IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 5, 2019, accepted August 17, 2019, date of publication August 29, 2019, date of current version September 13, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2938212*

# Net Electricity Clustering at Different Temporal Resolutions Using a SAX-Based Method for Integrated Distribution System Planning

**MICHAEL EMMANUEL, (Member, IEEE), AND JULIETA GIRALDEZ, (Member, IEEE)**
National Renewable Energy Laboratory, Golden, CO 80401, USA

Corresponding author: Michael Emmanuel (michael.emmanuel@nrel.gov)

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Solar Energy Technologies Office under Contract No. DE-AC36-08GO28308 with Alliance for Sustainable Energy, LLC, the operator of the National Renewable Energy Laboratory, and Hawaiian Electric Companies under Agreement CRD-16-00630. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

**ABSTRACT** This paper addresses a major utility and regulator concern of characterizing customer net electricity consumption profiles to realize integrated distribution system planning. This is pivotal in assessing the capability of the power system to accommodate net load variability and its impacts on the grid such as voltage rise, narrowing peak demand duration, and reducing the cost of energy storage. Although the extant literature has focused on load clustering, this paper uses a symbolic aggregate approximation-based (SAX-based) dimensionality- reduction and k-means techniques to cluster net consumption of smart meter data for more than 3500 residential customers in a month at different temporal resolutions. This study proposes the use of cumulative explained variance in the principal component analysis to determine the optimal number of segments and dimensionality of the transformed space during discretization while retaining the data integrity instead of using intuition, as proposed by the extant literature. Also, this paper describes a screening methodology to determine the distribution of high-voltage customers among the resulting clusters of customers with and without on-site solar photovoltaic generation at different time resolutions.

**INDEX TERMS** Symbolic aggregate approximation, clustering, net electricity consumption, principal component analysis.

## I. INTRODUCTION

The integration of distributed energy resources (DERs), such as wind and solar photovoltaic (PV) systems, with the electric power system is projected to increase at an unprecedented rate [1], [2]. This is largely driven by increased taxation of greenhouse gas emissions, DER technology improvement, and business model innovations [3]–[5].

The penetration of DERs into the electric power system, however, introduces a changing paradigm because the demand that utilities need to plan for is no longer only the consumption of residential and commercial loads-, but also the difference between the loads consuming power and the generation of customer-sited DERs, which is called the "net load". The net load, defined here as the total normal load demand minus the DER generation, gives the demand that must be met by the traditional, dispatchable generation and incorporated into the evolving, integrated distribution system planning. Managing power system dynamics requires the consideration of the net load profile and variability coupled with its impacts, such as voltage rise, reducing the cost of energy storage, shifting and narrowing the system peak demand period [6]–[8].

A more active distribution system is increasing the need to perform electric distribution studies with more realistic characterization of customer loads to study the effect of load diversity with DERs. A key component in coming up with representative load profiles is the application of clustering techniques to load data from smart meters. The extant literature reports load time-series data clustering using widely studied techniques such as hierarchical clustering [9], self-organizing maps and K-means [10], [11], fast search-and-find of density peaks [12] and the orthogonal wavelet transform [13]. Further, such time-series data are high-dimension data sets often affected by the *curse of dimensionality*, resulting in performance degradation of

The associate editor coordinating the review of this article and approving it for publication was Arash Asrari.

clustering algorithms, high-biased estimates, and computational expense [13]. A variety of dimensionality-reduction techniques for time-series data exists in the literature, such as the symbolic aggregate approximation (SAX) [12], [14], discrete wavelet transform [15] and discrete Fourier transform [16].

Further, the recently developed SAX-based dimensionality method for electricity consumption data representation has been implemented in recent studies [12], [14] without a clear definition of how to determine the optimal size of the segment and dimensionality of the transformed space in the lower-dimensional patterns. The number of segments is one key parameter that determines the effective implementation of SAX. For instance, a small number of segment sizes will lead to a compact representation with less information, whereas a very large number of segments could result in noise in the load time-series data set [17].

Few studies have highlighted different approaches for determining the optimal number (or sizes) of segments for SAX-based times-series approximation. Fotso *et al.* [17] proposed a parameter-free heuristic based on intuitive ideas from time-series classification to determine the optimal value of the number of segments. Zan *et al.* [18] used Shannon sampling theorem and adaptive hierarchical segmentation to determine the optimal size for SAX-based discretization of the time-series data; however, this article proposed the use of cumulative explained variance in the principal component analysis (PCA) to determine the optimal number of segments and dimensionality of the transformed space during discretization while retaining the data integrity, instead of using intuition, as proposed by the extant literature. The cumulative explained variance, obtained from the PCA eigen values, provides the number of principal components that retain significant portion of the full time -series data. Because this metric provides a measure of how much variance of the data can be retained, it can be used to determine the optimal number of segments for a SAX-based dimentionality-reduction technique. For smart meter data sets with high-dimensional space, it is pivotal to have a methodology that can be efficiently used to determine the optimal number of segments to produce a compact representation of the time-series data without noise and the loss of critical inherent information.

Further, existing studies have focused entirely on load profiling without considering the impact of DERs on customer load profiles. We propose approaching the customer loads as net loads, which now form an integral part of the evolving power system, with customer-sited resources. Net load profiling can help system operators identify customers for demand response programs and provide useful information for generation scheduling, integrated system planning, and power system flexibility evaluations [7]. Fig. 1 shows the net consumption behind-the-meter, illustrating its different components, such as gross load, net load (total load - total PV generation), exported and self-consumed generation, which are pivotal for integrated distribution planning operations.

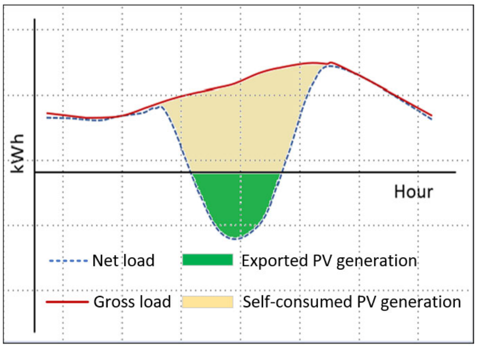

**FIGURE 1.** A typical behind-the-meter net consumption profile.

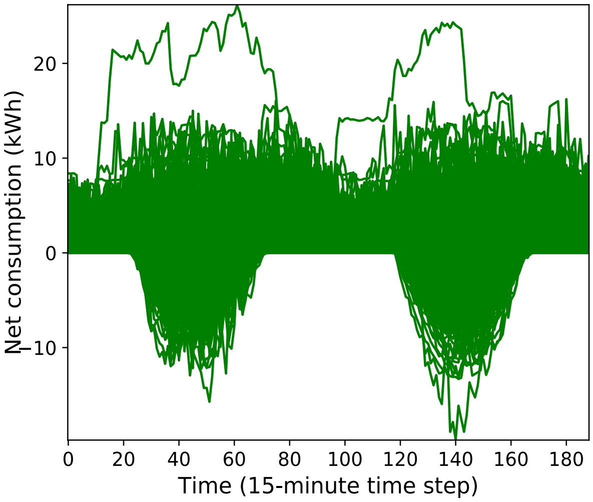

**FIGURE 2.** Net consumption profiles at 15-minute resolution.

Apart from the variability inherent in these net consumption profiles, the utility need to deal with different peaks occurring at various times of the day. Another consideration for integrated planning, is the shifting of the peaking periods caused by the presence of on-site PV generation. Figs. 2 and 3 show net consumption profiles for the residential customers used in this study at 15-minute and 1-hour resolutions respectively. Net consumption profiling is, therefore, necessary to identify typical profiles among these varying patterns that can be used for integrated distribution planning and operations and to evaluate the grid-readiness to accommodate back- feed introduced by customer-sited generation. The contributions of this paper are as follows:

- The cumulative explained variance in the PCA is proposed to determine the optimal number of segments for SAX-based discretization while minimizing noise and maintaining times-series compactness.
- Application of SAX-based dimensionality-reduction and k-means to perform net consumption profiling at different resolutions.
- The application of the proposed method for net consumption profiling is analyzed and discussed.

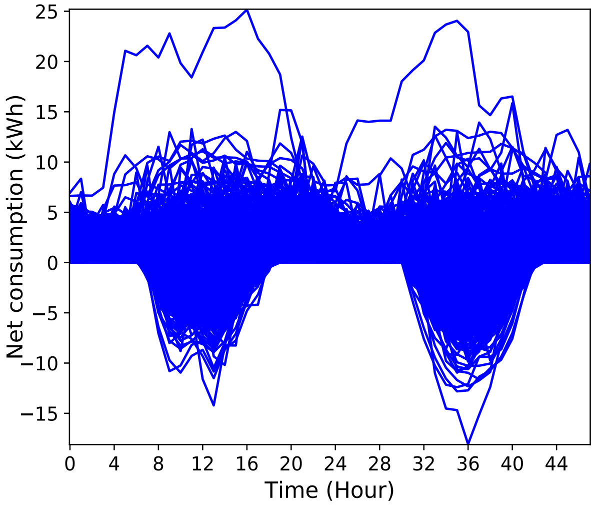

**FIGURE 3.** Net consumption profiles at 1 hour resolution.

- This paper proposes a screening methodology to determine the distribution of high-voltage customers among the resulting clusters of customers with and without on-site solar PV generation at different time steps.

## II. METHODOLOGY FOR NET CONSUMPTION PROFILING

The proposed methodology for net consumption clustering is grouped into five steps, as depicted in Fig. 4. In the first stage, smart meter time-series data obtained from the data repository are passed through preprocessing operations, such as data cleaning, removal of outliers, net load computation, and normalization for PCA and clustering.

The second step describes a methodology for finding the number of segments for SAX-based discretization while minimizing noise and maintaining times-series compactness at different resolutions. The third step reduces the dimensionality of the net load profiles using SAX, whereas the next step uses the k-means clustering technique to identify typical net consumption profiles. Finally, the last step describes a screening methodology to determine the distribution of high-voltage customers among the resulting clusters of customers with and without on-site solar PV generation. This articles uses tslearn package, which is a toolkit dedicated for clustering time series data [26]. The details of these processes are discussed in the following subsections.

### A. DATA PREPARATIONS

Data preparation performed in this study includes detection and removal of bad data as a result of movement or temporal disconnection of host meters. Also, outliers were removed using statistical measures, such as mean and standard deviation. Data standardization was performed to center and normalize the data set using standard scaler, which transforms the net consumption profiles by removing the mean and scaling to the unit variance as given in (1):

$$z = \frac{X - \bar{X}}{X_{std}} \tag{1}$$

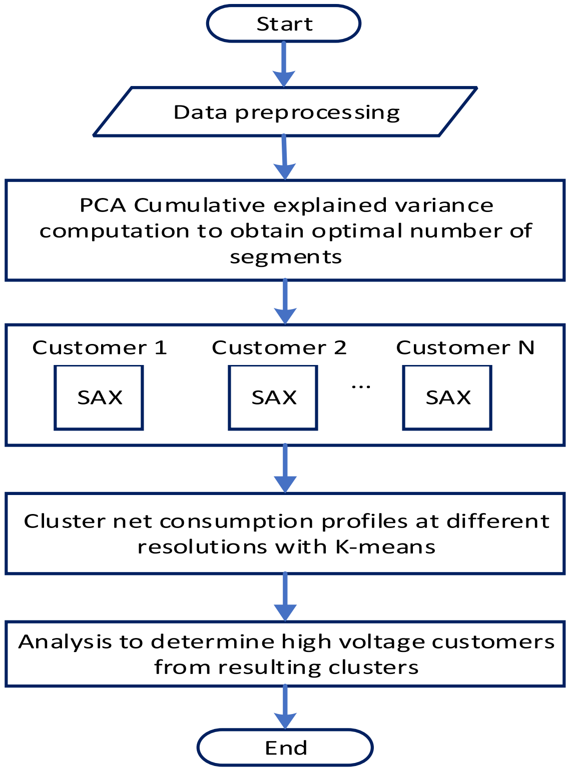
Start
Data preprocessing
PCA Cumulative explained variance computation to obtain optimal number of segments
Customer 1
Customer 2
Customer N
SAX
SAX
...
SAX
Cluster net consumption profiles at different resolutions with K-means
Analysis to determine high voltage customers from resulting clusters
End

**FIGURE 4.** Net consumption profiling procedures.

where z represents the standard score of a sample X, and $\bar{X}$ and $X_{std}$ are its mean and standard deviation [19], [20].

### B. SAX FOR NET CONSUMPTION PROFILES

The SAX-based dimentionality-reduction technique is used to approximate the time-series data via symbolic representation, with the lower bounding of distance measures defined in the original time series. The main advantage of this lower bounding property is that it allows the reduced-size representation to index the original data without false negatives. SAX implementation consists of two steps: transforming the net consumption profiles into a piecewise aggregate approximation (PAA) and symbolizing the PAA representation into a discrete string [22], [23], [26].

Consider a time-series data X of length n with elements $x_1, x_2, \ldots, x_n$. SAX approximation reduces the dimension of the data to m-dimensional time-series Y with elements $y_1, y_2, \ldots, y_m$, using PAA, where $m << n$. The ith element of Y can be expressed as follows [12], [14], [22], [23]:

$$y_i = \frac{m}{n} \sum_{j=n/m(i-1)+1}^{(n/m)i} x_j \tag{2}$$

where i is the index of the transformed PAA net consumption data, and j is the index of the normalized net consumption data. According to (2), to reduce the times series data X with n-dimension to m-dimension, the original data is divided into m-sized frames. The average value of the data within each frame is estimated and a vector of these values represents the approximation of the original time-series data. This averaging of the PAA can be used to smooth out large variability inherent in load profiles. The lower bounding property in SAX ensures the distance measure, D, as given in (3) is satisfied:

$$D(Y_1, Y_2) \leq D(X_1, X_2) \tag{3}$$

where $D(Y_1, Y_2)$, which is the lower bounding distance measure of the SAX representation data is given as follows:

$$D(Y_1, Y_2) = \sqrt{\frac{n}{m}}\sqrt{\sum_{i=1}^{m}(y_{1i} - y_{2i})^2} \tag{4}$$

$D(X_1, X_2)$), which could be an Euclidean distance measure for the original time-series data is given as follows:

$$D(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2l})^2} \tag{5}$$

The lower dimensional time-series data Y obtained using PAA is then transformed into a discrete representation through a number of equiprobable symbols. This is done using a Gaussian distribution because the normalized time-series data has a highly Gaussian distribution feature. The Gaussian curve is partitioned into "a" equal-sized areas using breakpoints $(\beta_1, \beta_2, \ldots, \beta_{a-1})$ in ascending order to have symbols with equal probability [12], [14].

A key parameter that determines the effective implementation of SAX, is the choice of the number of segments. A small number of segments will result in a compact representation of the time-series with less information, whereas a very large number of segments could lead to noise in the data set. This article proposes a methodology to determine the optimal number of segments for SAX-based dimensionality-reduction.

### C. EVALUATION OF OPTIMAL NUMBER OF SEGMENTS

The implementation of PCA for dimension reduction of the time-series data is not the focus of this article, and therefore, will not be discussed in detail; however, its application for determining the optimal number of segments is described. A key parameter that affects the effective implementation of SAX-based dimensionality reduction is the number of segments, which the time-series data is partitioned, to produce a compact representation and reduce data numerosity with minimal loss of information. The principal components provide the amount of variation in the data set which decreases as one moves from the first principal component to the last. The explained variance, which is the ratio of the eigen values for each principal component to the sum of the eigenvalues of all principal components, is an important metric in determining the quality of reconstruction done by the PCA and how much information is retained or lost [21]. The cumulative explained variance, which is the cumulative version of the explained variance, is a measure used to determine how much of information is retained from the original time-series, and the importance of the components. This metric could have a maximum value of unity, which represents no loss of information during dimension reduction in the transformed space. For the time-series data, Fig. 5 shows that keeping 500 principal components retains 90% of the information, whereas 10% is lost for the 15-minute time step net consumption data.

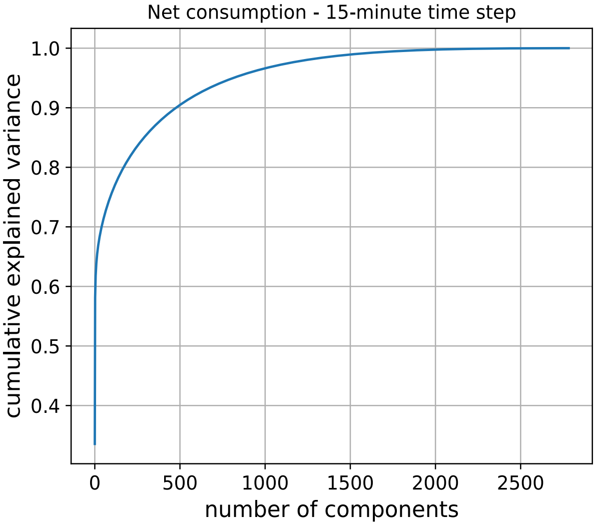

**FIGURE 5. Cumulative explained variance at 15-minute resolution.**

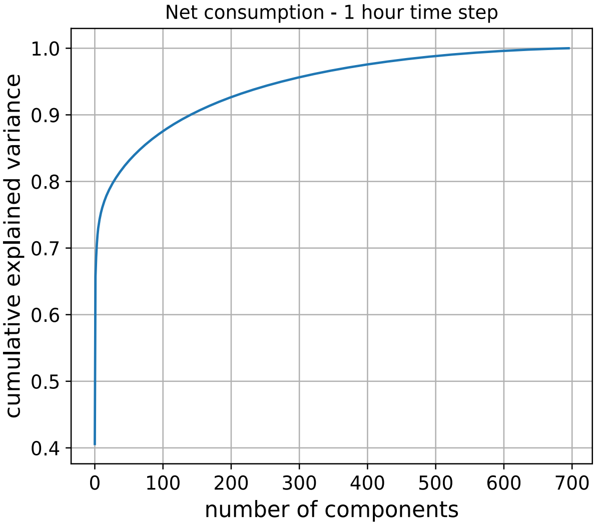

**FIGURE 6. Cumulative explained variance at 1-hour resolution.**

To determine the optimal number of segments, the number of principal components is varied iteratively from 500 principal components until the approximated time-series aligns with the original data. This is done to ensure that the maximum possible information is retained from the original time-series. This study uses 1000 components which retains about 98% of the information in the time-series data as shown in Fig. 5 for the 15-minute resolution data. The same procedure is applied to the 1-hour resolution net consumption data, shown in Fig. 6.

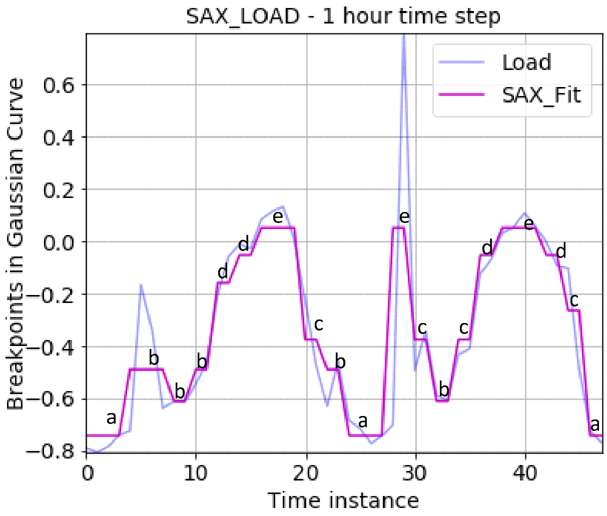

**FIGURE 7.** Net consumption profile and its SAX representation at 1-hour resolution.

According to (2), it is required that the number of m-sized to divide the original time-series into has to be determined while retaining data integrity in the transformed space. For 1-hour data resolution, the cumulative variance plot as depicted in Fig. 6, shows that keeping 300 principal components would retain about 95% of the information. Because SAX approximation depends on PAA representation, the calculated number components is used to determine the optimal PAA number of segments. For example, as shown in Fig. 7 with the SAX representation of the net consumption profile for a particular customer at 1-hour resolution, all the PAA coefficients that are less than the smallest breakpoint are mapped to symbol "a", whereas all coefficients greater than or equal to the smallest breakpoint and less than the second smallest breakpoint are mapped to symbol "b," and so on.

The symbolization of the time series data is done in ascending order of breakpoints, and the concatenation of symbols is defined as word [14]. The SAX representation for this profile can be represented as "abbbddecbaecbcdedca" with 5 symbols, 5 alphabet size, and 19 word size. The same procedure can be repeated for the 15-minute resolution SAX representation, as shown in Fig. 8.

## III. CLUSTERING OF CONSUMPTION PROFILES AND EVALUATORS

### A. K-MEANS CLUSTERING ON SAX REPRESENTATION

Clustering net consumption profiles at different resolutions is needed to differentiate and identify typical profiles of customers with and without on-site generation to help in planning operations, such as generation scheduling. In addition, clustering provide a means to locate a group of customers causing high voltage, and this can be used by the utility for further analysis. The extant literature shows that the application of the k-means clustering technique on SAX approximation of the original time series yields a good result [14], [23]. This paper uses this clustering method to identify typical net consumption profiles.

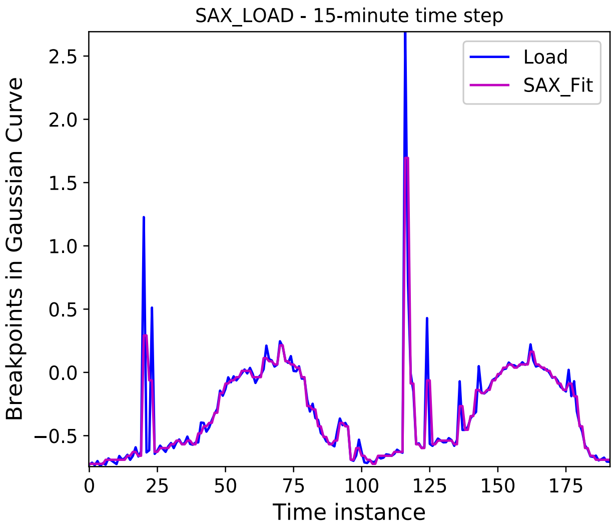

**FIGURE 8.** Net consumption profile and its SAX representation at 15-minute resolution.

For example, for n-number of time-series data sets, $X_k$, k = 1, 2, . . . , n to be clustered into C number of clusters. k-means initializes C number of cluster centroids and defines an objective function, J, which minimizes the sums of each data point to cluster centroids distances iteratively over all clusters, as given in (6) [14], [24]:

$$J = \sum_{i=1}^{C} \sum_{k=1}^{n} ||X_k - m_i|| \tag{6}$$

where $m_i$, i = 1, 2, . . . , C are C-number of cluster centroids. C is updated iteratively until the minimum value of J is obtained.

### B. EVALUATORS FOR K-MEANS CLUSTERING

This paper evaluates the quality of k-means clustering on net consumption profiles at 15-minute and hourly resolutions using the following metrics:

1) Silhouette coefficient (s): This is used where the ground truth labels are unknown, and for each sample, it is defined as follows:

$$s = \frac{d - e}{max(d, e)} \tag{7}$$

where "d" represents the average value between a sample and every other point in the same cluster and "b" is the average distance between a sample and every other point in the next neighboring cluster. The silhouette index ranges between the interval [−1, 1], with higher values indicating better defined clusters, −1, +1, and scores near zero represent incorrect, highly dense and overlapping clustering, respectively.

2) Calinski-Harabaz Index ($s_k$): This metric, also referred to as the variance ratio criterion (VRC), is defined as the ratio of the between-clusters dispersion ($S_{BW_m}$) and the within-cluster dispersion ($S_{WT_m}$).

The Calinski-Harabaz Index, s, for m clusters is defined as follows [25], [27]:

$$s(m) = \frac{S_{BW_m}}{S_{WT_m}} * \frac{N-m}{m-1}, \quad m \geq 2 \tag{8}$$

$$S_{WT_m} = \sum_i^m \sum_{j \in C_i} ||(j - y_i)||^2 \tag{9}$$

$$S_{BW_m} = \sum_i^m \sum_{j \in C_k} ||(j - y_i)||^2, \quad i \neq k \tag{10}$$

where N is the total number of data points in the time-series data, j is a data point within cluster i, $C_i$ is the ith cluster, $y_i$ is the centroid of cluster i, and $||(j - y_i)||$ is the Euclidean distance between j and $m_i$. Better defined clusters are expected to return a higher index.

3) Davies-Bouldin Index (DBI): This is a measure of the average similarity, $R_{i,j}$, between each cluster $C_i$ for i = 1, ..., m and its most similar one $C_j$, given as follows [25], [27]:

$$DBI = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i} R_{ij} = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i} \frac{W_i + W_j}{B_{ij}} \tag{11}$$

$$B_{ij} = ||(B_i - B_j)|| = \sum_{m=1}^{n} |b_{m,i} - b_{m,j}|^{(1/2)} \tag{12}$$

$$W_i = \left[\frac{1}{N_i}\sum_{m=1}^{N_i} |S_m - B_i|^{(2)}\right]^{(1/2)} \tag{13}$$

where $B_{ij}$ represents the between-cluster-distance for centroids i and j, $b_{m,i}$ is the mth element of $B_i$, $N_i$, is the total number of data points in cluster i, and $W_i$ is the within-cluster-distance for each data point, $S_m$, in cluster i and its centroid $B_i$.
A lower DBI score shows a better defined cluster separation, with zero being the lowest possible value.

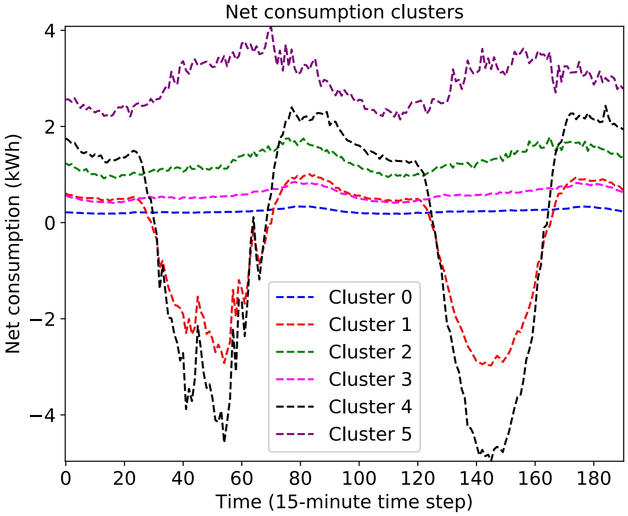

**FIGURE 9.** Net consumption clustering at 15-minute resolution.

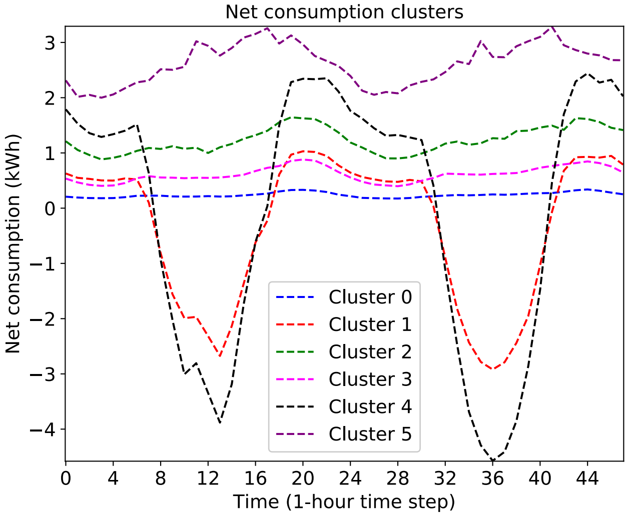

**FIGURE 10.** Net consumption clustering at 1 hour resolution.

## IV. CASE STUDY

The net consumption data set used in this article contains electricity consumption of 3,569 customers participating in an advanced metering infrastructure (AMI) pilot project in Hawai'i, of which 747 customers have on-site solar PV generation. Among these data, 155 bad load profiles and outliers were removed, which is a very small sample of the whole data set. Bad load profiles have zero measurements caused by meter movement or swapping whereas outliers are data points less than (mean − 3*standard deviation) and any points more than (mean + 3*standard deviation).

### A. DIMENSIONALITY REDUCTION OF THE TIME-SERIES USING SAX

For a 2-day (48-hour) net consumption profile with 15-minute and hourly resolutions data, the lengths of the time-series are 192 and 48, respectively. Using SAX, net consumption profiles can be reduced to a smaller dimension, w, where "w" represents the SAX word size. The compression ratio obtained using SAX is defined as the ration of the length of the original time series, n, and the word size, w. For the 15-minute and hourly resolution time-series, with 100, and 19-word SAX representations, the compressions achieved for a 2-day net consumption profiles are 1.92 and 2.52, respectively. A reduced data set requires less computational resources in clustering, while maintaining significant characteristics of the original data set.

### B. NET CONSUMPTION PROFILES CLUSTERING

Net consumption time-series data were clustered into six representative load and net load profiles using the SAX-based k-means algorithm, and the results are presented in Figs. 9 and 10 for 15-minute and 1-hour resolutions, respectively. Tables I and II provide a detailed analysis of these clusters for the considered time steps over 1-month.

The clustering results are very similar when comparing 15-minute and 1-hour resolutions in terms of cluster

**TABLE 1.** Time-series cluster analysis at 15-min resolution.

| Cluster | Peak Demand | | | $P_A$ | Avg. | $\% \ C_m$ | $\% < 0.95$ | $\% > 1.05$ |
|---|---|---|---|---|---|---|---|---|
| | kW | Time | Day | | kWh/day | | p.u | p.u. |
| 0 | 0.39 | 20:30 | 18 | 1.61 | 5.82 | 41.4 | 32.5 | 11.1 |
| 1** | 1.44 | 20:00 | 18 | n/a | -6.94 | 12.6 | 5 | 37 |
| 2 | 2.10 | 20:00 | 19 | 1.58 | 31.99 | 10.8 | 17.5 | 0.0 |
| 3 | 1.08 | 20:00 | 18 | 1.75 | 14.79 | 26.5 | 20 | 18.5 |
| 4** | 3.26 | 20:15 | 18 | n/a | 0.161 | 5.9 | 10 | 33.3 |
| 5 | 4.56 | 16:45 | 19 | 1.49 | 73.57 | 2.8 | 15 | 0 |

** Clusters with PV ; $C_m$-cluster membership; $P_A$-Peak-to-average ratio

**TABLE 2.** Time-series cluster analysis at 1-hour resolution.

| Cluster | Peak Demand | | | $P_A$ | Avg. | $\% \ C_m$ | $\% < 0.95$ | $\% > 1.05$ |
|---|---|---|---|---|---|---|---|---|
| | kW | Time | Day | | kWh/day | | p.u | p.u. |
| 0 | 0.39 | 20:00 | 18 | 1.62 | 5.78 | 38 | 66.7 | 0 |
| 1** | 1.52 | 20:00 | 18 | n/a | -5.98 | 12 | 0 | 30 |
| 2 | 2.05 | 21:00 | 20 | 1.59 | 30.96 | 12 | 0 | 0 |
| 3 | 1.08 | 20:00 | 18 | 1.7 | 15.21 | 28 | 6.7 | 30 |
| 4** | 3.35 | 20:00 | 18 | n/a | 2.88 | 6 | 0 | 40 |
| 5 | 3.88 | 20:00 | 24 | 1.41 | 66.12 | 4 | 26.7 | 0 |

** Clusters with PV; $C_m$-cluster membership; $P_A$-Peak-to-average ratio

membership and load profile characteristics, such as peak load, peak-to-average ratio, and energy consumption. What is mostly affected by the time resolution is the number of voltage violations above or below 1.05 p.u. and 0.95 p.u., respectively. Hourly resolution does not capture as many voltage violations as the 15-min resolution.

The highest population cluster is Cluster 0, with approximately 40% of the AMI customers. This characterizes very low consumption users, which are typical of tropical climates with no heating or cooling, and solely plug-loads, lighting, and wet appliances as major contributors to the residential demand. The peak demand driven by lighting occurs very late, close to 8:30 p.m. in the evening. In fact, Hawai'i has the nation's lowest residential sector energy consumption [28].

The next most populated cluster is Cluster 3, with approximately 30% of the AMI meters. It is also a low-consumption cluster, similar to Cluster 1 with a peak of 1,080 Watts occurring late in the evening, at 8 p.m. Following the two low-consumption clusters, with 12% of AMI meters is Cluster 1, characterizing net load profiles with low consumption that have installed PV systems that zero out the energy consumption during the course of a day. The peak demand of 1.44 kW at 8 p.m. is comparable to Cluster 3 but three times higher than Cluster 0. The negative peak from the PV power exported (-3 kW) is three times higher than the peak demand. This is relevant to utilities experiencing high-penetration rooftop PV because as increasing numbers of users install rooftop PV, the thermal rating of the distribution service transformer can be violated as a result of negative power export from net load profiles, such as the ones in Cluster 1.

Cluster 2, with 12% of the AMI customers, represents higher energy and power consumption loads, with a peak demand of 2 kW occurring also late into the evening (8 p.m. and 9 p.m.). Next is Cluster 5, which is the least populated cluster, with 3%-4% of the customers, representing higher consumption houses not that common in tropical islands, with a peak demand of 3-4.5 kW occurring earlier in the evening between 5 p.m. and 8 p.m., which is very likely caused by customers that have heating, ventilating, and air-conditioning cooling systems that start cooling when consumers get home from work.

Finally, Cluster 4 is another representative net load profile of higher end energy consumers that is less common than Cluster 1 but still represents 6% of the metered customers that also net zero their consumption daily, with a peak of 3.35 kW and a negative export peak power of 4.5 kW, so with a installed.

With regard to undervoltage violations, Cluster 0, which has very low demand consumption, experiences the highest number of undervoltage violations at both 15-min and hourly time resolutions. This could be due to primary voltage regulation lowering voltages during the day to accommodate PV voltage rise and/or as a result of neighbouring high-consumption customers causing undervoltage violations.

With regard to overvoltage violations, the trends are clear at both 15-minute and hourly resolutions. Higher voltages are driven by customers with PV systems, and there are comparable voltage violations in Cluster 1 and Cluster 4, but Cluster 1 has twice the number of PV customers than Cluster 4. This leads to the attribute of the larger exporting profile of Cluster 4, or customers installing larger PV systems in that cluster to cause more over-voltage violations. Finally, the lower consumption Clusters 0 and 3 also experience over-voltage violations. This could be because neighboring customers with PV systems also drive high voltages for low-consumption customers, whereas high-consumption customers in Clusters 2 and 5 are not effected by neighboring PV systems. Another explanation could be that high consumption customers tend to be located in newer neighborhoods that are often underground, and more susceptible to undervoltages but not overvoltages.

Further, in terms of membership for corresponding clusters (e.g., Cluster 4 in Figs. 9 and 10, respectively), Tables I and II indicate 5.9% and 6% membership for 15-minute and 1-hour resolutions, respectively. Also, for corresponding Cluster 1 in Figs. 9 and 10, respectively, Tables I and II show 12.6% and 12% membership for 15-minute and 1 hour resolutions respectively. This shows that the application of k-means on SAX representation of the net consumption profiles produces very close and comparable results without the loss of significant information in the original data set. This implies that for some applications with huge data sets and high computational requirements, down-sampling from 15-minute to 1 hour can

**TABLE 3. Clustering evaluation.**

| Metric | Time series resolution | |
|---|---|---|
| | 15-minute | 1 hour |
| Silhouette | 0.26 | 0.27 |
| VRC | 981.7 | 1395.2 |
| DBI | 1.55 | 1.39 |

still preserve the integrity of cluster membership; however, Tables I and II indicate that down-sampling could shift the peak load time, which has the potential to affect distribution planning operations, such as generation scheduling.

Table III shows the values for the three internal evaluators used in this study to assess the quality of clustering at different time steps. The higher values of the silhouette and VRC indices for 1-hour resolution show better defined clusters than the 15-minute time step. Also, the smaller DBI value for 1-hour resolution indicates a denser intra-cluster property and larger inter-cluster distances, which result in better clustering.

## V. POTENTIAL APPLICATIONS OF NET CONSUMPTION PROFILES CLUSTERING

The extant studies have focused entirely on clustering load time-series data sets without considering the impact of local generation on customer demand profiles. As distribution system planning gradually evolves using an integrated approach, it is now pivotal to be able to characterize net load on the network. This impact is no longer limited to the distribution network because at high levels of DER penetration, the net load characteristics can have a significant impact on the transmission domain and bulk power system operation.

Net electricity consumption clustering can help identify typical clusters of DER installations that can exacerbate adverse impacts on the grid, such as voltage rise and reverse power flow. The ability to localize such impacts can be used for further network analysis, detailed impact assessment of integrated DER units, and improved system operations. A direct application of the clustering net load profiling is to use the representative cluster load profiles to add load diversity to distribution modeling and simulation studies. Most distribution modeling and simulation studies currently use quasi-static time-series power flow simulation techniques, leveraging SCADA substation data to approximate the load profile. However, it is increasingly important to capture and represent the load diversity at the primary and secondary levels to understand the local impact of customer-sited resources such as rooftop PV in distribution planning and operation studies. The net load clustering proposed in this paper can be used to better represent customer loads with and without PV in distribution power flow models.

Also, clustering net demand can help system operators and planners identify customers for demand response programs and provide useful information for generation scheduling, integrated system planning, and net load forecasting. The output of net electricity consumption clustering can be a useful resource in ensuring load diversity for performing a more realistic integration studies.

## VI. CONCLUSION

As the power system continues to evolve through integrated network planning considering the impact of local generation, there is an increasing need to have a more realistic characterization of customer load for effective system operations, such as generation scheduling. The net load provides the amount of demand that is visible to system operators and must be met by the traditional, dispatchable generation.

This paper uses SAX-based dimensionality-reduction and k-means techniques to cluster the net consumption of smart meter data for more than 3,500 residential customers in a month at different temporal resolutions. This study shows that the application of k-means on the SAX reduced representation of the original time-series data can be used to cluster net demand profiles. This study proposes the use of cumulative explained variance in the principal component analysis to determine the optimal number of segments and dimensionality of the transformed space during discretization while retaining the data integrity, instead of using intuition as proposed by the extant literature.

The clustering results are very similar when comparing 15-min and 1-hour resolutions in terms of cluster membership and load profile characteristics, such as peak load, load factor, and energy consumption. What is mostly affected by the time resolution is the number of voltage violations because the hourly resolution does not capture as many voltage violations as the 15-min resolution. The results of the methodology presented in this paper can be used to determine the distribution of high-voltage customers among the resulting clusters of customers with and without on-site solar PV generation at different time steps.

## ACKNOWLEDGMENT

The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

## REFERENCES

[1] U. Singh, V. Zamani, and M. Baran, "On-line load estimation for distribution automation using AMI data," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Boston, MA, USA, Jul. 2016, pp. 1–5.

[2] H. Wang and N. N. Schulz, "Using AMR data for load estimation for distribution system analysis," *Electr. Power Syst. Res.*, vol. 76, pp. 336–342, Mar. 2006.

[3] M. Yazdani-Damavandi, N. Neyestani, G. Chicco, M. Shafie-khah, and J. P. S. Catalão, "Aggregation of distributed energy resources under the concept of multienergy players in local energy systems," *IEEE Trans. Sustain. Energy*, vol. 8, no. 4, pp. 1679–1693, Oct. 2017.

[4] X. Han, K. Heussen, O. Gehrke, H. W. Bindner, and B. Kroposki, "Taxonomy for evaluation of distributed control strategies for distributed energy resources," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5185–5195, Sep. 2018.

[5] T. Hong, M. Gui, M. E. Baran, and H. L. Willis, "Modeling and forecasting hourly electric load by multiple linear regression with interactions," in *Proc. IEEE PES Gen. Meeting*, Jul. 2010, pp. 1–8.

[6] A. Bergman, P. Denholm, D. Steinberg, and D. Rosner, "Maintaining the reliability of the modern power system," US Dept. Energy, Washington, DC, USA, Tech. Rep., Dec. 2016. [Online]. Available: https://www.hsdl.org/?view&did=806857

[7] S. Sreekumar, K. C. Sharma, and R. Bhakar, "Gumbel copula based aggregated net load forecasting for modern power systems," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 19, pp. 4348–4358, Oct. 2018.

[8] Y. V. Makarov, P. V. Etingov, J. Ma, Z. Huang, and K. Subbarao, "Incorporating uncertainty of wind power generation forecast into power system operation, dispatch, and unit commitment procedures," *IEEE Trans. Sustain. Energy*, vol. 2, no. 4, pp. 433–442, Oct. 2011.

[9] G. J. Tsekouras, P. B. Kotoulas, C. D. Tsirekis, E. N. Dialynas, and N. D. Hatziargyriou, "A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers," *Electric Power Syst. Res.*, vol. 78, no. 9, pp. 1494–1510, 2008.

[10] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 596–602, May 2005.

[11] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, "Development of low voltage network templates—Part I: Substation clustering and classification," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3036–3044, Nov. 2015.

[12] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2437–2447, Sep. 2016.

[13] H. Zhang, T. B. Ho, Y. Zhang, and M.-S. Lin, "Unsupervised feature extraction for time series clustering using orthogonal wavelet transform," *Informatica*, vol. 30, no. 3, pp. 305–319, 2006.

[14] M. J. E. Alam, K. M. Muttaqi, and D. Sutanto, "A SAX-based advanced computational tool for assessment of clustered rooftop solar PV impacts on LV and MV networks in smart grid," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 577–585, Mar. 2013.

[15] Y. Xiao, J. Yang, H. Que, M. J. Li, and Q. Gao, "Application of Wavelet-based clustering approach to load profiling on AMI measurements," in *Proc. IEEE China Int. Conf. Electr. Distrib. (CICED)*, Shenzhen, China, Sep. 2014, pp. 1537–1540.

[16] S. Zhong and K.-S. Tam, "Hierarchical classification of load profiles based on their characteristic attributes in frequency domain," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2434–2441, Sep. 2015.

[17] V. S. S. Fotso, E. M. Nguifo, and P. Vaslin, "Parameter free piecewise dynamic time warping for time series classification," in *Proc. ICML Time Ser. Workshop*, Sydney, NSW, Australia, 2017.

[18] C. T. Zan and H. Yamana, "Dynamic SAX parameter estimation for time series," *Int. J. Web Inf. Syst.*, vol. 13, no. 4, pp. 387–404, 2017.

[19] 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, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[20] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, "Phase identification in electric power distribution systems by clustering of smart meter data," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 259–265.

[21] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[22] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," *Knowl. Inf. Syst.*, vol. 3, no. 3, pp. 263–286, 2001.

[23] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: A novel symbolic representation of time series," *Data Mining Knowl. Discovery*, vol. 15, no. 2, pp. 107–144, 2007.

[24] T. W. Liao, "Clustering of time series data—A survey," *Pattern Recognit.*, vol. 38, pp. 1857–1874, Nov. 2005.

[25] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist., Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[26] R. Tavenard. (2017). *tslearn: A Machine Learning Toolkit Dedicated to Time-Series Data*. [Online]. Available: https://github.com/rtavenar/tslearn

[27] F. Molin, "Cluster analysis of European banking data," M.S. thesis, KTH Roy. Inst. Technol., Stockholm, Sweden, 2017.

[28] *State Energy Data System, Table C13, Energy Consumption Estimates Per Capita by End-Use Sector, Ranked by State*, U.S. EIA, Washington, DC, USA, 2016.

**MICHAEL EMMANUEL** received the B.Sc. degree in electrical/electronic engineering from the University of Ibadan, Ibadan, Nigeria, in 2008, the master's degree in electronics and telecommunications engineering from Jadavpur University, Kolkata, India, in 2012, and the Ph.D. degree from the Victoria University of Wellington, Wellington, New Zealand, 2018. He was a Graduate Engineer with the Power Holding Company of Nigeria-Sub-Transmission and as a HUAWEI BSS Technical Assistant Centre Engineer with Globacom Telecommunications, Lagos, Nigeria. He is currently with the National Renewable Energy Laboratory (NREL), Golden, CO, USA, as a Postdoctoral Researcher with the Power System Engineer Center. His research interests include distributed energy resources integration with the electric power systems with the high penetration of variable renewable energy systems, bulk power system operations, production cost modelling, power system planning, and smart grid communication networks.

**JULIETA GIRALDEZ** received the bachelor's degree in technical mining engineering from the Polytechnic University of Madrid, Spain, and the master's degree in electrical engineering from the Colorado School of Mines, Golden, CO, USA. She is currently pursuing the Ph.D. degree in systems engineering with Colorado State University. She was a Key Technical Contributor to the DOE Arizona Public Service High-Pen PV Project and the Duke Energy Case Study Project using advance inverters and a Distribution Management System for feeder voltage regulation. She is currently with the National Renewable Energy Laboratory (NREL), Golden, as a Senior Research Engineer with the Power System Engineer Center where she also leads the Microgrid and Smart Grid and Grid Integration Related Projects. She is also leading a DOE study on microgrid costs in USA and a project with HECO to simulate distribution feeder operations with advanced inverters. Her current research interests include integrating emerging technologies such as PV, energy storage, and microgrids in distribution systems.