishman Construction Management Program, Dept. of Civil and Environmental Engineering, Univ. of Michigan, 2350 Hayward St., Ann Arbor, MI 48109. E-mail: kyleand@umich.edu
[3]Associate Professor, Tishman Construction Management Program, Dept. of Civil and Environmental Engineering, Univ. of Michigan, 2350 Hayward St., Ann Arbor, MI 48109. E-mail: shdpm@umich.edu
[4]Assistant Professor, Gerald R. Ford School of Public Policy, Univ. of Michigan, 735 S. State St., Ann Arbor, MI 48109. E-mail: kraimi@umich.edu
[5]Associate Professor, Dept. of Communication Studies, Program in the Environment, Univ. of Michigan, 105 S. State St., Ann Arbor, MI 48109. E-mail: solhart@umich.edu

ABSTRACT

Normative messaging interventions have proven to be a cost-effective strategy for promoting pro-environmental behaviors. The effectiveness of normative messages is partially determined by how personally relevant the comparison groups are as well as the lag of feedback. Using readily available energy use data has created opportunities to generate highly personalized reference groups based on households' behavioral patterns. Unfortunately, it is not well understood how data granularity (e.g., minute, hour) affects the performance of behavioral reference group categorization. This is important because different levels of data granularity can produce conflicting results in terms of group similarity and vary in computational time. Therefore, this research aims to evaluate the performance of clustering methods across different levels of temporal granularity of energy use data. A clustering analysis is conducted using one-year of energy use data from 3,000 households in Holland, Michigan. The clustering results show that behavioral reference groups become the most similar when representing households' energy use behaviors at a six-hour interval. Computationally, less granular data (i.e., six and twelve hours) takes less time than highly granular data which increases exponentially with more households. Considering the enormous scale that normative messaging interventions need to be applied at, using less granular data (six-hour intervals) will permit interveners to maximize the effectiveness of highly personalized normative feedback messages while minimizing computation burdens.

INTRODUCTION

Residential energy consumption accounts for approximately 21% of all energy expenditures in the United States and over 25% in the Europe Union (EIA 2018; EC 2018). Within residences occupant behavior has a significant impact on energy consumption (Bahaj et al. 2007). Consequently, behavioral interventions targeting home energy use have become an increasingly popular topic for research over the last decade. Energy use feedback interventions, and normative feedback interventions in particular, have been found to be cost effective methods to promote pro-environmental behaviors (Anderson et al. 2017). Energy use feedback inform residents of their consumption over a previous period and normative messages also supply residents with consumption rates of a peer group (e.g., neighbors). More specifically, normative messages provide individual energy consumption information in conjunction with mean energy use data of other households (i.e., social norm).

Across several large-scale experiments in different U.S. cities, it was found that adding normative feedback to energy bills resulted in a 2% reduction in consumption (Allcott 2011). These studies supplied participants with normative messages using generic peer reference groups. While a 2% reduction in use might not seem significant at first, even small reductions in household energy use consumption, 1% for instance, when considered in the aggregate across the residential sector result in enormous net reductions of energy use (1250 tWh) and carbon emissions (150 million metric tons CO_2e). However, it is hypothesized that normative feedback campaigns could create even larger reductions in consumption by creating more personally relevant normative reference groups and increase norm adherence; this in turn would increase the effectiveness of normative messages (Goldstein et al. 2008). The most common personalized comparison groups are based on geographical proximity (e.g., street, city) and housing characteristics (e.g., housing size, heating type), both of which have proven useful in reducing home energy consumption (Darby 2006). Unfortunately, using traditional methods for creating personally relevant comparison groups requires a households' participation to collect the data (e.g., surveys and home energy audits), or prohibitively costly data collection. Thus, in practice, it is currently not financially viable to deploy personalized normative messaging interventions on a large scale.

The use of readily available energy use data offers new opportunities to create highly personalized behavioral reference groups for normative comparisons. For example, smart energy metering technologies can capture energy use data in real-time, making household energy use profiles readily available. As energy use profiles are significantly dependent on households' behaviors, it is possible to classify households into several meaningful groups based on similarities in behavior patterns where energy use norms are created (Richardson et al. 2010). Further, since such energy use profiles can be collected without any active participation from residents it permits the widespread application of personalized normative messaging interventions.

Unfortunately, despite the potential benefit of using readily available data, it is not well understood how the temporal granularity of energy use data affects the performance of behavioral reference group categorization. Understanding which time scale of data aggregation should be applied is important because different levels of data granularity can produce conflicting results in terms of group similarity and vary in computational time (Gouveia et al. 2018). Granell et al. (2015) investigated the impact of data granularity on residential electricity load profiles, but averaged clustering results regardless of the number of clusters. Considering that the best number of clusters is generally unknown in given datasets, it remains unclear how the data granularity affects the clustering performance across all the possible number of clusters.

Therefore, the objective of this research is to evaluate the performance of clustering methods across different level of temporal granularity of energy use data. A data mining-based behavioral reference group categorization framework is constructed to classify objects into several meaningful groups (Han et al. 2012) and perform a clustering analysis across various levels of data granularity.

BEHAVIORAL REFERENCE GROUP CATEGORIZATION FRAMEWORK

To classify households into several meaningful behavioral reference groups based on daily energy use profiles, a data mining-based categorization framework is proposed. The proposed categorization framework has three major components: 1) data preprocessing, 2) application of clustering algorithms and 3) clustering performance evaluation (Fig. 1).

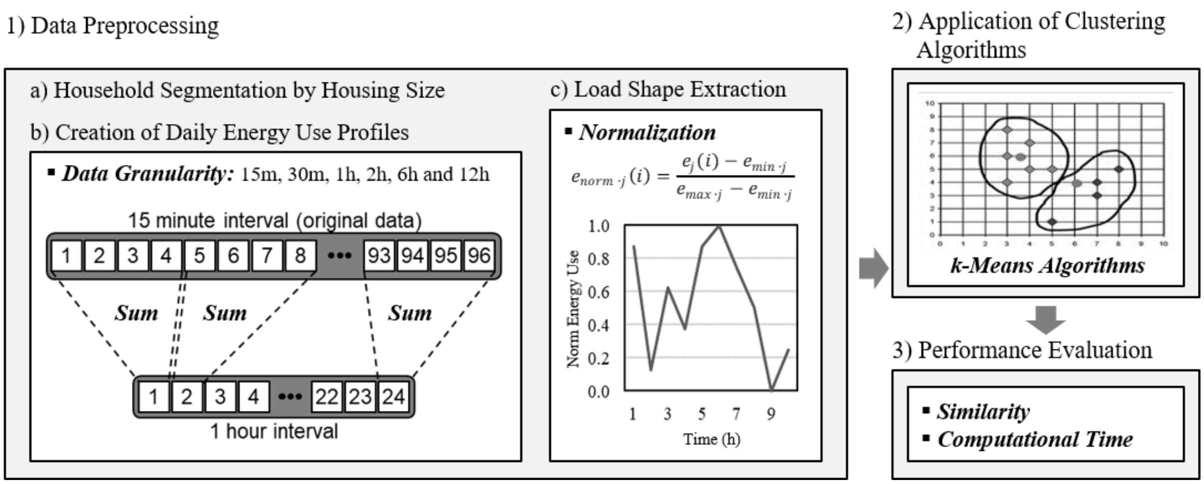

Figure 1. Framework of Behavioral Reference Group Categorization

DATA PREPROCESSING

The first step of data preprocessing is to separate households into five housing size categories. Since housing size is one of the most significant determinants of household energy use, this enhances the validity of the normative refence groups by comparing energy consumption of households with similar size (Santin et al. 2009). Also, housing size data is readily available from public record, so it is possible to create behavioral reference groups in a non-invasive way. Further, it is assumed that houses of the similar size in neighborhoods near each other with build dates are likely to comparable in energy efficiency because the data on home energy efficiency is not available or unknown. The second step is to create typical daily energy use profiles for households by averaging the values corresponding to the same time of all days of 1 week. The time intervals investigated include: 15 minutes, 30 minutes, one hour, two hours, six hours, and twelve hours. The last step is to extract load shapes from the typical daily energy use profiles by normalizing energy consumption at each time to a value between 0 and 1. This is important because households in a behavioral reference group should have similarities in the time of energy consumption (i.e., load shapes) but variations in the amount of energy consumption for normative comparison. Considering both the load shape and volume of daily energy use profiles during the clustering process makes it possible to categorize differing profiles into the same group as similarity is determined based on distance between data points (Chicco 2012). In order to overcome categorization problems, Min-Max normalization was applied. The normalized value of energy $e_j(i)$ spent at time i is calculated as follows:

$$e_{norm\cdot j}(i) = \frac{e_j(i) - e_{min\cdot j}}{e_{max\cdot j} - e_{min\cdot j}} \quad \text{(Eq. 1)}$$

where, $e_j(i)$ is the existing value of energy consumed at time i, $e_{min\cdot j}$ is the minimum amount of energy consumption for household j, $e_{max\cdot j}$ is the maximum amount of energy consumption for household j.

APPLICATION OF CLUSTERING ALGORITHM

A k-means clustering (KC) algorithm is applied to the preprocessed dataset to create personalized behavioral reference groups. The KC is a partitioning clustering method which distribute objects into k clusters (Han et al. 2012). The KC begins with k initial cluster centers which are arbitrarily determined as the initial centroids of groups. Each object is assigned to the closest cluster. Given that the difference between an object p and the centroid c_i of a cluster C_i is measured by Euclidean distance, the quality of clusters can be evaluated by the following objective function E:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \qquad \text{(Eq. 2)}$$

This objective function attempts to create the k clusters as compact and as divergent as possible. The KC updates the new centroids of clusters and evaluates them based using the objective function repeatedly until there are no new variations in similarities.

CLUSTERING PERFORMANCE EVALUATION

After applying the KC, all the results are evaluated to find the best number of behavioral reference groups for the given dataset. Here, the Davies-Bouldin Index (DBI) is adopted as a cluster validity index. The DBI calculates the mean value of the ratio of within-cluster scatter to between-cluster separation (Davies and Bouldin 1979). A lower value of DBI indicates better clusters and a better clustering method. The DBI is calculated using the following equation:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max_{j \neq i} \left\{ \frac{\overline{d_i} + \overline{d_j}}{d_{ij}} \right\} \qquad \text{(Eq. 3)}$$

where, k is the number of clusters, d_i is the average distance between all objects in the i^{th} cluster and the centroid of the i^{th} cluster, d_j is the average distance between all objects in the j^{th} cluster and d_{ij} is the distance between the centroids of the i^{th} and j^{th} clusters.

DATA COLLECTION AND CLEANING

Data was collected from 3,000 households in Holland, Michigan. All residences are all equipped with smart meters which capture electrical energy use data at a 15-minute interval. Electricity data was collected from January 1, 2016 through December 31, 2016. In addition to the electricity data, housing size data was collected from public records. Energy use data for 503 households were excluded from analysis due to erroneous data–missing or abnormal values, residents moved, or smart meter failure. Additionally, residence size was unobtainable for 319 households resulting in a total sample size of 2,248 households (Table 1).

Table 1. Number of Samples by Housing Size

	Housing Size Category				
	HS1	HS2	HS3	HS4	HS5
Floor Area (ft^2)	~1,000	1,000~1,500	1,500~2,000	2,000~2,500	2,500~
# of Samples	525	1,083	435	132	73

RESULTS AND DISCUSSION

A clustering analysis was conducted using the proposed data mining-based categorization framework to investigate how the temporal granularity of energy use data affects the performance of behavioral reference group categorization. In order to set the same condition of clustering analysis, 73 households were randomly selected from each housing size category. Then, all the clustering results (i.e., DBI and computational time) were averaged depending on data granularity.

The lowest average values of DBI were found when daily energy use profiles are represented using six-hour intervals except for when households are categorized into nine behavioral reference groups (Figure 2-a). The average values of DBI tend to increase for data with lower temporal granularity. These clustering results can be explained by the following two conflicting facts. First, as the dimensionality of data (i.e., data attributes) increases, the number of data points that make up a group become increasingly sparse (Ding et al. 2008). Thus, it is difficult to create meaningful groups in a high dimensional space. Second, using a small number of data attributes makes it difficult to distinguish between groups.

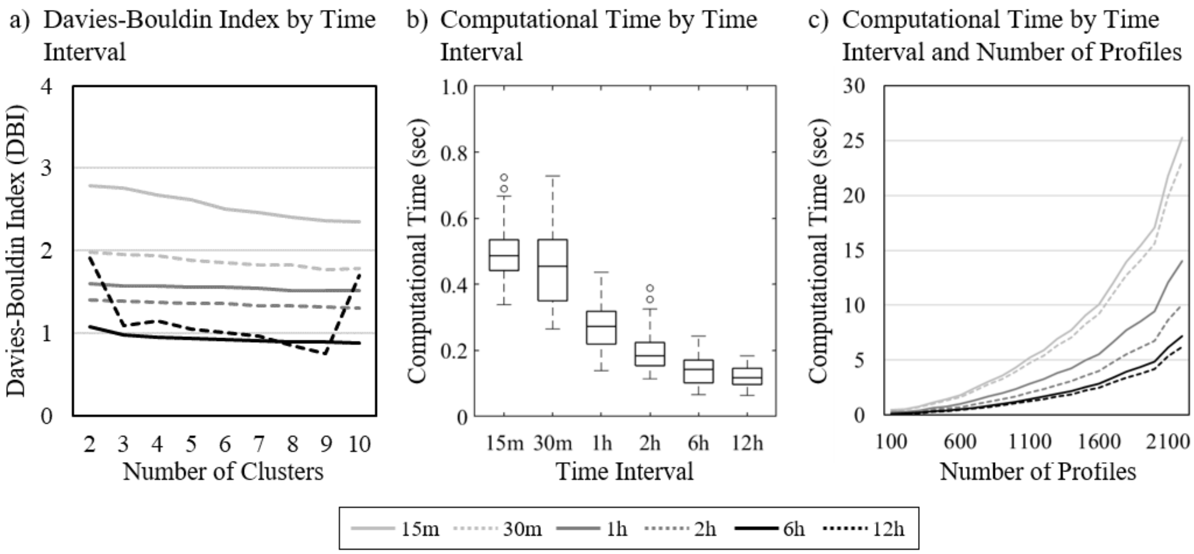

Figure 2. Average values of DBI and computational time by time interval

When investigating the average computational time by temporal granularity of energy use data, it was observed that creating behavioral reference groups using less granular energy data (i.e., six and twelve hours) takes less time than highly granular data (Figure 2-b). Further, it was discovered that the difference in computational time by data granularity increase significantly with more households (Figure 2-c). This significant increase does not concur with the previous opinions that the time complexity of KC is linearly correlated with the number of data objects (Hartigan and Wong 1979). This gap in the literature can be caused by the number of data objects and the number of iterations. With the small number of data objects, clustering results are shortly converged to the global optimum after a dozen iterations (Vattani 2011). However, as the number of data objects increases, the KC can be very slow to converge due to the larger number of iterations. For this reason, recent studies have suggested polynomial or superpolynomial time complexity of KC (Arthur and Vassilvitskii 2006).

Representing households' energy use behaviors at a six-hour interval will allow interveners to create more personalized normative comparison groups while minimizing computational burdens (Figure 2-b). While the computational time is nominal (a few seconds to a few minutes) for analysis of datasets with a few thousand to tens of thousands of residences, this will not be true when using much larger datasets with millions of homes. Computational time increases exponentially with the number of residences and can potentially become a consideration when analyzing a significant number of homes. Especially, since using insufficient group members can undermine the validity of normative comparisons within the group, it is inevitable to cluster a large number of residential populations for personalized normative messaging intervention. For instance, categorizing one million homes into ten behavioral reference groups will take approximately 400 times longer than categorizing ten thousand homes: 58 minutes compared to 8.7 seconds (time complexity of *k*-means proposed by Vattani 2011). If energy use feedback should be supplied every week, computational time will need to be considered when determining which level of data granularity should be used. Further, since households' energy use behaviors can change over time (e.g., week, month, season), it is necessary to update behavioral reference groups every billing period. Therefore, computational time would be a may be a consideration during intervention periods of a large number of residential populations.

CONCLUSION AND FUTURE RESEARCH

In this study we evaluated the performance of clustering methods at different levels of temporal granularity of energy use data. It was found that behavioral reference groups become the most similar when representing households' energy use behaviors at a six-hour interval. Computationally, data granularity is not an important factor when clustering tens of thousands of homes. However, it is observed that with a large volume of households, less granular data (i.e., six and twelve hours) contributes to minimizing the latency of normative behavioral feedback.

This research contributes to the literature by enhancing our knowledge of how the temporal granularity of energy use data affects the similarity of behavioral reference groups in terms of energy use profiles. In addition, this research contributes to the literature by developing a data mining-based categorization framework which permits interveners to maximize the effectiveness of highly personalized normative feedback messages while minimizing computational burdens. Future research efforts should consider the aggregation level of energy use data while creating behavioral reference groups in order to capture typical energy use profiles by reducing the variability in energy consumption due to irregular schedules (e.g., business trips).

ACKNOWLEDGEMENT

The work presented in this paper was supported financially with a National Science Foundation Award (No. CBET-1705273). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation

REFERENCES

Allcott, H. (2011). "Social norms and energy conservation." *J. Public Econ.*, 95, 1082-1095.

Anderson, K., Song, K., Lee S., Krupka, E., Lee, H. and Park, M. (2017). "Longitudinal analysis of normative energy use feedback on dormitory occupants." *Appl. Energy*, 189, 623-639.

Arthur, D. and Vassilvitskii, Sergei. (2006). "How slow is the k-means method?." *Proc. Twenty-second Annu. Symp. Comput. Geom.*, New York, NY, 144–153.

Bahaj, A. S., Myers, L. and James, P. A. B. (2007). “Urban energy generation: Influence of micro-wind turbine output on electicity consumption in buildings.” *Energy Build.*, 39(2), 154–165.
Chicco, G. (2012). “Overview and performance assessment of the clustering methods for electrical load pattern grouping.” *Energy*, 42(1), 68-80.
Darby, S. (2006). “The effectiveness of feedback on energy consumption: A review for DEFRA of the literature on metering, billing and direct displays.” Environmental Change Institute, University of Oxford, (April).
Davies, D. L. and Bouldin, D.W. (1979). “A cluster separation measure.” *IEEE Trans. Pattern Anal. Mach. Intell.,* 2, 224–227.
Ding, H., Trajcevski, G., Scheuermann, P., Wang, X. and Keogh, E. (2008). “Querying and Mining of Time Series Data: Experimental Comparison of Representation and Distance Measures.” *Proc. VLDB Endow.,* 1(2), 1542-1522.
Energy Information Administration (EIA). (2018). “Total Energy: Energy Consumption by Sector.” <https://www.eia.gov/totalenergy/data/monthly/index.php> (Nov. 18, 2018).
European Commission (EC). (2018) “Statistical pocketbook 2018” <https://ec.europa.eu/energy/sites/ener/files/documents/PocketBook_ENERGY_2015%20PDF%20final.pdf> (Nov. 18, 2018).
Festinger, L. (1954). “A theory of social comparison processes.” *Human Relations,* 7(2), 117-140.
Goldstein, N. J., Cialdini, R. B. and Griskevicius, V. (2008). “A room with a viewpoint: Using social norms to motivate environmental conservation in hotels.” *J. Consum. Res.*, 35(3), 472–482.
Gouveia, J. P., Seixas, J. and Long, G. (2018). “Mining households' energy data to disclose fuel poverty: Lessons for Southern Europe.” *J. Clean. Prod.*, 178, 534-550.
Granell, R., Axon, C. J. and Wallom, D. C. H. (2015). “Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles.” *IEEE Trans. Power Syst.,* 30(6), 3217-3224.
Han, J., Kamber, M. and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Waltham, MA.
Hartigan, J. A. and Wong, M. A. (1979). “Algorithm AS 136: A K-Means Clustering Algorithm.” *J. R. Stat. Soc.*, 28(1), 100-108.
Richardson, I., Thomson, M., Infield, D. and Clifford, C. (2010). “Domestic electricity use: a high-resolution energy demand model.” *Energy Build.*, 42(10), 1878–1887.
Santin, O. G., Itard, L. and Visscher, H. (2009). “The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock.” *Energy Build.*, 41(11), 1223-1232.
Vattani, A. (2011). “*k*-means requires exponentially many iterations even in the plane.” *Disc. Comput. Geom.*, 45(5), 596-616.