Semantic search in household energy consumption segmentation through descriptive characterization

Milad Afzalan
Virginia Tech
Blacksburg, VA, USA
afzalan@vt.edu

Farrokh Jazizadeh
Virginia Tech
Blacksburg, VA, USA
jazizade@vt.edu

ABSTRACT

With the widespread adoption of smart metering infrastructures, household energy consumption segmentation is receiving increasing attention. The objective is to transform the large volume of household daily load shapes into representative patterns through clustering methods, with the aim of program targeting and customer engagement. In the literature, there exists a high variation in the number of clusters that different studies have adopted. In order to address the challenge in the trade-off between cluster accuracy and ease of interpretation, in this paper, we introduce a data-driven characterization scheme for resultant clustered load shapes, with the aim of facilitating information retrieval of load shapes with specific semantic attributes. The characterization scheme extracts descriptive features from load shapes to explain their temporal pattern. Using segmentation results on a sample data set from Pecan Street Dataport, we show the feasibility of obtaining the semantic representation of load shapes and performing query analysis by accounting for their similarities. Furthermore, as an application case study, we demonstrated the identification/retrieval of suitable households with specific load types for the adoption of PV-battery system, with average self-sufficiency of 80%.

CCS CONCEPTS

• **Hardware** → **Smart grid**; • **Information systems** → *Clustering*; **Mathematics of computing** → *Time series analysis*;

KEYWORDS

Segmentation, clustering, demand response, data classification

ACM Reference format:

Milad Afzalan and Farrokh Jazizadeh. 2019. Semantic search in household energy consumption segmentation through descriptive characterization. In *Proceedings of 6th ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys ' 19), New York, USA, November 2019*, 4 pages. https://doi.org/10.1145/3360322.3360865

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

BuildSys '19, November 13–14, 2019, New York, NY, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7005-9/19/11...$15.00
https://doi.org/10.1145/3360322.3360865

1 INTRODUCTION

Smart metering at the national scale has provided unprecedented bulk of temporal data for residential customers with opportunities for energy analytics. Accordingly, future smart grid accounts for consumption patterns of customers to improve the efficient operation of the power system with the high penetration of distributed energy resources (DERs). To this end, load shape clustering can be carried out for the task of customer segmentation to reveal the high variation in consumption patterns [1], which could unleash potential applications such as identification of customers suitable for demand response (DR) [2], integration of renewables [3], and improved load forecasting [4]. In segmentation, an important factor is the proper selection of the number of clusters. A low number can lead to a coarse-level representation of load shapes and accuracy loss by obtaining cluster centroids that are not necessarily representative of their members. On the other hand, a larger number of clusters could increase accuracy to better reveal distinguishable load shape patterns. However, a larger number of clusters poses challenges on the interpretability of segmentation and derivation of meaningful implications for energy planning.

For studies with large-scale data sets (# of load shapes>20K- e.g., [2, 5-7]), high variations in selecting representative clusters have been observed from low numbers (i.e., 3 for ~2 million load shapes [8] and 8 for ~700K load shapes [6]) to high numbers (e.g., ~270 for ~66 million load shapes [2]) for household energy use. Majority of the previous research efforts, with limited clusters, have *not* presented proper metrics (e.g., error in consumption or peak overlap) to evaluate the representativeness of cluster centroids for their associated profiles, while these metrics are necessary to quantify the accuracy of the segmentation [3]. Higher accuracies are specifically critical for the analysis of household energy usage and applications like DR, integration of renewables, leveraging smart and flexible loads, and tariff determination. However, in large-scale data sets, the higher accuracy is concomitant with a larger number of clusters, which makes deriving meaningful insights more difficult. As an example, Kwac et al. [2] developed a segmentation scheme to create a library of load shapes for analysis of household lifestyle. They showed a large sample of load shapes (~66 million) can be reasonably encoded into a library of more than 270 representative patterns. However, given the large size of the resultant library in large datasets, it could be still difficult to use the clustered data for further analyses and interpretations.

2 TOWARDS QUERY ANALYSIS IN SEGMENTATION

Considering the aforementioned problem, in this study, we have introduced a characterization scheme for facilitated interpretation of the resultant clusters for load shape segmentation to enable query analysis for information retrieval. The characterization (i.e., semantic labeling) scheme could be used for retrieving clusters with similar semantic characteristics, which are difficult to be carried out through the mathematical process of clustering. Such characteristics (see Figure 1) are defined as descriptive features based on the magnitude or the pattern of load shapes (such as the timing of peak consumption) with the following advantages:

- Providing semantic information on segmentation results could enable high accuracy of clustering while facilitating the information retrieval from large segmentation libraries for targeted and customer-centered energy program design.
- They enable query analysis for specific applications. As an example, one could query households with double peak energy distribution at noon (when solar energy is abundant) and evening (when wholesale electricity market price is higher) to be engaged in programs for adopting PV (photovoltaic)-battery system.

Upon defining the characterization scheme, using a sample data set from Pecan Street Dataport [9], we have demonstrated the applicability of the method on household segmentation results. Furthermore, as an example scenario, the applicability of using the semantic characterization for identifying the proper household to adopt PV-storage system through query analysis has been presented.

3 SEMANTIC CHARACTERIZATION OF CLUSTERED LOAD SHAPES

Given a set of clustered load shapes (C_i, $i \in [1{:}K]$; K is the number of clusters) obtained from N daily time-series load shapes $S = \{s_1(t), s_2(t), \ldots, s_N(t)\}$, the objective is to extract a sample of descriptive features for C_i that contain semantic information about load shapes. Different descriptive features (as visually shown in Figure 1) can be extracted from the cluster centroid. Each feature describes an aspect related to temporal shape or magnitude of consumption, which collectively characterizes a cluster. In other words, we characterize the level of consumption, the demand distribution pattern, and the timing, severity, and duration of peak demand. Through empirical observations, we have defined a set of features that could provide a semantic characterization of clusters. Therefore, we have established the baseline for feature selection while taking into account that the complete feature analysis requires a comprehensive investigation. The features, used in this study, are described as follows:

- **Energy consumption level (L,M,H)**: Some load shapes are associated with high energy demand while others are considered moderate or low demand. Therefore, the energy demand is considered as an indicator of variability of load shapes. The clusters could, therefore, be labeled as low (L), medium (M), and high (H). Upon segmentation to K clusters, the energy demand for each one is calculated by numerical integration of the load shape curve ($\int_{j=1}^{T} s_j(t)$) (see Figure 1, attribute 1). Considering 25th and 75th quantiles for the obtained values, clusters lower than 25th are labeled as L, higher than 75 labeled as H, and those between this range are labeled as M.
- **Demand distribution pattern (NM,UM,BM,MM)**: One important aspect of the segmentation process is the peak occurrence, specifically at times of grid congestion. Accordingly, load shapes could be divided into a few categories: (*i*) no modes (indicating baseline loads or a consistent consumption magnitude) (NM), (*ii*) unimodal (UM), (*iii*) bimodal (BM), and in less common cases (*iv*) multimodal (MM) (Figure 1, attribute 2). A data point is marked as a local peak if it is larger than its two adjacent neighboring points and its *prominence* is higher than the threshold τ. The prominence of the peak (the vertical line in Figure 1, attribute 4) is identified as follows: (*i*) From the peak location, a horizontal line is extended to both sides till it either reaches the end of the daily profile or intersects with a higher peak, (*ii*) the minimum of the daily profile on both sides of the intervals in the last step is selected, (*iii*) the higher of the two minimum values defines the reference level, in which the relative height of the peak compared to the reference is the prominence. The thresholds τ for each group of (L,M,H) are fixed as constant values, which are selected based on empirical observations on the power demand magnitude of household energy use.
- **Peak timing (MN,M,N,AN,E,NT)**: The temporal location of peak demand is of importance for different applications such as DR and the integration of renewable resources. We consider the timeframes of midnight (MN from 00:00 to 6:00), morning (M from 6:00 to 11:00), noon (N from 11:00 to 14:00), afternoon (AN from 14:00 to 17:00), evening (E from 17:00 to 20:00), and night (NT from 20:00 to 24:00) for the timing of peak $P(t)$ (Figure 1, attribute 3).
- **Peak intensity (NS,MS,S)**: The intensity of peak indicates its relative energy demand, compared to adjacent timeframes. Accordingly, sharper peaks could imply the simultaneous operation of high power-draw appliances or intense use of air conditioning systems (compare two peaks in Figure 1, attribute 4) which make them suitable for DR. Considering the centroid of a cluster C_i as μ_i and the standard deviation of its data points as σ_i, the peak $P(t)$ is labeled as non-significant (NS) if $P(t) < \mu_i + \sigma_i$, moderately significant (MS) if $\mu_i + \sigma_i < P(t) < \mu_i + 2\sigma_i$, and significant (S) if $P > \mu_i + 2\sigma_i$.

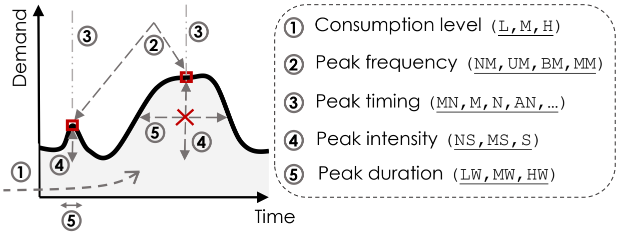
Demand
③
②
③
①
④
⑤
④
⑤
Time
① Consumption level (L,M,H)
② Peak frequency (NM,UM,BM,MM)
③ Peak timing (MN,M,N,AN,…)
④ Peak intensity (NS,MS,S)
⑤ Peak duration (LW,MW,HW)

Figure 1: Sample of a clustered load shape and its representative extracted features

- **Peak duration** (**LW,MW,HW**): The peak duration has implications for effective control of demand (compare two peaks in Figure 1, attribute 5). Accordingly, due to the requirements of DR events which may need different timespans, low width (LW) peak is considered to last for less than 1 hour, medium width (MW) has a duration between 1 to 2 hours, and high width (HW) has a duration of more than 2 hours. Width is defined as the length of the horizontal line drawn at the half prominence level, enclosed by other data points in the daily profile (see Figure 1, attribute 5).

Through determining the abovementioned features, the clustered load shapes are semantically labeled by concatenating the features in a sequence. For example, for the representation in Figure 1, assuming medium level consumption, bimodal energy distribution with peak times in the morning and evening, the peak intensity of non-significant and significant, and the peak duration of less than 1 hour and several hours, the cluster is characterized as (M)-(BM)-(MN,E)-(NS,S)-(LW,HW).

4 CASE STUDY ANALYSIS

We show the performance of the characterization method on the segmentation in addition to discussing a preliminary implication scenario for identifying households for PV-battery adoption.

4.1 Dataset

We considered the sample of households from the Pecan Street Dataport project [9] with available data in July and August 2016. After pre-processing for creating load shapes and performing median filtering to reduce the noise impact, our sample included 436 households with a total number of 27032 daily profiles. Each daily profile contains 15-minute resolution data (96 per daily profile).

4.2 Segmentation and load shape characterization

For segmentation, different clustering algorithms such as self-organizing map (SOM), K-means, and adaptive K-means can be applied (see e.g., [2, 3, 6]). Due to its capability for electricity load segmentation, we employed SOM in this study. Using the within sum of square error objective function and evaluating the error from the elbow curve based on different ranges of k, the daily profiles were segmented into k =120 clusters. For each resultant cluster, the descriptive features were extracted. Figure 3 shows example clusters and their semantic labels (shown above each subplot). As this figure shows, the characteristics of temporal shapes and magnitude of load shapes in each case have been successfully retrieved by our approach. Therefore, the proposed features could be used in query analysis of large cluster libraries in consumer segmentation. To demonstrate the query analysis capabilities, we performed an analysis to identify clusters that share similar semantic features while using a large number of clusters. Figure 4 shows examples in the segmentation library in which different load shapes that follow similar extracted features are retrieved from the dataset.

4.3 PV-battery adoption case study

As a case example, we considered identifying households that could potentially benefit more from installing PV-battery systems according to their segmentation results. To this end, we retrieved labels from the library that identify such households. The labels were defined based on similarity to the pattern of PV generation and/or benefiting battery implementation (i.e., peak demand after PV generation diminishes). To evaluate the suitability of the identified households for PV-battery adoption, we used the commonly used *self-sufficiency* metric as follows:

$$\textit{self-sufficiency} = \frac{M}{P} \tag{1}$$

in which M is the self-consumed energy (amount of PV energy used at the time of generation and its surplus value stored in the battery and used later) and P is the total energy load of household. The PV system for each household was simulated with NREL PV Watts platform [10]. Using the data for July and August for Austin, TX (where data was collected), hourly PV generation was extracted and interpolated in 15-minute resolution. An initial DC system size of 5kW was assumed. Using the ratio of the total load to the total PV generation as a weight factor, the PV output at each time was recalculated to make the PV generation tailored to each household [3]. For battery, the energy capacity of 14kWh and peak power of 7kW for charging/discharging was assumed,

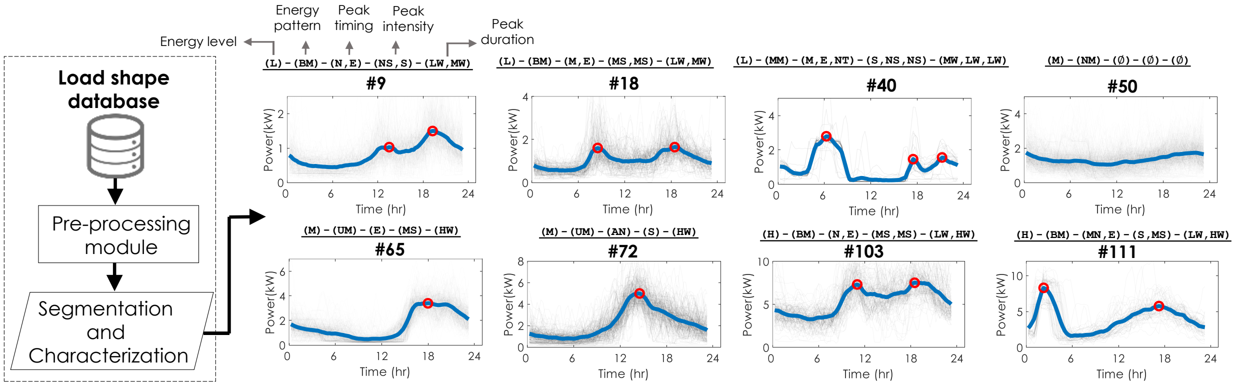

Figure 2: Examples of obtained clusters from segmentation and the extracted semantic labels. For each subplot, the centroid of the cluster is shown in blue, the associated daily profiles are shown as wiggly grey lines, and detected peaks are shown with red circles. The cluster number and the extracted semantic label are depicted on top of each subplot.

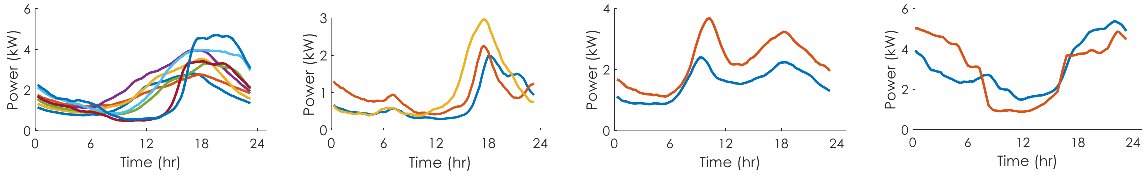

Figure 3: Samples of retrieved clusters from segmentation with similar descriptive features

similar to the specification of Tesla Powerwall 2. To simulate battery storage management, we used a simplified model [11]. The battery is charged when PV generation is higher than demand and discharged later when demand is higher than PV. Figure 4 shows an example of result over 4 days for 1 household in which demand is the ground truth data and generation and battery outputs were simulated. Following this procedure, the *self-sufficiency* for each household over 2 months of data was calculated.

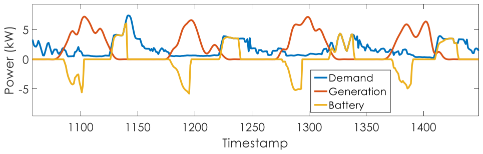

Figure 4: Demand, PV generation, and battery storage (charging and discharging are shown with – and + values).

Realistically, for subsequent days, each household consumption style changes and might be attributed to different clusters (not necessarily one specific cluster). Therefore, to identify appropriate households for PV-battery system, the clusters mode (i.e., cluster with highest occurrence) over the 2-month was assigned. Then, we retrieved households that were associated with clusters with either of these characteristics: **(1)** Unimodal energy distribution pattern (UM) with morning (M) or noon (N) as the peak time (see cluster #72 in Figure 2 as an example), or **(2)** Bimodal energy distribution pattern (BM) with combination of peak timing of morning (M) or noon (N) as the 1st peak, and noon (N), afternoon (AN), evening (E), or night (NT) as the 2nd peak (see the 3rd plot in Figure 3 as an example). Such characteristics in clusters were assumed to align better with the pattern of PV generation or PV-battery management. Other features (i.e., energy level, peak timing, and peak duration) were allowed to take different values due to their lower impact in this example scenario.

Using the above queries, a sample of households were identified. We measured the self-sufficiency for each household belonging to (1) identified sample described above and (2) the complementary sample (rest of households). Figure 5 compares the distribution of the self-sufficiency metric. As shown, the identified samples from query analysis are better prosumers given their self-consumption outcome. Specifically, the self-sufficiency of the identified group shown to be 80% while it was 72% for the complementary group (*p-value*=2*e-14 from two-sample t-test). Furthermore, the distribution shows that a considerable number of households in the identified sample achieves the self-sufficiency of above 80%.

5 Conclusion

In this paper, we introduced a semantic characterization scheme for household energy segmentation result. Different features based on magnitude and temporal pattern of load shapes were defined and extracted that altogether provide semantic information on segmentation results. The objective was to facilitate the inference and interaction with segmentation library for the large-scale database of households. An evaluation on the Pecan Street project showed the applicability of the method for successful characterization of segmentation results. A case example was presented to show the applicability of identifying suitable household for PV-battery adoption. Future directions of this research comprise of: (1) comprehensive analysis for different demand-side management implication such as tariff determination and (2) presenting quantified results for segmentation performance through merging clustered load shape with similar features.

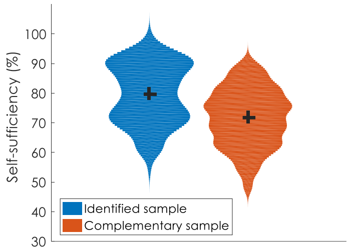

Figure 5: Distribution of self-sufficiency for identified sample (N=153 households) and complementary sample (N = 283 households)

REFERENCES

[1] M. Afzalan and F. Jazizadeh, "Residential loads flexibility potential for demand response using energy consumption patterns and user segments," *Applied Energy,* vol. 254, p. 113693, 2019.

[2] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Transactions on Smart Grid,* vol. 5, no. 1, pp. 420-430, 2014.

[3] S. Xu, E. Barbour, and M. C. González, "Household segmentation by load shape and daily consumption," in *Proc. ACM SigKDD 2017 Conf. Halifax, Nov. Scotia, Canada, August 2017*, 2017.

[4] E. Barbour and M. González, "Enhancing household-level load forecasts using daily load profile clustering," in *Proceedings of the 5th Conference on Systems for Built Environments*, 2018, pp. 107-115: ACM.

[5] T. Teeraratkul, D. O'Neill, and S. Lall, "Shape-based approach to household electric load curve clustering and prediction," *IEEE Transactions on Smart Grid,* vol. 9, no. 5, pp. 5196-5206, 2018.

[6] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Applied energy,* vol. 141, pp. 190-199, 2015.

[7] S. Iyengar, S. Lee, D. Irwin, and P. Shenoy, "Analyzing energy usage on a city-scale using utility smart meters," in *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*, 2016, pp. 51-60: ACM.

[8] J. Y. Park, X. Yang, C. Miller, P. Arjunan, and Z. Nagy, "Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset," *Applied Energy,* vol. 236, pp. 1280-1295, 2019.

[9] Source: Pecan Street Inc. Dataport, 2017.

[10] N. P. Watts. Available: https://pvwatts.nrel.gov/pvwatts.php

[11] R. Luthander, J. Widén, D. Nilsson, and J. Palm, "Photovoltaic self-consumption in buildings: A review," *Applied energy,* vol. 142, pp. 80-94, 2015.