Big Data Analytics for Demand Response: Clustering Over Space and Time

Charalampos Chelmis
Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
chelmis@usc.edu

Jahanvi Kolte
Institute of Technology
Nirma University
Ahmedabad, Gujarat, India
jahanvi1612@gmail.com

Viktor K. Prasanna
Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
prasanna@usc.edu

***Abstract*—The pervasive deployment of advanced sensing infrastructure in Cyber-Physical systems, such as the Smart Grid, has resulted in an unprecedented data explosion. Such data exhibit both large volumes and high velocity characteristics, two of the three pillars of Big Data, and have a time-series notion as datasets in this context typically consist of successive measurements made over a time interval. Time-series data can be valuable for data mining and analytics tasks such as identifying the "right" customers among a diverse population, to target for Demand Response programs. However, time series are challenging to mine due to their high dimensionality. In this paper, we motivate this problem using a real application from the smart grid domain. We explore novel representations of time-series data for BigData analytics, and propose a clustering technique for determining natural segmentation of customers and identification of temporal consumption patterns. Our method is generizable to large-scale, real-world scenarios, without making any assumptions about the data. We evaluate our technique using real datasets from smart meters, totaling $\sim 18,200,000$ data points, and show the efficacy of our technique in efficiency detecting the number of optimal number of clusters.**

***Keywords*-demand response; pattern mining; time-series; cyber-physical systems**

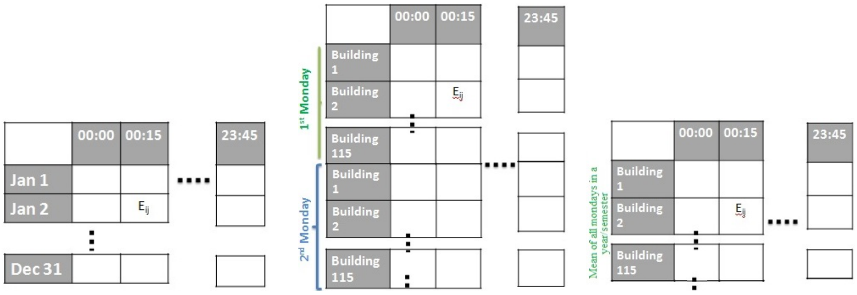
I. Introduction

The ubiquitous deployment of Advanced Metering Infrastructure (AMI) by utilities have enabled electricity usage sensing and bi-directional communication between consumers and electric utilities [1]. This provides ample opportunities to efficiently deal with peak demands, and reduce energy consumption during peak demand periods using pricing incentives as in Demand Response (DR) programs [2], [3]. The growing availability of high resolution, high-dimensional electricity consumption data offers unique opportunities in developing forecasting models [4], [5], but has also offered a data goldmine for data analytics which are crucial in helping consumers understand their electricity consumption footprints and utilities unlock the potential benefits of investing into smart meters by unraveling customer behavior in fine granularities. As customers usage varies widely based on their needs, defining and describing subsets of customers whose usage patterns are in some way similar from sensed data is of paramount importance to Smart Grid applications [2].

Analysis of energy meter data has received wide attention recently [5], [6]. Energy consumption recorded at fine granularity and the use of two-way communication between smart meters and utilities has enabled applications such as DR [7], [8], customer segmentation [9], [5], [10], consumer behavior prediction [9], energy consumption estimation from customer characteristics [5], customer preferences and socioeconomic characteristics derivation [11], and detection of consumption anomalies [12], [13]. Principal component analysis (PCA) has been extensively used in the literature to discover correlations with consumption data [14] and for variable selection among a large set of predictors [15], as well as to predict electricity consumption [16], [17].

In this work, we focus on uncovering patterns from large-scale AMI data over a large population and across various temporal granoularities. Intuitively, energy consumption is expected to be periodic, as it is governed by human activities which usually follow some schedule (e.g., daily or weekly). For example, a person is very likely to be at the same place on Monday mornings, and therefore it is also likely that an emerging behavior can be recorded; in this case the kwh of a building will be similar on Monday mornings even if occupied by multiple tenants or hosts hundreds of office spaces. However, usage is likely to differ by few half hours earlier or later due to natural irregularities in behavior (e.g., someone returned home at 6:30 p.m. instead of 6 p.m.). So how does a utility go about uncovering such patterns for hundreds of thousands or millions of customers automatically? To address this question, we use PCA to uncover (i) **temporal** patterns between consumption values for each customer individually, and (ii) **spatial** patterns, i.e., patterns common to several customers. We propose new representations of time-series data to mine such patterns across various temporal granularities. To the best of our knowledge, we are the first to address the challenge of automatic discovery of patterns from large-scale AMI data over a large population and across various temporal granularities.

The remainder of this paper is organized as follows. In Section II, we describe electricity consumption data representations to uncover implicit patterns from a large-scale real-world corpus. In Section III, we analyze consumption patterns and motivate the need for a dimensionality reduction



(a) Customer-wise

(b) Stashed for each day of the week

(c) Mean (customer-wise)

Figure 1: Matrix representation of 15-minute energy consumption data.

technique. In Section IV, we conduct cluster analysis in the principal component space, and present a detailed analysis on the results in Section V. We conclude in Section VI.

II. Energy Consumption Data

Smart Grids are facing a data explosion, with millions of customers getting upgraded to smart meters and power utilities contending with over 100 energy consumption data points per customer, per day, sampled every 15-mins that they need to analyze and use for intelligent grid operations [2], [3], [9], [10]. Insights from smart meter data enable utilities to maintain efficient and reliable grid operations, while also allowing consumers to use energy more effectively. However, statistical techniques for analyzing electricity consumption data may yield different results when applied at different granularities. Next we discuss two dimensions for which the level of resolution is important:

- **Temporal**: Appropriate “arrangement” (e.g., daily, weekly or monthly) of fine-grained 15-minute data (similarly hourly or daily data) can capture different patterns due to lifestyle, environmental, structural, and customer features.
- **Spacial**: Different consumption trends can be identified when analyzing data at the fine-grained household level or at the aggregated feeder level.

Thus, a set of questions arises: What is a good representation for energy consumption data? What kind of patterns should one expect to emerge out of a corpus of energy consumption data depending on data representation? Next, we set forth to answer such questions.

A. Stashing Consumption Data

We consider a daily observation of 15-minute energy consumption data $E^c = [e_1, \ldots, e_{96}]$, where e^c_j is the energy consumed by customer c in the j^{th} 15-minute interval of the day. We begin by arranging daily observations into a matrix $\mathbf{E^c}$ per customer c (Figure 1a), such that rows represent days in a year and columns represent 15-minute intervals of the day when energy consumption values were recorded[1]. In this case, the size of each matrix $\mathbf{E^c}$ is 365×96. We use this representation to study daily patterns per building, as well as to examine temporal variations in demand.

Next, we form a matrix of aggregate yearly[2] observations from all customers $c \in [1, N]$ for a specific day of the week (e.g., Monday). It follows that matrix $\mathbf{E^d}$, where $d \in [1, 7]$ denotes the day of the week, consists of rows which represent daily observations obtained for each customer c as shown in Figure 1b. In this case, the size of each matrix $\mathbf{E^d}$ is $52 \times N \times 96$. We use this representation to study variations in electricity demand over time (for the same day of the week) per customer and also to identify similarities (for the same day of the week) between customers.

The aforementioned matrix representations constitute fine-grained consumption data stashing strategies. For coarser representations we considered averaging energy consumption values row- or column- wise. For simplicity, we present here a representation according to which the contents of the matrix are obtained by considering the mean of energy consumption values for each day of the week accordingly at a specific 15-minute interval over the period of a day. Following, the notation used for the representation discussed in Figure 1b, we obtain that $\hat{e}^d_{cj} = \frac{1}{|D|}\sum_{k=1}^{|D|-1} e^d_{(N(k-1)+c)j}$ for day d, customer c, $|D|$ number of distinct d days (e.g., number of Mondays) over the course of a year, and j^th 15-minute interval of the day. In this case, the size of each matrix $\mathbf{\hat{E}^d}$ is $N \times 96$. We use this representation to study coarse-grain similarities in behavior between customers as well as statistically understand how their consumption changes on average by the day of the week.

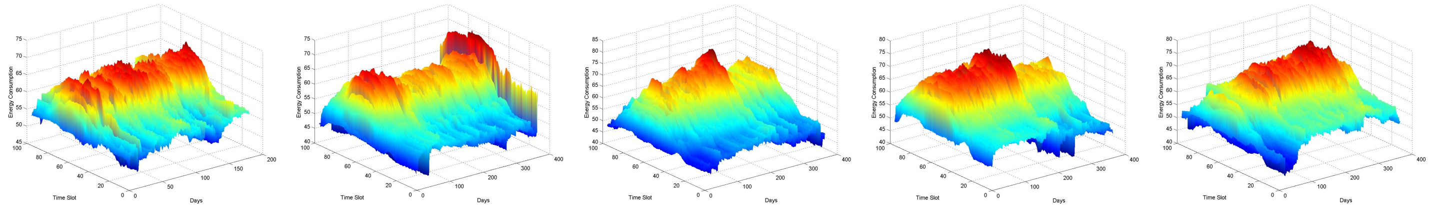
B. Data Set

The dataset used in this study was obtained from the University of Southern California campus microgrid[3]. The

[1] We also considered a representation where rows represent weeks in a year and columns represent energy consumption values over a week period.

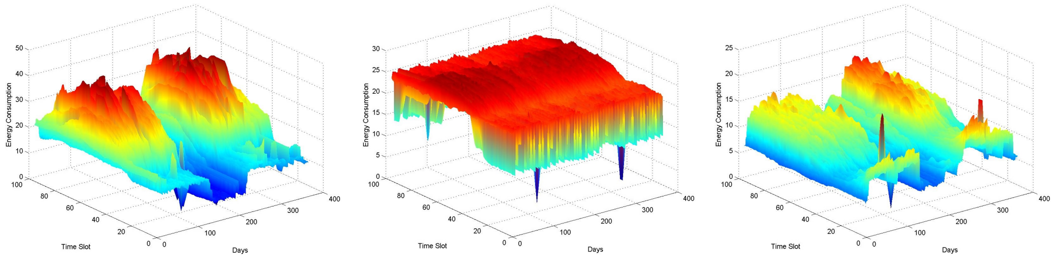
[2] We also experimented with semester-based segmentations.

[3] The dataset is available upon request for academic use from the USC Facility Management Services (FMS).



(a) 2009 (b) 2010 (c) 2011 (d) 2012 (e) 2013

Figure 2: Smart meter data for an individual building measured over five years.



(a) Building 54 (b) Building 68 (c) Building 106

Figure 3: Smart meter data for three buildings of different functions measured over a period of one year.

dataset comprises of a collection of observed electricity consumption values (measured in kWh at every 15 minutes) from 115 buildings, collected over a period of five years (January 1, 2009 to December 30, 2013), totaling **18,127,680** data points across all smart meters. The dataset contains a diverse set of building types: academic buildings with teaching and office space, residential dormitories, and administrative buildings. Building names have been obfuscated for privacy.

Figure 2 shows smart meter data for a specific building measured over five years. The x-axes represents the days of year, the y-axes denotes interval of the day, and the z-axes shows energy consumption. This visualization corresponds to the matrix representation of customer-wise 15-minute energy consumption data presented in Section II-A (see Figure 1a). Our assumption is that by grouping daily observations together (i) daily patterns can be observed, (ii) the persistence or sift of such patterns over the course of a month, semester, or year can be studied, (iii) consumption can be compared over the years.

Despite some variability, Figure 2 demonstrates a distinguishable pattern that persists across days and also across the years: consumption increases during the course of a day and peaks around the 60-th time slot ($\sim$ 3pm). Periods of reduced consumption (e.g., during Spring 2013) compared to the average behavior, or inversely, increased consumption (e.g., during Summer 2012) can also be identified. Continuing our analysis, we present daily consumption observations over a course of a year for three buildings of different types in Figure 3. From Figure 3, it can is seen that the electricity consumption of building 54 drops significantly during summer. Instead, building 68 demonstrates a considerably stable consumption pattern throughout the year, whereas, building 106 exhibits a different consumption pattern with its peak consumption period being in the evening and late at night.

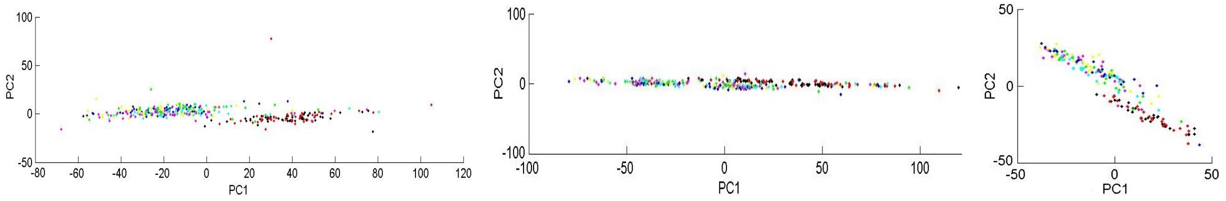
Visualizing the high resolution historical data confirms our hypothesis that valuable behavioral patterns can be mined and their evolution can been tracked over time to understand shifts in behavior, lifestyle or other customer characteristics. It also motivates the need for an automated, principled way to perform such analysis.

III. Consumption Patterns Mining

A. Principal Component Analysis

Given many vectors in a D-dimentional space, how can we visualize them when dimentionality D is high? More importantly, is it possible to group high resolution electricity consumption data acquired over a number of years for a large number of customers efficiently? We argue here that even though clustering methods can be directly applied to raw electricity consumption data, this is inefficient as it requires storage and processing of high dimensional and high volume data. Hence, it would be beneficial to cluster consumption data in a space of reduced dimension. To address this gap we apply Principal Component Analysis (PCA) [18] on our large-scale, real-world dataset using the representations in Section II-A. Our goal is to express electricity consumption data in a way that enables the identification of tacit patterns, highlights their similarities and magnifies their differences. As a side effect, we use PCA for data compression.

PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal



(a) 2009

(b) 2011

(c) 2013

Figure 4: PCA on energy consumption data in matrix format (see Figure 1a) per building, over a period of three years. Color scheme (better seen in color): Monday, Tueday,Wednesday, Thursday, Friday, Saturday, Sunday.

components. Particularly, PCA transforms the data to a new coordinate system such that the first principal component accounts for as much of the variability in the data as possible, and each succeeding component in turn has the highest variance possible under the constraint that it is uncorrelated with the preceding components.

We experimented with the various representations detailed in Section II-A. We discuss our findings next.

Our first experiment involved performing PCA on each matrix formed for each building where rows represent days and columns time of the day (see Figure 1a). Figure 4 shows the results plotted on the first two principal components for a Building 1 for three years (due to space limitations). Data points are colored to indicate days of the week so as identify underlying patterns across days. Since no distinct cluster formation is observable, Figure 4 supports our assumption that energy consumption values across days are similar and do not vary significantly. We do notice a distinct separation between weekdays and weekends however; this means that energy consumption follows very different consumption patterns during the week and weekends. Intuitively, this makes sense for a campus building which thrives with students during the week but has limited activity during the weekend. The differentiation between weekdays and weekends is consistent across the year.

Next, we performed PCA on each matrix formed for each day of the week for all buildings in our dataset simultaneously (see Figure 1b). Figures 5, 6, and 7 summarize the results plotted on the first two principal components for each day, for all buildings, for year 2012. Due to space limitations we refrain from presenting results for all years and also for other principal components[4]. Our findings are consistent for all five years in our dataset however, hence we assume them to be robust. Data points represent daily observation vectors across the two main components and are colored to distinguish between buildings.

Our goal is to uncover patterns (i) across days for each building, and (ii) between buildings for a day of the week. Figures 5, 6, and 7 demonstrate two interesting patterns in our dataset. First, the electricity consumption values for any given building form a very well formed group, suggesting that the energy consumption needs remain similar across a semester (e.g. Spring) for a given day of the week (e.g. every Monday). The same result can be verified for the course of the year by comparing the data point clouds corresponding to individual buildings for Spring, Summer, and Fall for the same day as in Figure 8. Furthermore, a observable variation between the energy needs of buildings during weekdays and weekends can be observed, further validating our discussion around Figure 4.

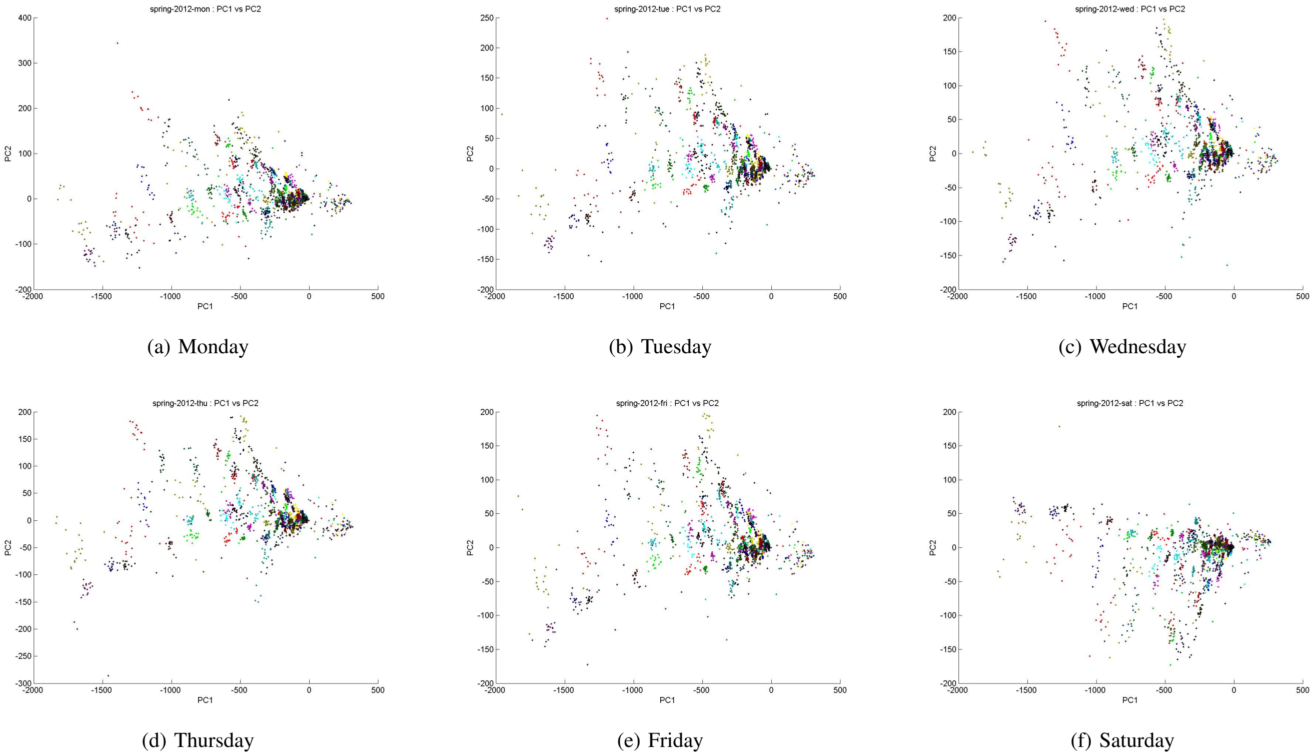
Figures 5, 6, and 7 also expose similarities in consumption between buildings; this means that buildings that naturally cluster together in the first two principal components share similarities on certain levels. For example, buildings with same function type (such as classrooms, office buildings, or dormitories) are expected to follow similar schedules in an academic environment thus exhibiting similar characteristics in their consumption. Lifestyle, appliances or other household characteristics can thus be predicted by consumption data [9] as inferred by clustering consumption time series on an appropriately transformed space.

We conjecture that PCA of appropriately organized data exposes hidden trails in electricity consumption data which would remain hidden and therefore unexploited otherwise. Moreover, instead of relying on 96 dimensions for our analysis, four dimensions (actually two principal components provide an adequately good approximation) are sufficient for describing 95.83 % of the data (97.9 % of the variance lies in the first two principal components) and the implicit patterns in it resulting into $95,8\%$ compression (i.e., 4 instead of 96 dimensions).

IV. Clustering of Consumption Patterns

In Section III we manually annotated Figure 7a to highlight major clusters. We argued there that a distinct separation between such clouds can be observed indicating different consumption patterns among buildings but also similarities between (i) the consumption characteristics of a given building for various instances of the same day of the week (e.g., Monday), and (ii) between daily observation vectors of different buildings. In this section, we propose to automate this tedious and laborious process using clus-

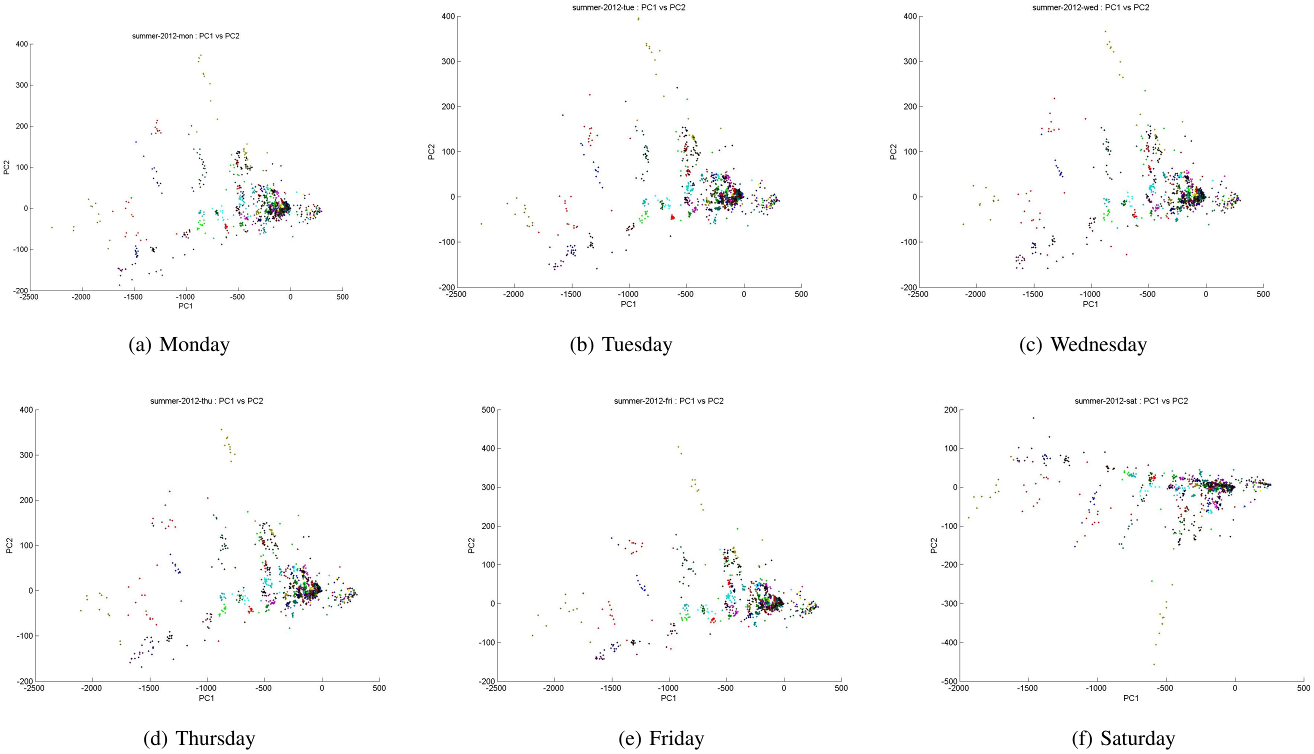
[4] We found that the first two principal components account for 90% of the data variability.



(a) Monday (b) Tuesday (c) Wednesday

(d) Thursday (e) Friday (f) Saturday

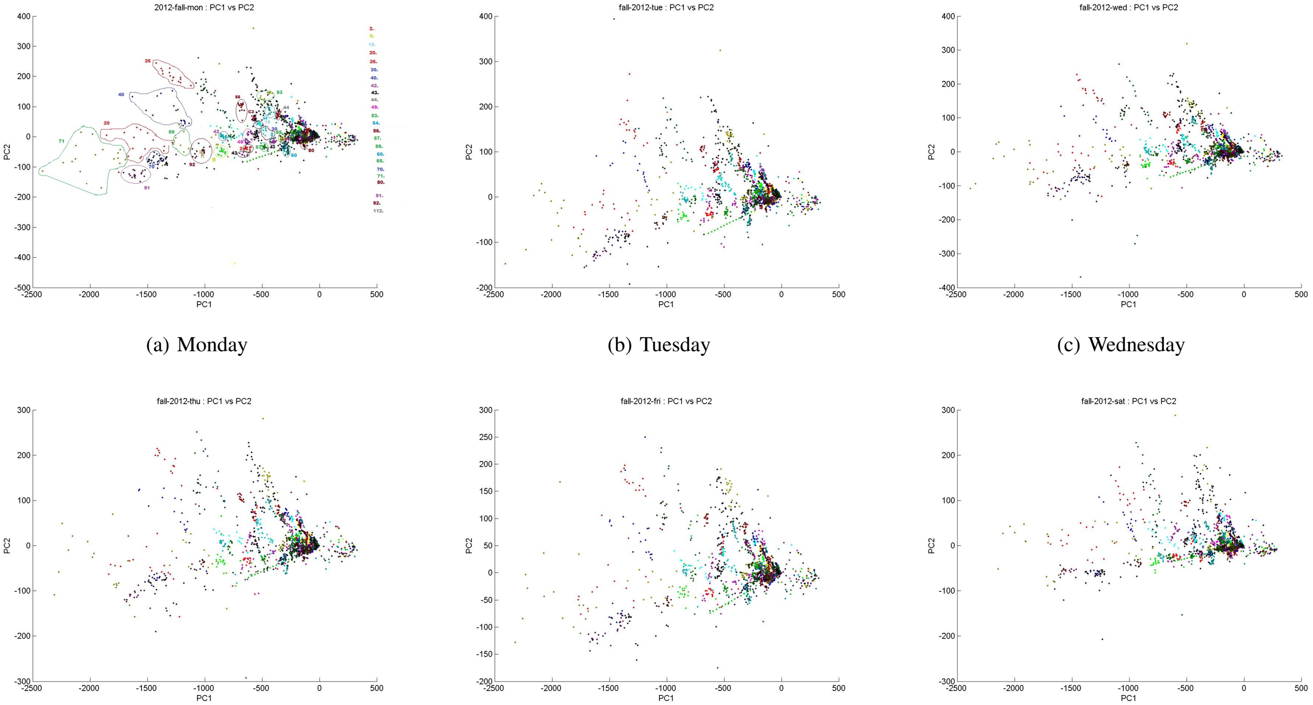
Figure 5: PCA on energy consumption data stashed for each day-of-the-week (see Figure 1b) for Spring semester of 2012.



(a) Monday (b) Tuesday (c) Wednesday

(d) Thursday (e) Friday (f) Saturday

Figure 6: PCA on energy consumption data stashed for each day-of-the-week (see Figure 1b) for Summer semester of 2012.



(a) Monday

(b) Tuesday

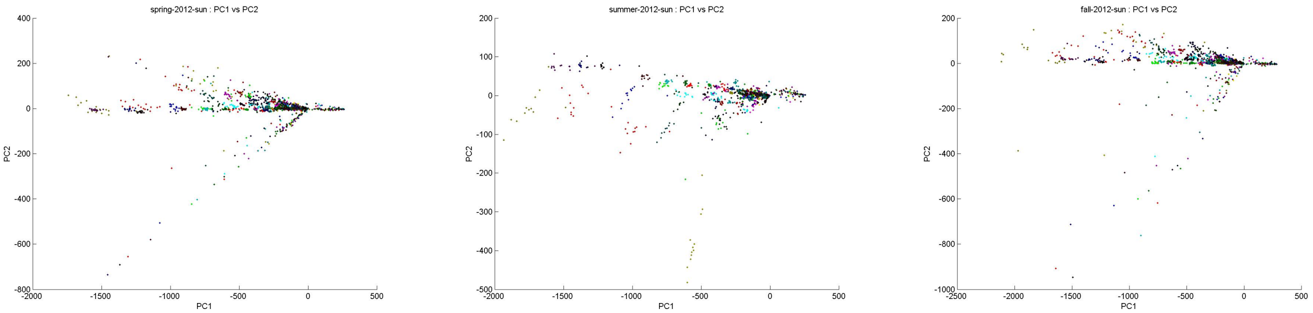
(c) Wednesday

(d) Thursday

(e) Friday

(f) Saturday

Figure 7: PCA on energy consumption data stashed for each day-of-the-week (see Figure 1b) for Fall semester of 2012.



(a) Spring

(b) Summer

(c) Fall

Figure 8: PCA on Sunday energy consumption data (see Figure 1b) for Fall semester of 2012.

ter analysis to identify buildings with similar consumption characteristics.

Clustering electricity consumption data in K groups such that the demand curves of the days belonging to a cluster are similar among them and dissimilar to the curves of those days belonging to other clusters according to some distance is challenging for numerous reasons. First, there is a great number of distance metrics that can be considered. Second, the number of possible patterns is unbounded. Third, we argued in Section II-A that multiple levels of granularity and representation may result in different clustering configurations and a plethora of interpretations. To address these challenges we venture to address the questions of which clustering technique should be chosen, how many clusters should be created by considering a variety of clustering methods.

A. Clustering Methods

1) K-means Clustering: K-means [19] partitions N observations into k disjoint subsets such that the intra-cluster distance between observations belonging to a cluster and the point designated as cluster centroid is minimized. Specifically, K-means partitions the data space into Voronoi cells such that the distance between a data point and the geometric center of its Voronoi cell is lees than the distance to the centers of other cells [5]. Euclidean distance is used as the distance metric, and variance as a measure of cluster scatter. K-means is a greedy algorithm and as such its performance depends on the appropriate selection of the initial cluster centers [20]. A proper number of clusters K is also hard to be determined beforehand; setting K to some value without proper reasoning is not appropriate.

2) Hierarchical Clustering: Hierarchical Clustering [21] is typically used to build a binary tree representation of a dataset by successively merging similar groups of observations without requiring a predetermined number of clusters. There are two approaches for hierarchical clustering: agglomerative and divisive. The agglomerative hierarchical clustering, which we use here, recursively combines clusters until a single data point remains.

3) Hausdorff-based K-medoids Clustering: K-medoids [22] is similar to K-means, but K-medoids is more robust to noise and outliers as compared to K-means due to the fact that it minimizes the sum of pairwise dissimilarities instead of a sum of squared Euclidean distances. In contrast to K-means, a medoid is chosen as the representative item for each cluster at each iteration by identifying an observation within the cluster that minimizes the sum of distances to all other objects in the cluster.

To avoid clustering individual consumption values for each customer for individual time slots, we propose a modified K-medoids algorithm based on Hausdorff distance [23]. Our proposed algorithm proceeds as the standard K-medoids algorithm except for evaluating the cluster centroids differently. Specifically, instead of considering the distance matrix that K-medoids employs, we instead consider the absolute distance values between electricity consumption observations as computed by the Hausdorff distance.

Hausdorff distance: Hausdorff distance [23] measures how far two subsets are from each other; two sets are close if every point of either set is close to some point of the other set. Let $A = \{a_1, a_2, \ldots, a_m\}$ and $B = b_1, b_2, \ldots, b_n$ be two non-empty subsets of a metric space. Their Hausdorff distance $dH(A, B)$ is calculated by computing the shortest distance between each feature a_i in set A with respect to features in set B, and then maintain the largest value. In other words, Hausdorff distance is the greatest of all distances from a point a in one set to the closest point b in the other set. Formally,

$$d_H(A,B) = \max\{\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in B} d(a,b)\}, \quad (1)$$

where sup represents the supremum and inf the infimum. As it stands, $d_H(A, B)$ is not always symmetric. Therefore, we consider the Hausdorff distance to be: $d_H(A, B) = \max\{d_H(A, B), d(B, A)\}$[5].

B. Voronoi Decomposition

Voronoi Decomposition [24] is used to partition a space into regions based on distance from a set of points (often referred to as seeds) which is specified beforehand. For each such seed there is a corresponding region, called Voronoi cell, consisting of all points closer to that seed than to any other. The Voronoi decomposition is dual to Delaunay triangulation according to which three nearest data points are computed and triangles are formed iteratively. Voronoi decomposition tries to maximize the minimum angle in the triangles formed. A circumcircle is drawn for each triangle. The circumcenter may or may not lie in interior of a triangle. Circumcenters lying in adjacent triangles are connected using line segment. Such line segments form a closed region called Voronoi Region. A 2-D Delaunay triangulation ensures that the circumcircle associated with each triangle contains no other point in its interior. This definition extends naturally to higher dimensions.

We are particularly emphasizing on the use of Voronoi Decomposition as we found it very efficient in the task of detecting outliers. Intuitively, the presence of outliers can be visualized using Voronoi diagrams. More importantly, outlier detection can be automated by identifying large and perhaps unbounded regions in the Voronoi Decomposition. This is in turn useful for data denoising, which can be applied as a pre-processing step before time-series clustering.

C. Optimal Number of Clusters

The most challenging problem of clustering has invariably been to select the right number of clusters [20], [5]. When ground trough is unavailable then the best number of clusters is impossible to find. For these reasons, we apply three indices to determine the optimal number of clusters in a clustering configuration task and also evaluate cluster validity, instead of using an arbitrary a-priori number of clusters K. We detail these indices in the following paragraphs.

Dunn Index (DI): Despite the plethora of cluster validity indexes, we select Dunns Index [20] as a standard metric for cluster evaluation when ground trough is unavailable. Dunn Index, an internal evaluation scheme (i.e., the result is based on the clustered data itself), evaluates clusters based on two criteria: (i) minimum intra-cluster distance and (ii) maximum inter-cluster distance. For a given clustering assignment, a higher Dunn index indicates better clustering.

We first derive the minimum distance between points of different clusters: $d_{min} = \min_{k\neq k'} d_{kk'} = \min_{i\in I_k j\in I_{k'}} \|M_i^k - M_j^{k'}\|$, where $M_1, \ldots, M_n$ are the data points to be clustered, and $d_{kk'}$ is the distance between clusters C_k and $C_{k'}$ as measured by the distance between their closest points. For each cluster C_k, we further compute the largest within-cluster distance $d_{max} = \max_{1\leq k\leq K} D_k = \max_{i\neq j\in I_k} \|M_i^k - M_j^{k'}\|$, where D_k is cluster's k diameter, i.e., the largest distance

[5] We also experimented with the mean of the minimum pairwise distances between data points of two subsets.

separating two distinct points in the cluster. The Dunn index is then calculated as the quotient of dmin and dmax.

Calinski Harabasz Index (CHI): Calinski Harabasz index [20] measures data variance by considering between-cluster (SSB) and within-cluster variance (SSW). The optimal number of K is obtained when the value of CHI(K) is maximized. Formally, $CHI(K) = \frac{\sum_{i=1}^{K} n_i \|m-m_i\|^2}{\sum_{i=1}^{K} \sum_{x \in c_i} \|x-m_i\|^2} \times \frac{(N-K)}{K-1}$, where $SS_B = \sum_{i=1}^{K} n_i \|m - m_i\|^2$ is the overall between-cluster variance, overall within-cluster variance is denoted by $SS_W = \sum_{i=1}^{K} \sum_{x \in c_i} \|x - m_i\|^2$, N is the number of observations, K is the number of clusters, c_i is the i^{th} cluster, m_i is the centroid of cluster i, m is the overall mean of the sample data, x is a data point, and $\| \cdot \|$ denotes the l^2-norm.

Energy Variance Index (EVI): We introduce a domain specific metric for cluster evaluation that measures variability in energy consumption values between observations belonging to same cluster. Intuitively, daily observation vectors should end up in the same cluster for low variable customers. Similarly, customers with similar consumption patters might be grouped together. By considering the energy difference between instances belonging to the same cluster, we can avoid grouping together customers that have similar consumption patterns but different magnitude scales. For example, two customers that exhibit the same observed pattern (e.g. mid-afternoon peak consumption) with one being ten times the magnitude of the other (e.g., a commercial and a residential customer) might have identical consumption behavior, but entail very different treatment by utilities for DR purposes. Utility would save time and money by focusing their efforts on customers who are not only positioned to reduce peak load when needed most (e.g., during mid-afternoon), but also by identifying customers with the highest potential impact in consumption shedding (i.e., large commercial entities instead of residential loads). Specifically, we compute EVI(K) as follows:

- **Step 1**: Cluster the electricity demand dataset using one of the methods described in Section IV-A
- **Step 2**: Identify data points belonging to clusters
- **Step 3**: For each cluster, formulate an energy difference matrix where (a, b) represents absolute difference in energy from data point a to data point b.
- **Step 4**: Compute sum of upper triangle in the matrix. If point (a, b) has been considered for evaluation, (b, a) is excluded to avoid redundant calculations.
- **Step 5**: Repeat Step 2 and Step 3 for all clusters.
- **Step 6**: Compute the intra-cluster sum of absolute energy differences.

V. Evaluation

In this section we discuss the clustering performance of the methods presented in Section IV-A. All experiments were carried out on a 64-bit Windows PC with 6 GB RAM, 2.5 GHz i5 processor. We used the indices from Section IV-C to access clustering validity and evaluate the optimal number of clusters in each case.

Next, we discuss the results we obtained for Monday, Spring semester in 2012. Although we performed experiments for each day of the week, for each of the three semesters, for all five years in our dataset, we refrain from presenting these results here due to space limitations. However, we note that our observations are consistent across experiments, hence we believe our conclusions to be robust. Figure 9 shows the results for K-means, Hausdorff-based K-medoids, and Voronoi Decomposition.

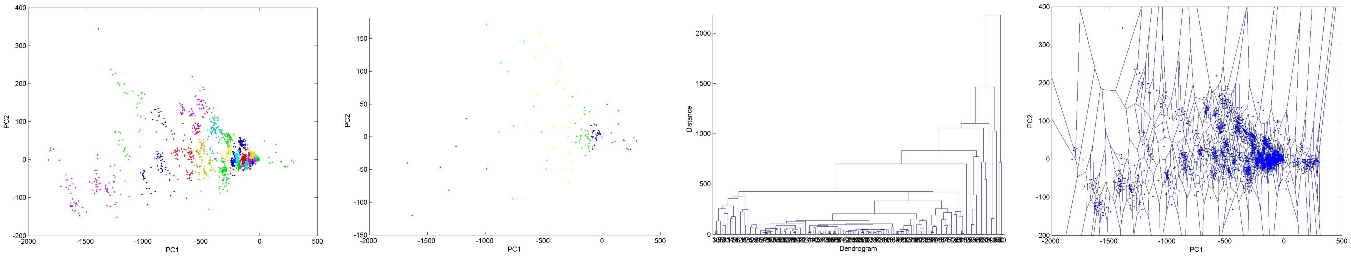
Figure 9 shows that points belonging to one building may be distributed to different clusters according to K-means, which ignores the fact that data points are correlated since they come from the same building, although from different instances for the same day of the week. Further, different K values may yield different results, whereas choosing an optimal seed set to initialize the algorithm is challenging. We leave this interesting research directions as future work.

To avoid clustering individual daily consumption observations for a given building, which can in turn result in a building participating in numerous clusters, we used the Hausdorff-based K-medoids algorithm to determine one point per building instead. We found the two variations we considered[6] to produce similar results.

We used agglomerative hierarchical clustering and the voronoi decomposition to empirically evaluate the clustering results. The agglomerative approach makes sense because unification of buildings into a cluster leads to a tree structure, the height of which can be controlled based on a variety of features (e.g., spatial distance). We further used the Voronoi decomposition as an effective way to identify outliers, which form open regions that cover more area than the norm. In addition to visual inspection, we performed validity analysis using the various indices discussed in section IV-C. Figure 10 shows the results. The optimal number of clusters is 6 according to DI (if we exclude 1 as the trivial solution), and 4 as per EVI. This agrees with the observed 6 well formed clusters in Figure 9b for Hausdorff distance K-medoids). Instead, according to CHI the optimal choice of k is 25.

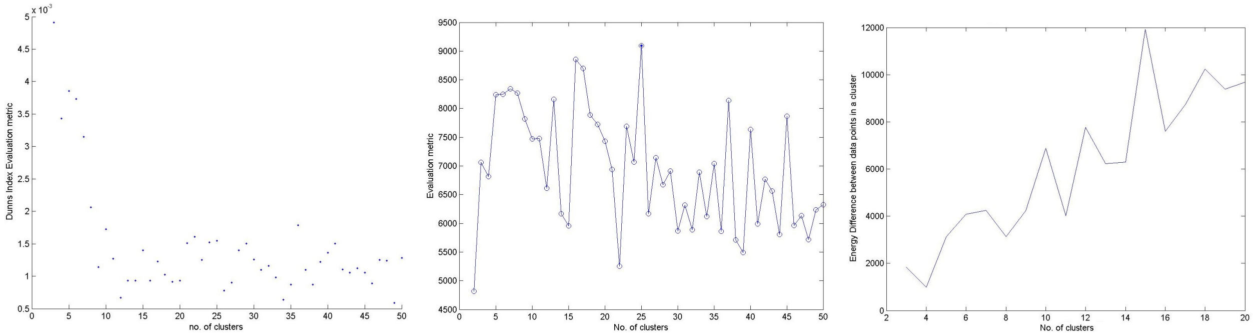
As there is no ground truth available for this dataset, we can only speculate about the results. Intuitively, naively applying k-means to the observation vectors does not leverage the fact that consumption observations for a building are correlated. Instead, the Hausdorff-based K-medoids is capable of identifying good clusters of similar buildings by operating on sets of observations and their respective distances rather than considering individual points. EVI is also useful to consider in this context as another measure of clutering validity, especially when the variability in the

[6] Based on Equation 1 or the mean distance to calculate the distance matrix (see Section IV-A3).



(a) K-means (b) K-medoids (c) Hierarchical Clustering (d) Voronoi

Figure 9: Clustering Results obtained from the methods described in Section IV-A.



(a) Dunn Index (b) Calinski Harabasz Index (c) Energy variance Index

Figure 10: Comparison of Clustering Results obtained from K-means using different cluster validity indices.

scale of energy consumption of consumers being grouped together. A methodology to efficiently divide the consumer base into appropriate bins using EVI is an interesting direction which we intent to explore in future work.

VI. Conclusions

We explored temporal patterns arising in electricity consumption time-series data using a real-world, large-scale dataset. We motivated the need for alternate representations of electricity consumption data, arguing that approached based on time-series representations are unable to mine implicit temporal patterns over a collection of high resolution consumption data from a diverse consumer base. We showed that usage behavior patterns can be identified at (i) different times-of-day, (ii) days-of-the-week, or (iii) at coarser granularities (i.e., by semester or yearly) for a customer. We also showed that similarities can be mined between customers with phenomenally different characteristics by appropriately clustering time-series data in a principal components space. We applied numerous clustering algorithms over a space of reduced dimensionality to segment daily consumption observations and buildings (i.e., consumers) alike.

We developed a novel algorithm for time series clustering based on Hausdorff distance that efficiently clusters buildings under our distance metric and data stashing technique. The proposed method scales to large data sets, and does not have to be confined to electricity consumption data. Instead, our stashing and clustering approach can be applied to any application that involves high dimensional time-series data.

Our findings have important implications for utility-side processing and storage of high velocity, high resolution electricity consumption time-series data. Beyond customer segmentation and pattern analysis, the entropy (i.e., variability) of consumption within a smart meter can yield further understanding of customers characteristics and lifestyles, which can ultimately be used for making more informed targeting decisions for Demand Response.

A limitation of our work is that clusters formed by the K-medoid (also the K-means algorithm) are highly dependent on the choice of seeds. Due to lack of standard methods for choice of seeds, this domain is open for interesting future work. As no ground truth for clusters is available in our dataset, choosing appropriate seeds becomes even more complicated. Although we have experimented with a variety of methods for seed selection, we did not reach conclusive results and hence we refrained from discussing them.

Finally, we experimented with applying Voronoi decomposition in the task of outlier detection with encouraging preliminary results. Even though in our analysis we did not observe differentiation in the value of Dunn index, the Calinski Harabasz index resulted in higher values of k, corroborating visual inspection. More importantly, we feel that efficiently dividing the consumer base into appropriate bins using EVI is an interesting direction, which we intent to explore in future work.

Acknowledgment

This material is based upon work supported by the United States Department of Energy under Award Number number DE-OE0000192, and the Los Angeles Department of Water and Power (LA DWP). The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, the LA DWP, nor any of their employees. The authors would like to thank Dr. Anand Panangadan and Ajitesh Srivastava for their constructive feedback on this work.

References

[1] Z. Fan, P. Kulkarni, S. Gormus, C. Efthymiou, G. Kalogridis, M. Sooriyabandara, Z. Zhu, S. Lambotharan, and W. H. Chin, "Smart grid communications: Overview of research challenges, solutions, and standardization activities," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 1, pp. 21–38, 2013.

[2] Y. Simmhan and M. U. Noor, "Scalable prediction of energy consumption using incremental time series clustering," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 29–36.

[3] J. Kwac and R. Rajagopal, "Demand response targeting using big data analytics," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 683–690.

[4] W. Shen, V. Babushkin, Z. Aung, and W. L. Woon, "An ensemble model for day-ahead electricity demand time series forecasting," in *Proceedings of the Fourth International Conference on Future Energy Systems*, ser. e-Energy '13. New York, NY, USA: ACM, 2013, pp. 51–62.

[5] C.-Y. Kuo, M.-F. Lee, C.-L. Fu, Y.-H. Ho, and L.-J. Chen, "An in-depth study of forecasting household electricity demand using realistic datasets," in *Proceedings of the 5th International Conference on Future Energy Systems*, ser. e-Energy '14. New York, NY, USA: ACM, 2014, pp. 145–155.

[6] H. Khadilkar, T. Ganu, Z. Charbiwala, L. C. Ming, S. Mathew, and D. P. Seetharam, "Algorithms for upgrading the resolution of aggregate energy meter data," in *Proceedings of the 5th International Conference on Future Energy Systems*, ser. e-Energy '14. New York, NY, USA: ACM, 2014, pp. 277–288.

[7] N. Armaroli and V. Balzani, "The future of energy supply: Challenges and opportunities," *Angewandte Chemie International Edition*, vol. 46, no. 1-2, pp. 52–66, 2007.

[8] K. Spees and L. Leve, "Impacts of responsive load in pjm: Load shifting and real time pricing," *Energy Journal*, vol. 29, no. 2, pp. 101–122, 2008.

[9] A. Albert and R. Rajagopal, "Smart meter driven segmentation: What your consumption says about you," *Power Systems, IEEE Transactions on*, vol. 28, no. 4, pp. 4019–4030, Nov 2013.

[10] T. Wijaya, T. Ganu, D. Chakraborty, K. Aberer, and D. Seetharam, "Consumer segmentation and knowledge extraction from smart meter and survey data," *Proceedings of the 2014 SIAM International Conference on Data Mining*, p. 9, 2014.

[11] C. Beckel, L. Sadamori, and S. Santini, "Automatic socio-economic classification of households using electricity consumption data," in *Proceedings of the Fourth International Conference on Future Energy Systems*, ser. e-Energy '13. New York, NY, USA: ACM, 2013, pp. 75–86.

[12] S. Depuru, L. Wang, and V. Devabhaktuni, "Smart meters for power grid: Challenges, issues, advantages and status," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 6, pp. 2736–2742, 2011.

[13] K. Palani, N. Nasir, V. C. Prakash, A. Chugh, R. Gupta, and K. Ramamritham, "Putting smart meters to work: Beyond the usual," in *Proceedings of the 5th International Conference on Future Energy Systems*, ser. e-Energy '14. New York, NY, USA: ACM, 2014, pp. 237–238.

[14] J. C. Lam, K. K. Wan, K. Cheung, and L. Yang, "Principal component analysis of electricity use in office buildings," *Energy and buildings*, vol. 40, no. 5, pp. 828–836, 2008.

[15] D. Ndiaye and K. Gabriel, "Principal component analysis of the electricity consumption in residential dwellings," *Energy and buildings*, vol. 43, no. 2, pp. 446–453, 2011.

[16] J. Wu, T. A. Reddy, and D. Claridge, "Statistical modeling of daily energy consumption in commercial buildings using multiple regression and principal component analysis," *Proc. 8th Syrup. Improving Building Systems in Hot and Humid Climates, Dallas, May 1992*, 1992.

[17] D. Ruch, L. Chen, J. S. Haberl, and D. E. Claridge, "A change-point principal component analysis (cp/pca) method for predicting energy usage in commercial buildings: the pca model," *Journal of solar energy engineering*, vol. 115, no. 2, pp. 77–84, 1993.

[18] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[19] J. Macqueen, "Some methods for classification and analysis of multivariate observations," *5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.

[20] K. Hammouda, "A comparative study of data clustering techniques," *International journal of computer science and information technology*, vol. 5, no. 2, pp. 220–231, 2008.

[21] P. Rodrigues, J. Gama, and J. P. Pedroso, "Lbf: Hierarchical time-series clustering for data streams," *Proceedings of the 1st International Workshop on Knowledge Discovery in Data Streams*, pp. 22–31, 2004.

[22] L. Kaufman and P. Rousseeuw, *Clustering by means of medoids*. North-Holland, 1987.

[23] R. T. Rockafellar and R. J.-B. Wets, "Variational analysis," p. 117, 1998.

[24] F. Aurenhammer, "Voronoi diagrams – a survey of a fundamental geometric data structure," *ACM Computing Surveys (CSUR)*, vol. 23, no. 3, pp. 345–405, 1991.