Alonso, A. M., Nogales, F. J., & Ruiz, C. (2020). Hierarchical Clustering for Smart Meter Electricity Loads Based on Quantile Autocovariances. *IEEE Transactions on Smart Grid*,11 (5), pp. 4522-4530.

DOI: 10.1109/tsg.2020.2991316

Hierarchical Clustering for Smart Meter Electricity Loads Based on Quantile Autocovariances

Andrés M. Alonso, F. Javier Nogales and Carlos Ruiz

Abstract

In order to improve the efficiency and sustainability of electricity systems, most countries worldwide are deploying advanced metering infrastructures, and in particular household smart meters, in the residential sector. This technology is able to record electricity load time series at a very high frequency rates, information that can be exploited to develop new clustering models to group individual households by similar consumptions patterns. To this end, in this work we propose three hierarchical clustering methodologies that allow capturing different characteristics of the time series. These are based on a set of "dissimilarity" measures computed over different features: quantile auto-covariances, and simple and partial autocorrelations. The main advantage is that they allow summarizing each time series in a few representative features so that they are computationally efficient, robust against outliers, easy to automatize, and scalable to hundreds of thousands of smart meters series. We evaluate the performance of each clustering model in a real-world smart meter dataset with thousands of half-hourly time series. The results show how the obtained clusters identify relevant consumption behaviors of households and capture part of their geo-demographic segmentation. Moreover, we apply a supervised classification procedure to explore which features are more relevant to define each cluster.

Index Terms

Quantile auto-variances, massive time-series, hierarchical clustering, smart meters.

I. Introduction

A. Background and Aim

MOVED by the need of improving the efficiency and sustainability of aging electrical systems, many countries worldwide are adopting new information and communication technologies, with special emphasis on the residential sector [1]. These technologies imply a new paradigm in the economic and technical operation of distribution networks, and create new business opportunities for all the companies that take part in the electricity supply chain.

It is very relevant the extended integration of advanced metering infrastructures (AMI) [2] with a special role played by households "smart meters". These devices allow recording electricity consumption data at a very high frequency rate and instantly transmit this information to the retailing and/or distribution companies.

Furthermore, as many electricity markets worldwide are open to competition in both the generation and retailing sectors, there is a growing interest by the electrical companies in using these data to increase their profit, their market share or the consumers' welfare. In this vein, the treatment of these new datasets requires the research and implementation of novel data science techniques, with practical applications on energy fraud detection, outlier identification, consumers profiling, demand response, tariff design, load forecasting, etc. [3].

A. M. Alonso is with the Department of Statistics and Institute Flores de Lemus; F. J. Nogales and C. Ruiz are with the Department of Statistics and UC3M-BS Institute for Financial Big Data (IFiBiD), University Carlos III de Madrid, Spain. (e-mails: andres.alonso@uc3m.es; fcojavier.nogales@uc3m.es; carlos.ruiz@uc3m.es)

Accepted version at *IEEE Transactions on Smart Grid*

https://doi.org/10.1109/TSG.2020.2991316

The special characteristics of the data stored by smart meters (hundreds of thousands, or even millions, of high frequency time series), and their combination with exogenous variables (meteorological, calendar, economical, etc.), open the possibility of designing specific clustering models for household consumers. Furthermore, these models can help to better understand the behavior of both aggregated and disaggregated electrical loads [4], and how this knowledge can be exploited to improve electrical systems.

In particular, clustering households with similar consumption patterns has many potential applications. One of the most relevant is customer segmentation (see [5]), where energy retailers may use the clusters to improve their business decision making by offering personalized tariffs depending on the seasons. Other interesting application where clustering methodologies could be applied is demand response (DR) policies, see [6]. Energy retailers can plan and conduct these policies through the different load profiles from the clusters, saving both on physical energy demand (by automatic control in the smart grid) and electrical bills for customers (by adjusting consumption on electricity price). These techniques can also be applied to detect and reduce electricity theft, through the analysis of some small and anomalous groups with atypical and potentially malicious consumption.

Another important application of clustering techniques is to combine the obtained clusters with a forecasting procedure. That is, instead of individually predicting each time series by selecting a model, we can exploit that the series belonging to the same group have a similar dynamic behavior and we can obtain a joint model for these series. For example, in [7], it is shown that a dynamic factorial model with cluster structure improves the prediction performance of the classical factorial model. Similarly, the forecasting model may benefit from the hierarchical structure of a cluster. For instance, in [8] a forecasting tool is applied incrementally to each node of a hierarchical clustering similar to the one used in this paper. This approach does not need to define in advance the number of clusters and allows for forecasting at different granularity levels.

In this work we propose different hierarchical-based clustering strategies based on a set of “dissimilarity” measures: quantile auto-covariance, and simple and partial autocorrelations. The use of these measures for load clustering is, as far as we know, new in the smart grid literature.

These strategies summarize each consumption time series in only a few representative features so that they are highly efficient, easy to automatize and scalable to hundreds of thousands of series, i.e., can be successfully implemented in large-scale applications that make use of smart meters datasets. Another advantage on the use of these measures is that they are robust against possible outliers and/or reading errors, that most standard clustering methods would not be able to handle, and very present in smart meter time series.

We test the performance of our clustering models by using a real-world dataset with thousands of electricity consumption time series. The results are promising as the obtained clusters not only identify relevant consumption patterns but also capture part of the geo-demographic segmentation of the consumers. Moreover, an innovative contribution of this work is that we implement a multiclass supervised classification algorithm, based on decision trees, in order to characterize the most important features conditioning each cluster.

B. Literature Review

A review of several clustering techniques to group similar electricity consumers is presented in [9]. It is shown that the overall performance of the different techniques is related to their ability to isolate outliers. A clustering method for household consumers based on K-means and Principal Components Analysis (PCA) is proposed in [10]. The resulting clusters are subject to a multiple regression analysis to identify relevant explanatory variables. The work in [11] addresses the consumer segmentation problem by normalizing the daily load shapes for each consumer, together with their total consumption, to apply an adaptive K-means algorithm. A clustering model based on K-means is proposed in [12] to, focusing on commercial and industrial electricity consumers, identify candidate users for energy efficiency policies and their businesses opening and closing hours. The work in [13] evaluates and compare three clustering techniques for smart meter data: k-medoid, K-means and Self Organizing Maps (SOM), to show that the latter presented to overall best performance. Traditional time series methods are applied in [14], like wavelets or autocorrelation analysis, to the raw smart meter data to enrich the input of a K-mean based clustering algorithm for consumers segmentation. [15] proposes to use dynamic information, in terms of transitions between adjacent time periods, for consumers segmentation. The resulting clusters are used to evaluate their potential for demand response policies.

Several works seek to identify relevant features that condition the dynamic patterns of electricity consumers. For instance, [16] proposed a methodology to examine smart meter data and identify important determinants of consumers electricity load. A mixture model framework, based on linear Gaussian approximations, is used by [17] to derive relevant load profiles from individual consumption patterns. With the same aim, a supervised Machine Learning (ML) model is proposed in [18] based on individual household consumption time series. A detailed analysis of household consumption data is presented in [19] to identify those time periods from which relevant consumption features can be extracted. Based on these features, a mixture-based clustering algorithm is proposed and evaluated by bootstrap techniques. To extend the number of features that can potentially be used for profiling consumers, [20] complement the smart meter data with door-to-door question surveys. It is shown how this new dataset improves the performance of a Ward's hierarchical clustering algorithm.

Clustering models pose relevant computational challenges when the number of time series increases. Hence there are several works that focus on improving the efficiency of the clustering algorithms. An efficient frequency domain hierarchical clustering model is proposed [21] to derive adequate load profiles. Moreover, [22] studies how the temporal resolution of the consumption time series may have a strong impact on both the quality and computational performance of the clustering techniques. [23] proposes a feature construction model for time series to cluster similar consumers. The model reduces the dimensionality of the problem by using conditional filters and profile errors. With a similar aim, [24] presents a two-level clustering methodology to derive representative consumptions profiles based on K-means. The first level is used to obtain local profiles that are generalized in the second level.

Clustering techniques have been used also to improve the accuracy of forecasting models. A clustering K-Means based algorithm is employed in [25] to household load curves to group similar consumers and enhance the performance of a nonparametric functional wavelet-kernel approach. Similarly, [26], implements a K-means based clustering algorithm to group similar consumers and then adjust a Neural Network (NN) forecasting model for aggregated loads. A K-means based algorithm is employed in [27] to derive consumption estimates and impute missing data. The cross-similarities between consumptions series is used by [28] to enhance the performance of a forecasting model, based on Long Short-term Memory (LSTM) networks. [29] also makes used of consumers segmentation through PCA and K-means clustering to identify typical daily consumption profiles that can improve the accuracy of a ML forecasting tool.

C. Contributions

We build part of our research on the original methodology presented in [30], which proposes to cluster time series based on quantile autocovariances distances. An extensive simulation analysis and a real-world application on daily financial time series show the ability of this approach to identify different dependence models among the series.

In the present work, and by the first time to the authors knowledge, we adapt and extend part of the methodology in [30] to identify relevant clusters from a real-world dataset including thousands of smart meters time series.

Moreover, by considering the state of the art presented in Section I-B, the main contributions of this work are fourfold:

1) To summarize each smart meter time series in a small set of meaningful features: autocorrelation coefficients, partial autocorrelation coefficients and quantile autocovariances.

2) To propose three hierarchical clustering models, based on Euclidean dissimilarity measures, computed over the previous features. The models are computationally efficient and robust against outlier observations.
3) To test the proposed methodology in a real dataset, including thousands of half-hourly load time series, to: a) characterize relevant electricity consumption profiles and b) to verify that the resulting clusters are able to capture, up to some extent, the geo-demographic segmentation of household consumers.
4) To make use of a supervised classification procedure (decision trees) to identify those variables (features) that have been more relevant to form the resulting clusters.

D. *Paper Organization*

This paper is organized as follows. In Section II the proposed hierarchical clustering methodology for smart meter time series is presented. The numerical results, based on a real-world dataset, are presented in Section III. Finally, Section IV presents the main conclusions derived from this work.

II. The Clustering Methodology

Let's assume that we observe N time series, $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$ where $\boldsymbol{X}_i = (X_{i,t_i}, X_{i,t_i+1}, \ldots, X_{i,T_i})$ and (t_i, T_i) denotes the first and the last times where the i-th time series is observed, respectively. In our dataset, the (t_i, T_i) are the same for all time series but our procedures do not require this condition since they are based on extracted features from the time series. As mentioned in the previous section, there are many interesting features to consider as "clustering" variables instead of using raw data. In our case, we consider three sets of features that capture different aspects of the time series dynamic behavior:

- The set of *autocorrelation coefficients* of orders $(1, 2, \ldots, K)$, that is, we calculate the correlation coefficient between the variables $X_{i,t}$ and $X_{i,t+j}$ for $j = 1, 2, \ldots, K$ defined by

$$\rho_i(t, t+j) = \frac{Cov\left(X_{i,t}, X_{i,t+j}\right)}{(Var(X_{i,t})Var(X_{i,t+j}))^{1/2}}. \tag{1}$$

- The set of *partial autocorrelation coefficients* of orders $(1, 2, \ldots, K)$, that is, we calculate the correlation coefficient between observations separated by j periods, $X_{i,t}$ and $X_{i,t+j}$, when we eliminate the linear dependence due to intermediate values. The partial autocorrelation coefficient will be denoted by $\pi_i(t, t+j)$.
- The set of *quantile autocovariances* of order j at quantile levels $(\tau, \tau') \in [0, 1]^2$ defined by

$$\begin{aligned}\gamma_{i,(\tau,\tau')}(t, t+j) &= \\ &Cov\left(I(X_{i,t} \leq q_{\tau,i}), I(X_{i,t+j} \leq q_{\tau',i})\right),\end{aligned} \tag{2}$$

  where $I(\cdot)$ denotes the indicator function and $q_{\tau,i}$ and $q_{\tau',i}$ are the $\tau-$ and $\tau'-$quantiles of $X_{i,t}$ and $X_{i,t+j}$, respectively.

An exhaustive study of these simple and partial autocorrelation coefficients can be found in the excellent book [31] and for quantile autocovariances we refer the reader to [32] and [33].

It is interesting to realize the differences among features (1) and (2) since both involve the calculation of a covariance between observations separated by j periods. In (1), the covariance term is estimated by

$$\begin{aligned}&\frac{1}{T_i - j} \sum_{t=t_i}^{T_i - j} X_{i,t} X_{i,t+j} \\ &\quad - \frac{1}{T_i - j} \sum_{t=t_i}^{T_i - j} X_{i,t} * \frac{1}{T_i - j} \sum_{t=t_i}^{T_i - j} X_{i,t+j},\end{aligned}$$

which involves the products $X_{i,t}X_{i,t+j}$ that can be distorted by extreme or outlier observations. For example, two very high loads observed at a distance of j periods would spuriously increase the correlation at the $j-$lag. On the other hand, the quantile autocovariance (2) is estimated by

$$\widehat{\gamma}_{i,(\tau,\tau')}(t,t+j) = \\ \frac{1}{T_i - j}\sum_{t=t_i}^{T_i-j} I(X_{i,t} \leq \widehat{q}_{\tau,i})I(X_{i,t+j} \leq \widehat{q}_{\tau',i}) \ - \ \tau\tau'. \tag{3}$$

The involved products $I(X_{i,t} \leq \widehat{q}_{\tau,i})I(X_{i,t+j} \leq \widehat{q}_{\tau',i}$ are bounded which imply a negligible effect of outliers. The expression (3) can be interpreted as a mean of the number of times that values at t below $\widehat{q}_{\tau,i}$ coincide with values at $t+j$ below $\widehat{q}_{\tau',i}$. The term $\tau\tau'$ is the number of coincidences that occur completely randomly. Therefore, a positive $\widehat{\gamma}_{i,(\tau,\tau')}$ means that the number of matches is greater (smaller) than expected by chance.

It should be noticed that the above characteristics, in general, depend on t and j, but if the time series are stationary, then they do not depend on t, which simplifies their analysis. For this reason, we first consider the logarithmic transformation and then the seasonal (daily) difference of the smart meter's time series since the resulting time series can be considered stationary. That is, as the time series that will be used in this paper present a half-hourly frequency, then $X_{i,t} = \ell_{i,t} - \ell_{i,t-48}$ are the series to be clustered, where $\ell_{i,t} = \log L_{i,t}$ denotes the logarithm of the load time series of the i-th time series. We should fix the largest lag, K, in the sets of autocorrelation and partial autocorrelation coefficients. We can fit autoregressive models to all the univariate time series, selecting the order by the BIC criterion, and take $K = \max_{1\leq i\leq N}(p_i)$, where p_i is the selected order for i-th time series. It is shown in [34] that this procedure provides an upper bound of the memory of N stationary linear time series. The selected K was 96. This selection allows us to capture the main linear dependencies in all time series. Also, for the set of quantile autocovariances, we should fix the lag and quantile levels. In this case, following the suggestions of [30], we use $j = 1$ and $\tau \in \{0.1, 0.5, 0.9\}$ since these values have shown that they are capable of capturing and differentiating different types of nonlinearities. Finally, the three clustering analyzes will be based on the following sets of features:

a) $\boldsymbol{\rho}_i = \{\rho_i(1), \rho_i(2), \ldots, \rho_i(96)\}_{i\in\{1,2,\ldots,N\}}$
b) $\boldsymbol{\pi}_i = \{\pi_i(1), \pi_i(2), \ldots, \pi_i(96)\}_{i\in\{1,2,\ldots,N\}}$
c) $\boldsymbol{\gamma}_i = \{\gamma_i,(0.1,0.1),\ \gamma_i,(0.1,0.5),\ \gamma_i,(0.1,0.9),\ \gamma_i(0.5,0.1),\ \gamma_i(0.5,0.5),\ \gamma_i(0.5,0.9),\ \gamma_i(0.9,0.1),\ \gamma_i(0.9,0.5),\ \gamma_i(0.9,0.9)\}_{i\in\{1,2,\ldots,N\}}$

Thus, the analysis will be based on 96×1 vectors of features for autocorrelation and partial autocorrelation coefficients and based on 9×1 vectors of features for quantile autocovariances. Once we have the vectors of features, we define a dissimilarity measure between time series $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ by the Euclidean distance of the corresponding vectors. That is:

a) $d_{AC}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \|\boldsymbol{\rho}_i - \boldsymbol{\rho}_j\|_2$
b) $d_{PAC}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \|\boldsymbol{\pi}_i - \boldsymbol{\pi}_j\|_2$
c) $d_{QC}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \|\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_j\|_2$

where $\|\cdot\|$ denotes de Euclidean distance and AC, PAC and QC stand for autocorrelation, partial autocorrelation and quantile autocorrelation coefficients, respectively.

The distances $d_M(\boldsymbol{X}_i, \boldsymbol{X}_j)$ will be obtained for all pairs (i,j) with $i \neq j$ to construct the following $N\times N$ dissimilarity matrix

$$\boldsymbol{DM}_M = \\ \begin{pmatrix} 0 & d_M(X_1,X_2) & \ldots & d_M(X_1,X_N) \\ d_M(X_2,X_1) & 0 & \ldots & d_M(X_2,X_N) \\ \vdots & \vdots & \ddots & \vdots \\ d_M(X_N,X_1) & d_M(X_N,X_2) & \ldots & 0 \end{pmatrix} \tag{4}$$

where $M \in \{AC, PAC, QC\}$. The dissimilarity matrix (4) can be used in any cluster procedure which requires this kind of input. In particular, we can apply hierarchical clustering since it allows us to identify clusters as well as hierarchies among the clusters. In hierarchical cluster procedures, to decide which groups should be combined, it is necessary to choose a measure of dissimilarity (linkage criterion) between sets. It is important to emphasize that this choice will influence the shape of the groups, since some sets could be close according to one distance and far according to another. The three best known measures are minimum or single-linkage (d_s), maximum or complete-linkage (d_c) and average linkage (d_a) defined by:

$$d_s(A,B) = \min\{d(X_i, X_j) : i \in A, j \in B\}$$

$$d_c(A,B) = \max\{d(X_i, X_j) : i \in A, j \in B\}$$

$$d_a(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{i=1}^{n_B} d(X_i, X_j),$$

where A and B are two sets of observations having n_A and n_B elements, respectively.

In this work, we prefer to use complete linkage as it ensures that the observations in a group are “similar” to all observations of the same group in the sense that once the cut-off point in the dendrogram has been set all the distances within of a cluster are smaller than this cut-off point.

For the comparison of the different clustering results, we use the adjusted Rand index, ARI, which is based on counting pairs. The adjusted Rand index proposed by [35] compares two different cluster partition, $C = (C_1, \ldots, C_k)$ and $C' = (C'_1, \ldots, C'_{k'})$ using the following formulas:

$$ARI(C, C') = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k'} \binom{\#(C_i \bigcap C'_j)}{2} - \sum_{i=1}^{k} \binom{\#(C_i)}{2} \sum_{j=1}^{k'} \binom{\#(C'_j)}{2} / \binom{n}{2}}{\left(\sum_{i=1}^{k} \binom{\#(C_i)}{2} + \sum_{j=1}^{k'} \binom{\#(C'_j)}{2} \right) / 2 - \sum_{i=1}^{k} \binom{\#(C_i)}{2} \sum_{j=1}^{k'} \binom{\#(C'_j)}{2} / \binom{n}{2}}.$$

The closer to one the index, the higher is the agreement between the two partitions.

Once we obtain the groups, an interesting question is to know which variables have been the most relevant to form these groups. This question can be addressed through the use of a supervised classification procedure where the labels of the observations will be the result of the clustering methodology. That is, if we have k clusters, we will assign the labels $\{1, 2, \ldots, k\}$ to the observations of the respective clusters. These labels and the features will be the input of the supervised classification procedure. In this work, we will use decision trees [36] for multiclass classification problem since for this procedure unbiased estimates of the predictor (feature) importance [37] are available. In a decision tree, the importance of a given feature (predictor), X, used to predict Y can be calculated as

$$I(X) = \sum_t \frac{N_t}{N} \delta i(t),$$

where the sum is on the nodes t that use feature X and $\delta i(t) = i(t) - \frac{N_L}{N_t} i(t_L) - \frac{N_R}{N_t} i(t_R)$, that is, the change of index i when the node t is splitted into the left node, t_L, and the right node, t_R. N_t, N_R and N_L are the number of observations at the nodes t, t_L and t_R, respectively. The index i is called impurity function and usual selection are the Gini index, the Shannon entropy, or the variance of Y (see, for instance, [38]). In addition, we can estimate the misclassification rate of the decision tree by a cross-validation procedure. The misclassification rate is the proportion of observations (time series) to which a different label from its “original” label is assigned. It should be noted that here the “original” labels are those created by the grouping procedure.

III. Numerical Results

In this section we use the public energy consumption dataset from [39]. It includes a sample of 5,567 households of London with their individual electricity consumption time series during 2013, in kWh (per half hour), date and time, and CACI ACORN segmentation (6 geo-demographic categories) [40]. In particular, to validate this work's clustering methodology, we will compare the resulting clusters with the geo-demographic aggregated categories coded as "ACORN_GROUPED", which classify households into three main groups: "Affluent", "Comfortable" and "Adversity". Moreover, the dataset is also divided into two subgroups of consumers:

i) *std tariff*: Consumers whose electricity tariff is fixed (standard) to a constant price during the time of the study.

ii) *tou tariff*: Consumers with "time of use" tariff for which the electricity price is different for each hour.

In order to better characterized the inherent consumption behavior of individual households, we have focused the following study on the std tariff consumers, as these are not influenced by a variable price signal. This initial group includes approximately 4500 time series from which some of them are discarded, due a high proportion of missing observation, rendering a final subsample of around 3200 time series (households) with readings from 01/01/2013 until 12/31/2013. It should be noted that some of the time series present different start or end dates within this range. However, the proposed clustering strategies do not need time series to be observed in exactly the same date range, and hence we can exploit the complete dataset. Nevertheless, the different correlation coefficients could also be obtained by restricting the observations to a shorter specific of the year, and hence distinguishing between seasonal consumption patterns. In this regard, the present approach can be considered as a joint measure that encompasses all seasons.

The first step of our procedure is to obtain the hierarchical structures using the QC, AC and PAC features and the complete linkage introduced in Section II. The usual way to represent these structures is through dendrograms that, for reasons of space, we have omitted but are available upon request to the authors. In the three graphs we have observed some clear groups of observations (time series) and also observations that join the hierarchical structure at large levels. Those observations have a dynamic "atypical" behavior and are grouped in clusters with less than 1% of the total number of time series. Once we discard the atypical observations, we find eight, six and seven large clusters for QC, AC and PAC, respectively. Moreover, the degrees of coincidence among these three clusters partitions are low as indicated by the adjusted Rand indexes (0.0941 when comparing QC and AC; 0.1432 when comparing QC and PAC and 0.2687 when comparing AC and PAC). This implies that the three approaches look at different characteristics of the time series.

Figures 1 - 3 illustrate these large clusters obtained with QC, AC and PAC, respectively. In the figures, we represent the mean of the features used to obtain the clusters. There are nine features, in the case of QC, corresponding to the covariance of quantiles 10%, 50% and 90%, i.e., (0.1 versus 0.1), (0.1 versus 0.5), $\ldots$, (0.9 versus 0.9). In the case of AC and PAC, we use the first 96 simple and partial autocorrelations, respectively. The clusters based on QC reveals differences in the median consumptions (.5 versus .5) and highest versus median consumptions (.9 versus .5). For instance, there is a remarkable difference between c^3_{QC} and c^7_{QC} versus c^1_{QC}, c^2_{QC}, c^4_{QC}, c^6_{QC} and c^8_{QC} at the median consumptions (.5 versus .5). The c^3_{QC} and c^7_{QC} have negative covariances and the c^1_{QC}, c^2_{QC}, c^4_{QC}, c^6_{QC} and c^8_{QC} have positive ones. That is, in the first group, a consumption below the median tends to be followed by consumption above the median, while the second group tends to maintain their consumption below the median. The groups by SAC and PAC show differences in the short-range dependencies but also in the way they are around the lag 48 (one day). We can focus on the first correlations coefficients that show different degrees of persistency in the consumptions. For instance, in the AC clusters, there is a clear order from high dependency at c^4_{AC}, c^1_{AC} and c^2_{AC}, medium at c^3_{AC} and c^5_{AC} and to low dependency at c^6_{AC}. At the PAC clusters, we can differentiate between clusters with negative second partial autocorrelation (c^1_{PAC}, c^2_{PAC}

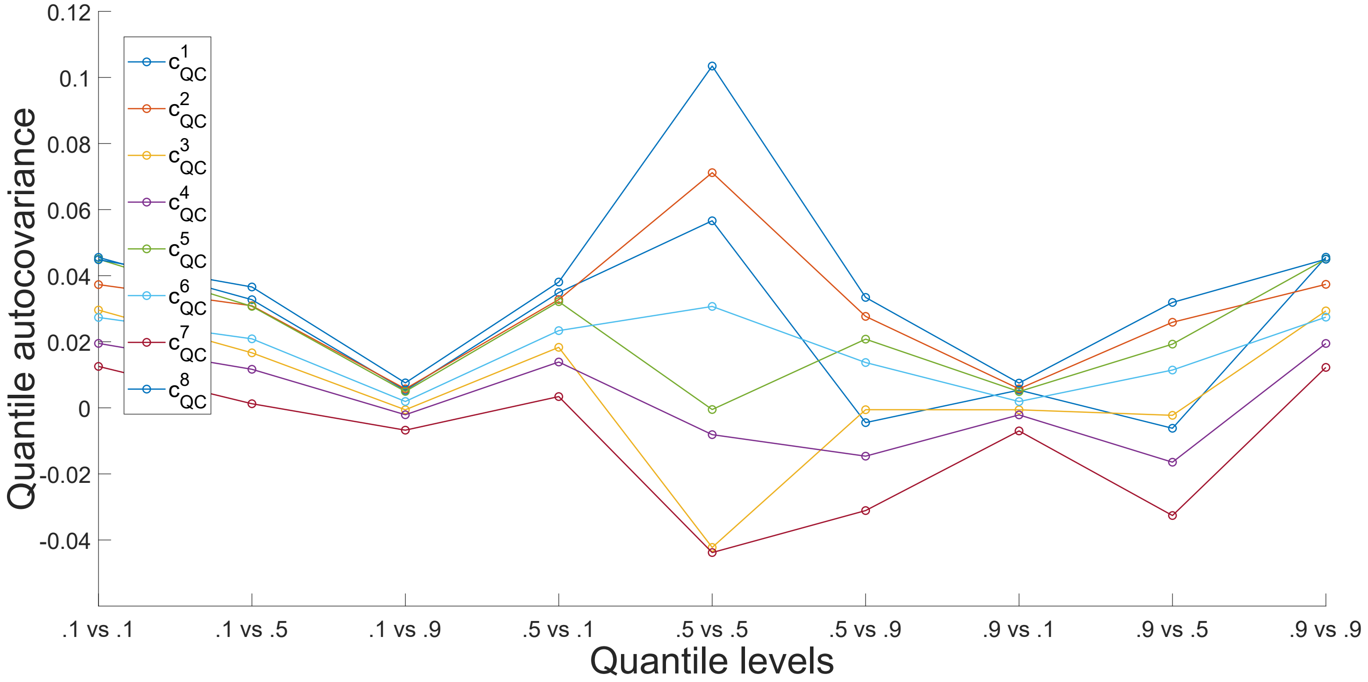

Figure 1. Main clusters obtained with QC and complete linkage.

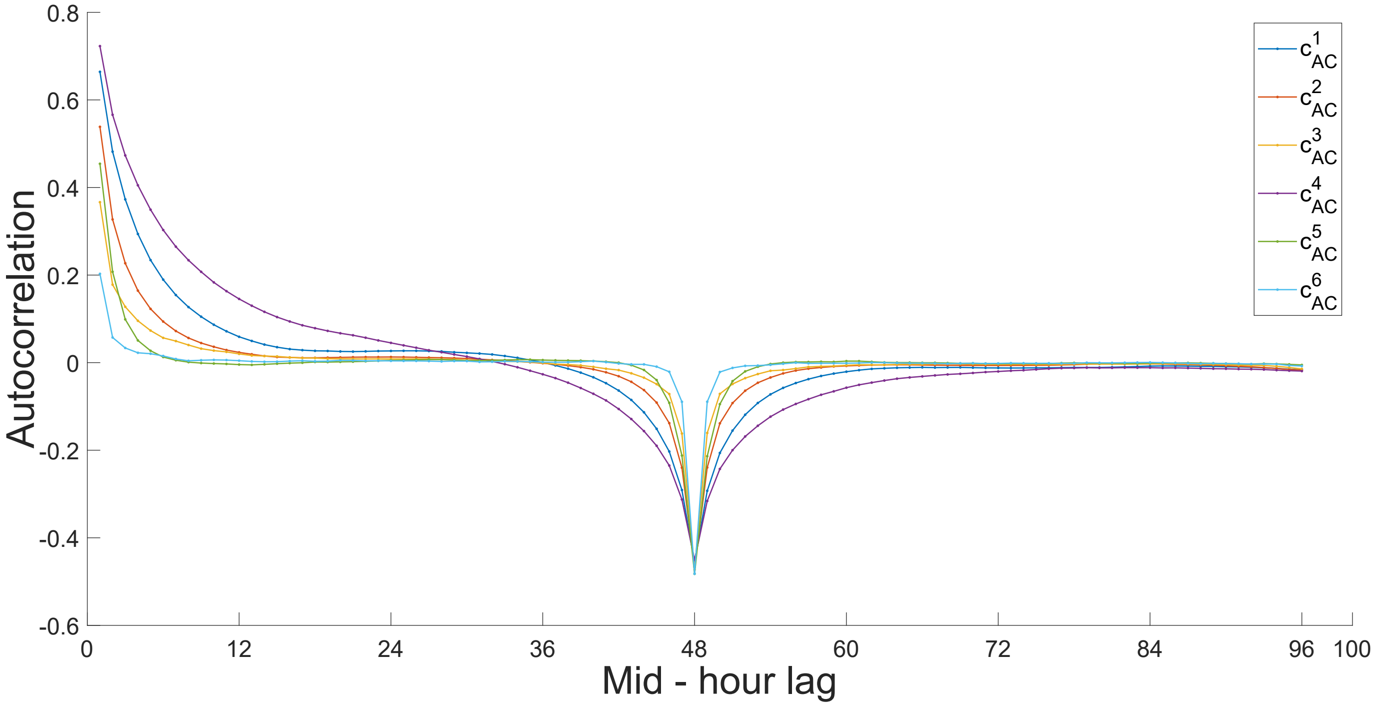

Figure 2. Main clusters obtained with AC and complete linkage.

and c^6_{PAC}), medium (c^3_{PAC}, c^4_{PAC}, and c^5_{PAC}) and high positive (c^7_{PAC}). That is, once we eliminate the first order correlation, there are negative (or positive) direct effects on the consumption at the two-step ahead period.

To perform a cluster evaluation, in Table I, we report the two internal validity criteria suggested by [41], the cophenetic correlation coefficient (CCC) which allows us to evaluate the hierarchy of clustering schemes and the normalized Γ statistic (NGS) which allows to evaluate the degree of agreement between a

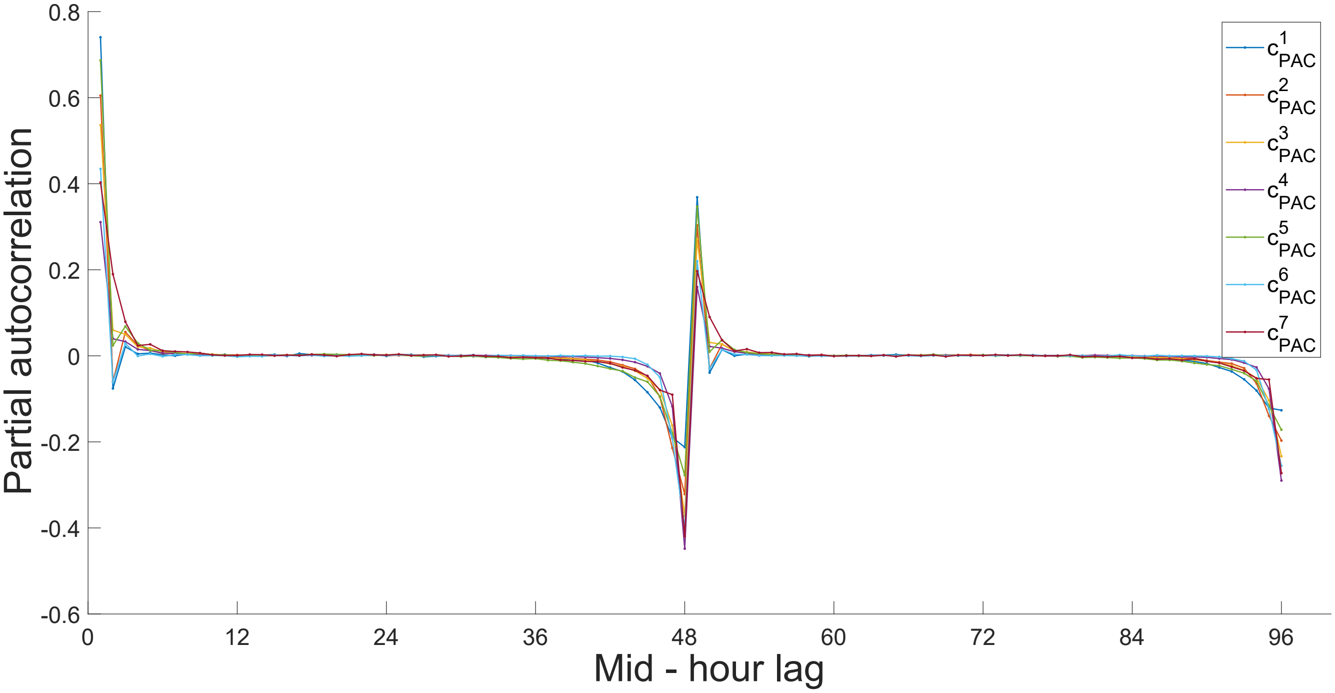

Figure 3. Main clusters obtained with PAC and complete linkage.

	QAC solution	SAC solution	PAC solution
CCC	.669	.435	.517
NΓS	.737	.713	.722
1-NN	.024	.120	.121

Table I
Validity metrics for the formed clusters

given clustering solution and the distance or dissimilarity matrix. Both measures are correlation coefficients having values in (-1,1), the closer the value to one the better the clustering solutions. Additionally, we perform a leave-one-out cross-validation procedure with one-nearest neighbor classifier (see, for instance, [42])[1]. We also report the miss-classification rates (1-NN) of this procedure at the table. The lower the rates, the better the clustering solutions. These values support the conclusion that the cluster solutions we have obtained are satisfactory representation of the underlying grouping structure.

Moreover, we compare our clustering solutions with the ones obtained with some state-of-the-art techniques such k-means and partition around medoids (PAM). For both methods, we select the same number of clusters than in our methods and we use 100 random initializations. In Table II, we report the Rand index between our solutions and the k-means and PAM's solutions. Also, we report the Rand index versus the crisp solution obtained with the fuzzy c-means clustering (FCMC), that is the observation is assigned to the cluster with highest degree of membership. These values support the conclusion that the cluster solutions we have obtained are similar to the ones obtained with other methods using the same features and number of clusters.

Of course, other meaningful cluster solutions can be found with different features and/or with different number of clusters. In this regard, we have opted to select, depending on the particular methodology, numbers of clusters that vary between 6 and 8. This is because we believe that this (or a comparable) range of values is particularly interesting in many smart grid applications. For instance, in the contest

[1] It should be noted that the 1-nn procedure looks for the nearest neighbor which would correspond to the single linkage in the hierarchical cluster and we are using the complete linkage.

	QAC solution	SAC solution	PAC solution
k-means	.791	.781	.745
PAM	.789	.754	.748
FCMC	.773	.773	.746

Table II
COMPARISON WITH OTHER CLUSTERING TECHNIQUES: RAND INDEX

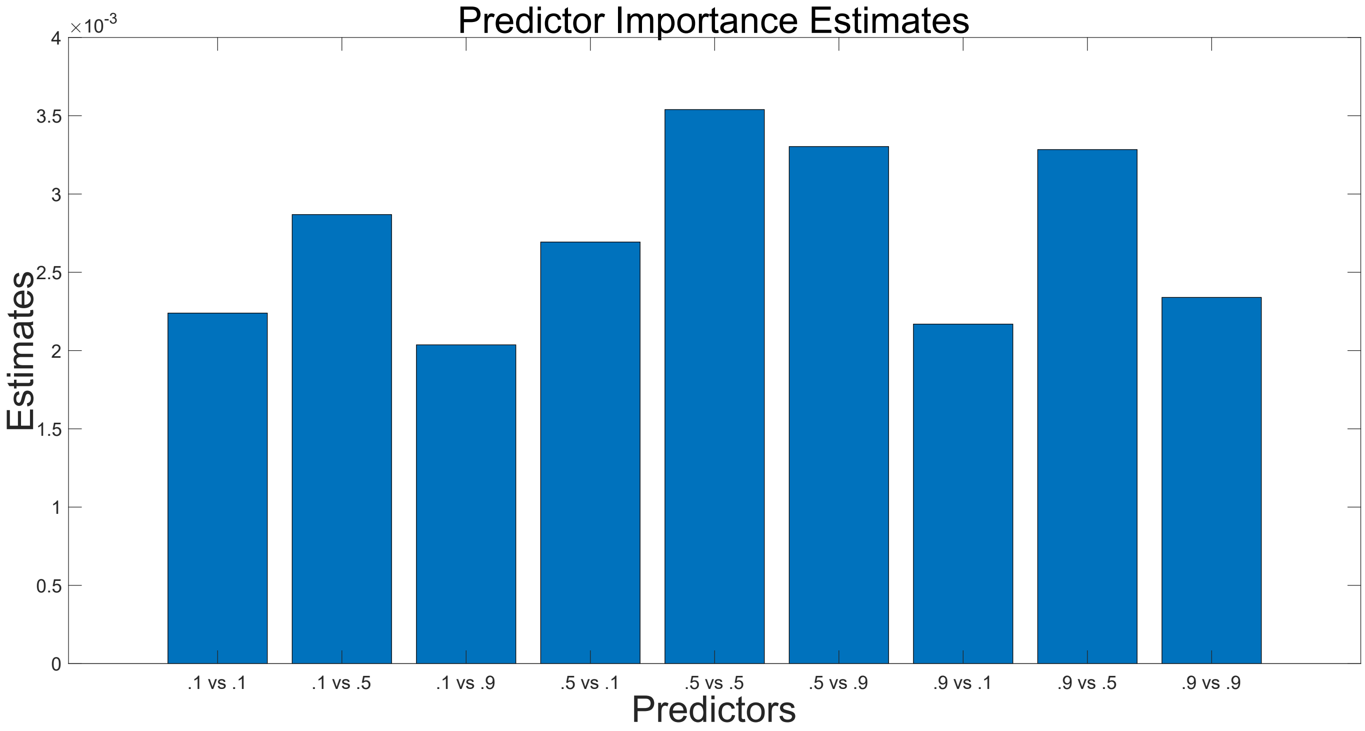

Figure 4. Predictor importance estimates for clusters based on QC.

of customer segmentation, energy retailers may use these clusters to improve their business decision making by offering personalized electricity price tariffs (see [5]). Hence, it is reasonable to assume that the total number of offered tariffs (one per cluster) will not be too small, to effectively discriminate among consumers, but not too large, to not over-complicate the business model.

Figures 4 - 6 provide the estimates of the predictor importance. It is clear that all features are relevant in the clustering based on QC but we can select the features in the clusters based on AC and PAC. In particular, for AC, the first fifteen lags and the four lags around the 48–lag appear to be relevant and, for PAC, the first four lags and the four lags before and two lags after the 48– and 96–lags as well as those daily "seasonal" lags. It is interesting to notice that the 48–lag is not relevant in the AC but this is due to the (daily) seasonal difference. However, there are still stationary seasonal behavior as reflected by relevant predictors/lags around the 48–lag. For PAC, the daily lags are highly relevant. The misclassification rates estimated by cross–validation for the three trained decision trees were 9.9%, 23.3% and 21.4% when using QC, AC and PAC, respectively. These low rates indicate that the errors are far from those obtained from a random classification procedure. This points out that the obtained trees are good approximations to the clustering mechanism. Of course, other supervised classification procedures such as random forest or neural networks can be used in order to obtain better approximations. However, this is not the purpose of this work as we rather aim to obtain an assessment of the importance of the features that have been used to obtain the unsupervised classification.

Tables III - V show the number of households on each cluster that are classified in the three ACORN_ GROUPED categories (with three possible values, adversity, comfortable and affluent). Note that they are

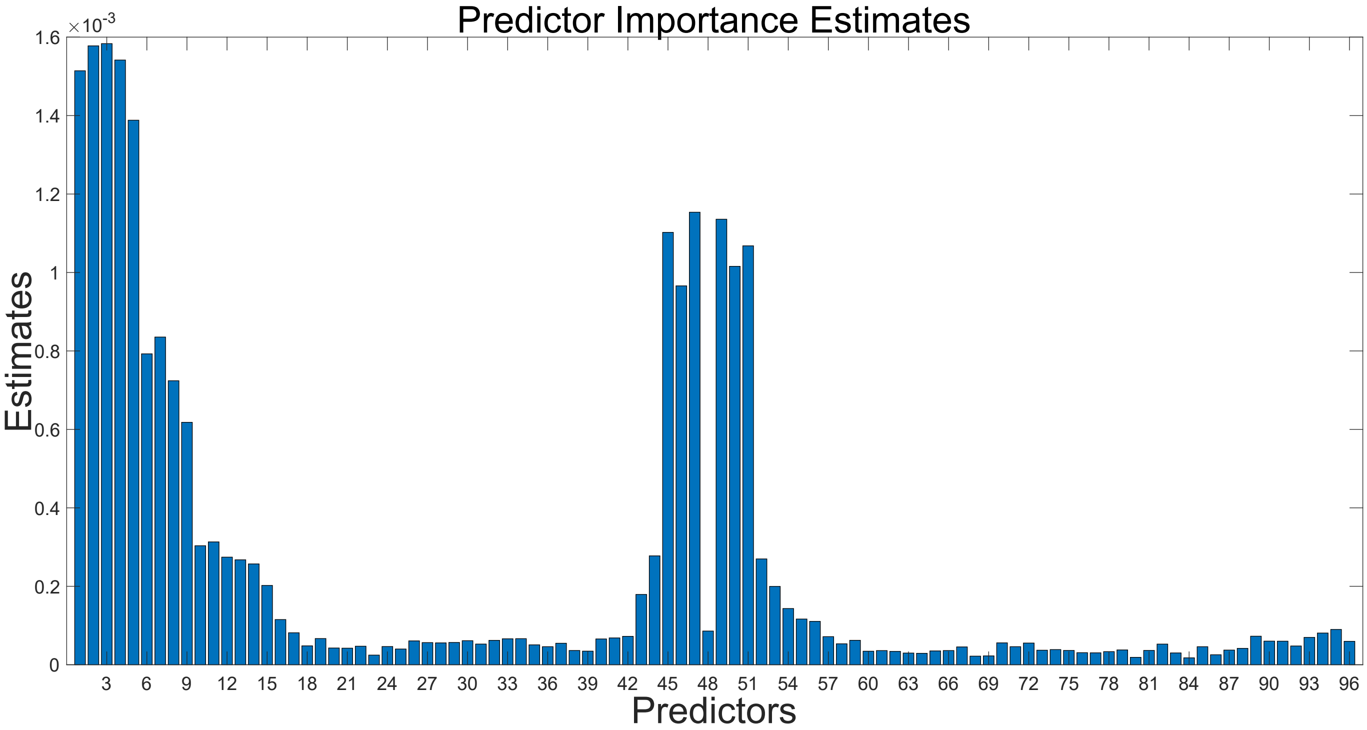

Figure 5. Predictor importance estimates for clusters based on AC.

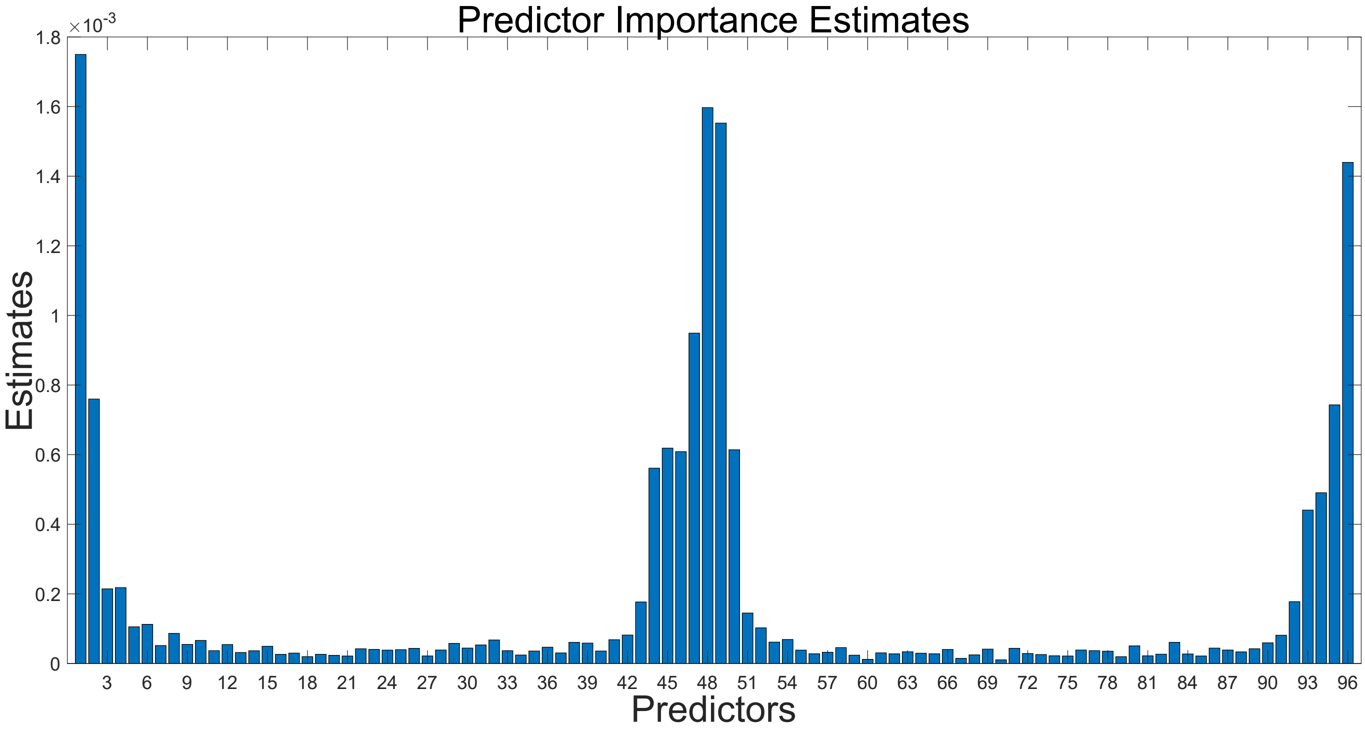

Figure 6. Predictor importance estimates for clusters based on PAC.

Cluster	Adversity	Comfortable	Affluent
c^1_{QC}	24	24	44
c^2_{QC}	293	278	360
c^3_{QC}	42	23	36
c^4_{QC}	52	40	37
c^5_{QC}	28	22	55
c^6_{QC}	482	343	358
c^7_{QC}	18	9	16
c^8_{QC}	146	160	258

Table III
CLUSTERS BY QC VERSUS ACORN_GROUPED

Cluster	Adversity	Comfortable	Affluent
c^1_{AC}	70	83	161
c^2_{AC}	426	404	544
c^3_{AC}	252	171	179
c^4_{AC}	16	23	59
c^5_{AC}	120	79	99
c^6_{AC}	214	140	120

Table IV
CLUSTERS BY AC VERSUS ACORN_GROUPED

unevenly distributed across clusters. Indeed, we have performed chi-squared tests in those tables and the results are highly significant in the three cases revealing that clustering is related to ACORN_GROUPED classification. We should point out that we do not intent to validate our cluster solutions with this analysis, but rather show that the cluster solutions and the demographic category are statistically related (using a chi-square test). In other words, the proposed clustering methodology is able to, up to some representative extend, provide insights on the geo-demographic characteristics of a household (Acorn groups), just by studying the time series dependencies.

Figures 7 - 9 show the prototype's half-hourly profile for each cluster. For clarity we represent only the medoid of each cluster, that is, the half-hourly profile for the element (smart meter) in the cluster

Cluster	Adversity	Comfortable	Affluent
c^1_{PAC}	27	16	61
c^2_{PAC}	66	67	111
c^3_{PAC}	397	376	501
c^4_{PAC}	454	305	261
c^5_{PAC}	39	51	129
c^6_{PAC}	83	54	68
c^7_{PAC}	18	29	39

Table V
CLUSTERS BY PAC VERSUS ACORN_GROUPED

with minimal average dissimilarity to all objects in that cluster. We can observe different characteristic consumption patterns associated to different types of consumers. For instance, the eight clusters obtained with QC and complete linkage in Figure 7 allow distinguishing between consumers with morning (clusters c^3_{QC}, c^6_{QC} and c^7_{QC}) and evening (clusters c^1_{QC}, c^2_{QC}, c^5_{QC} and c^8_{QC}) peak loads, and those with a more constant consumption pattern (cluster c^4_{QC}). This is also appreciated in the 6 clusters obtained with AC and complete linkage in Figure 8. In this case, clusters c^1_{AC} and c^4_{AC} capture those consumers with two intermediate peak loads in the morning and in the evening. Cluster c^5_{AC} represents consumers with a single peak consumption in the afternoon, and clusters c^2_{AC}, c^3_{AC} and c^5_{AC}, present consumers with less volatility.

Similarly, Figure 9 shows the clusters obtained with PAC and complete linkage. Clusters c^1_{PAC}, c^2_{PAC} and c^5_{PAC} characterize consumers with a steady increasing load that reach its maximum at midnight, while clusters c^3_{PAC}, c^4_{PAC}, c^6_{PAC} and c^7_{PAC}, represent consumers with two intermediate peaks in the morning and evening. In this case clusters from each type are mainly differentiated by the average load consumption levels.

We can also compare the main differences between the clusters obtained by QC, AC and PAC, in terms of the profiles of their bigger clusters (those that include more than 1000 smart meters). These are cluster 6 (c^6_{QC}) for QC (1183 smart meters), cluster 2 (c^2_{AC}) for AC (1374 smart meters), and clusters 3 (c^3_{PAC}) and 4 (c^4_{PAC}) for PAC (1274 and 1020 smart meters, respectively). In particular c^6_{QC} (Figure 7) includes consumers with a potential high load consumption during the mornings, and a second softer (and less volatile) peak consumption during the night. On the contrary, c^2_{AC} (Figure 8) and c^3_{PAC} (Figure 9) characterize consumers with low consumption during the morning and a single peak during the night, while c^4_{PAC} includes those with a more stable load profile.

IV. Conclusions

In this work we have presented three different hierarchical-based clustering strategies based on a set "dissimilarity" measures computed over: quantile auto-covariances, and simple and partial autocorrelations. The main advantage of this approach is that we can summarize each series in only a set of representative features which makes them very easy to implement (highly efficient), easy to automatize and scalable to hundreds of thousands of series, i.e., valid for real-world applications with large datasets of time series, as the ones obtained from smart meters. We evaluate the performance of these clustering models with thousands of electricity consumption time series. The results are promising: we are able to obtain highly representative clusters capturing different electricity load consumption patterns and identifying the level of influence of each of the models' features. Moreover, we have seen how the proposed clustering scheme can provide meaningful insights on the geo-demographic level of a household (Acorn groups), just by analyzing its time series dependencies (autocorrelations). Future lines of work would explore if these clustering strategies could be useful in the design of forecasting procedures for aggregated and disaggregated smart meters time series.

Acknowledgment

The authors gratefully acknowledge the financial support from the Spanish government through project MTM2017-88979-P and from Fundación Iberdrola through "Ayudas a la Investigación en Energía y Medio Ambiente 2018".

References

[1] R. Hierzinger, M. Albu, H. Van Elburg, A. J. Scott, A. Łazicki, L. Penttinen, F. Puente, and H. Sæle, "European smart metering landscape report 2012," *SmartRegions Deliverable*, vol. 2, 2012.

[2] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni, and N. Gudi, "Smart meters for power grid—challenges, issues, advantages and status," in *2011 IEEE/PES Power Systems Conference and Exposition*. IEEE, 2011, pp. 1–7.

[3] B. Yildiz, J. Bilbao, J. Dore, and A. Sproul, "Recent advances in the analysis of residential electricity consumption and applications of smart meter data," *Applied Energy*, vol. 208, pp. 402–427, 2017.

[4] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2018.

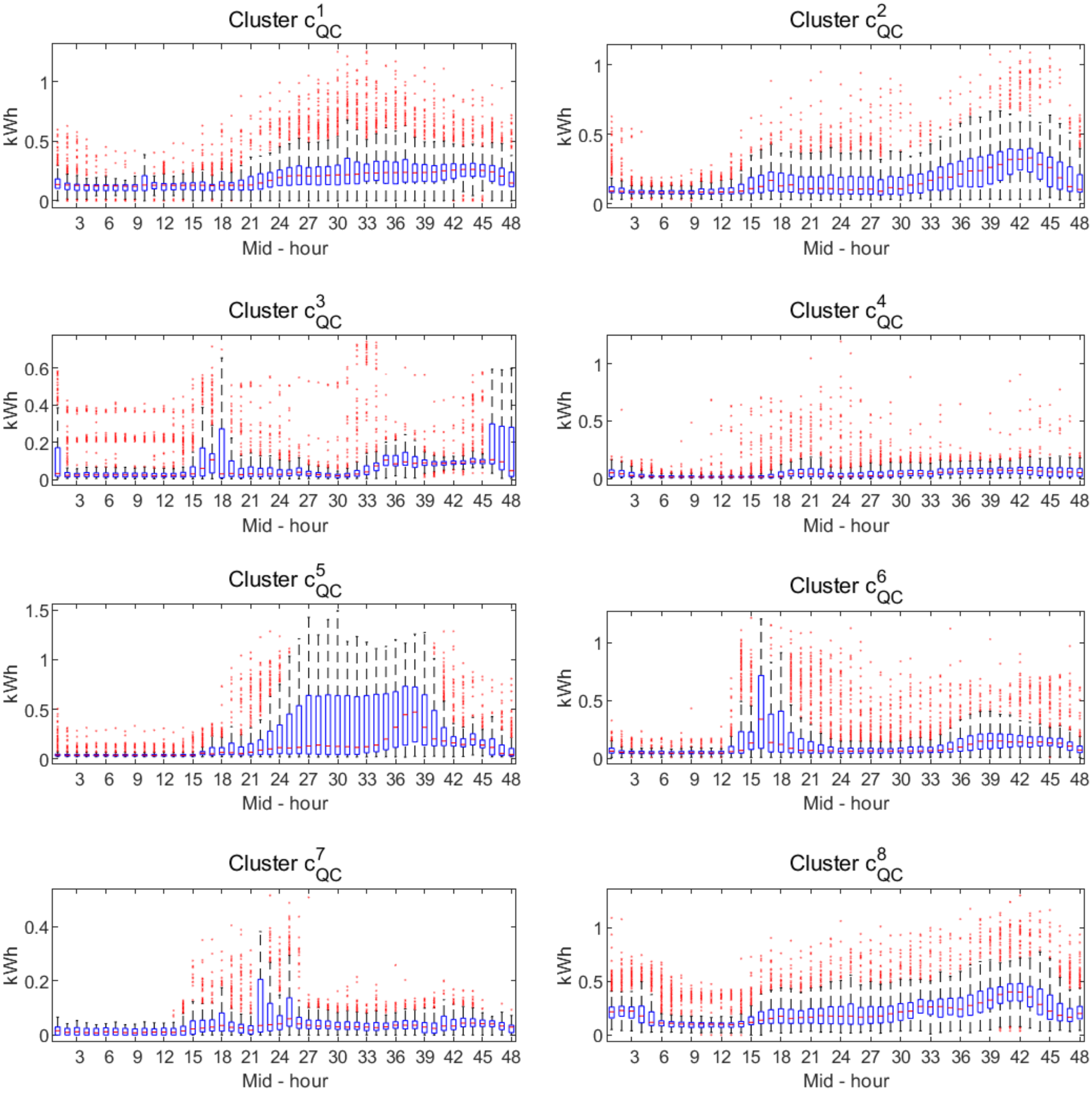

Figure 7. Prototype's half-hourly profile for clusters obtained with QC and complete linkage.

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

[6] J. L. Mathieu, D. S. Callaway, and S. Kiliccote, "Variability in automated responses of commercial buildings and industrial facilities to dynamic electricity prices," *Energy and Buildings*, vol. 43, no. 12, pp. 3322–3330, 2011.

[7] A. M. Alonso, P. Galeano, and D. Peña, "A robust procedure to build dynamic factor models with cluster structure," *Journal of Econometrics*, 2020.

[8] J. Gama and P. P. Rodrigues, "Stream-based electricity load forecast," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2007, pp. 446–453.

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

[10] M. Koivisto, P. Heine, I. Mellin, and M. Lehtonen, "Clustering of connection points and load modeling in distribution systems," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1255–1265, 2012.

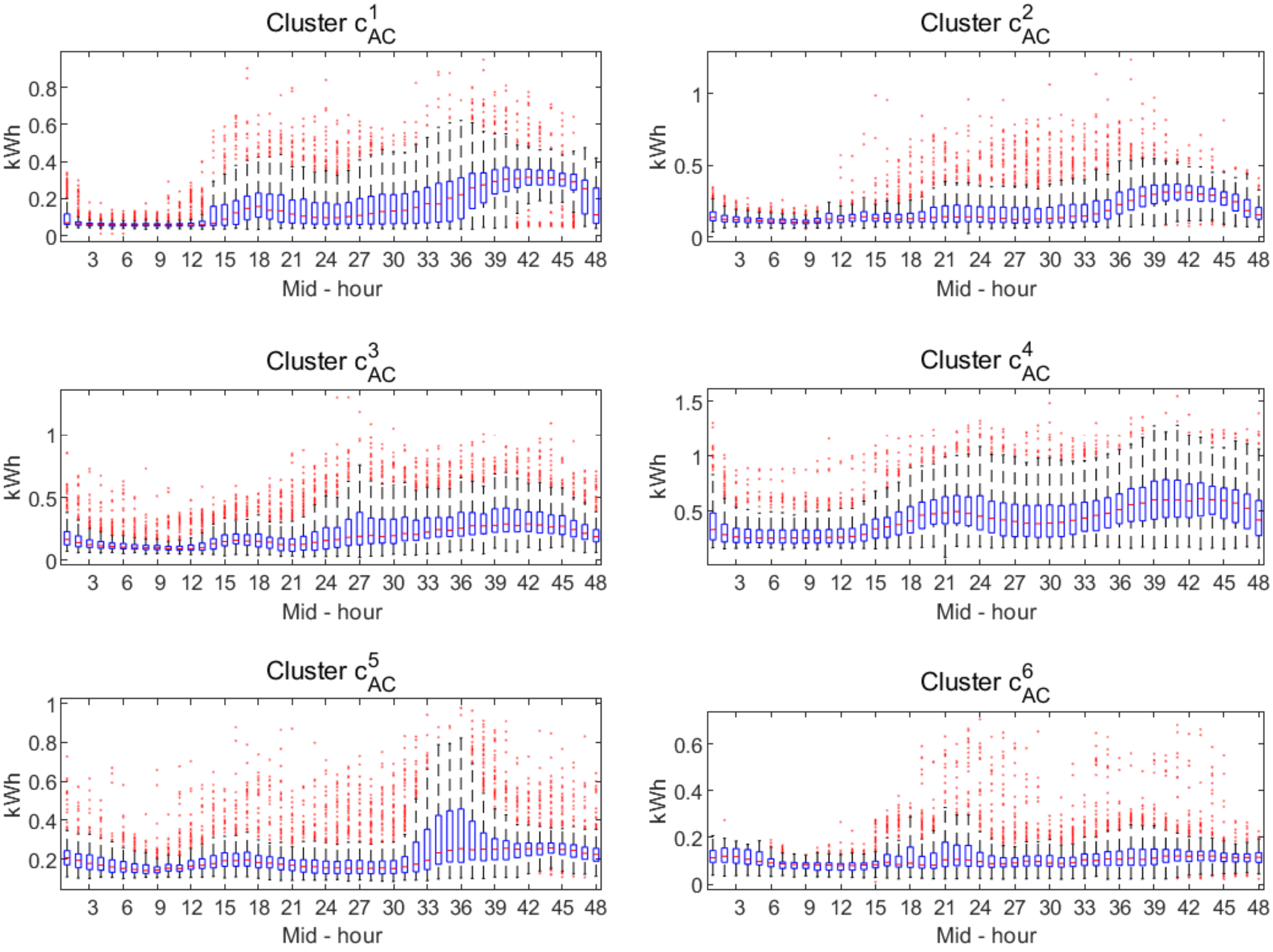

Figure 8. Prototype's half-hourly profile for clusters obtained with AC and complete linkage.

[11] 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.

[12] A. Lavin and D. Klabjan, "Clustering time-series energy data from smart meters," *Energy efficiency*, vol. 8, no. 4, pp. 681–689, 2015.

[13] 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.

[14] A. Tureczek, P. Nielsen, and H. Madsen, "Electricity consumption clustering using smart meter data," *Energies*, vol. 11, no. 4, p. 859, 2018.

[15] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE transactions on smart grid*, vol. 7, no. 5, pp. 2437–2447, 2016.

[16] A. Kavousian, R. Rajagopal, and M. Fischer, "Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior," *Energy*, vol. 55, pp. 184–194, 2013.

[17] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Järventausta, "Enhanced load profiling for residential network customers," *IEEE Transactions on Power Delivery*, vol. 29, no. 1, pp. 88–96, 2013.

[18] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397–410, 2014.

[19] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE transactions on smart grid*, vol. 7, no. 1, pp. 136–144, 2015.

[20] J. P. Gouveia and J. Seixas, "Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys," *Energy and Buildings*, vol. 116, pp. 666–676, 2016.

[21] S. Zhong and K.-S. Tam, "Hierarchical classification of load profiles based on their characteristic attributes in frequency domain," *IEEE Transactions on Power Systems*, vol. 30, no. 5, pp. 2434–2441, 2014.

[22] R. Granell, C. J. Axon, and D. C. Wallom, "Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles," *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3217–3224, 2014.

[23] R. Al-Otaibi, N. Jin, T. Wilcox, and P. Flach, "Feature construction and calibration for clustering daily load curves from smart-meter data," *IEEE Transactions on industrial informatics*, vol. 12, no. 2, pp. 645–654, 2016.

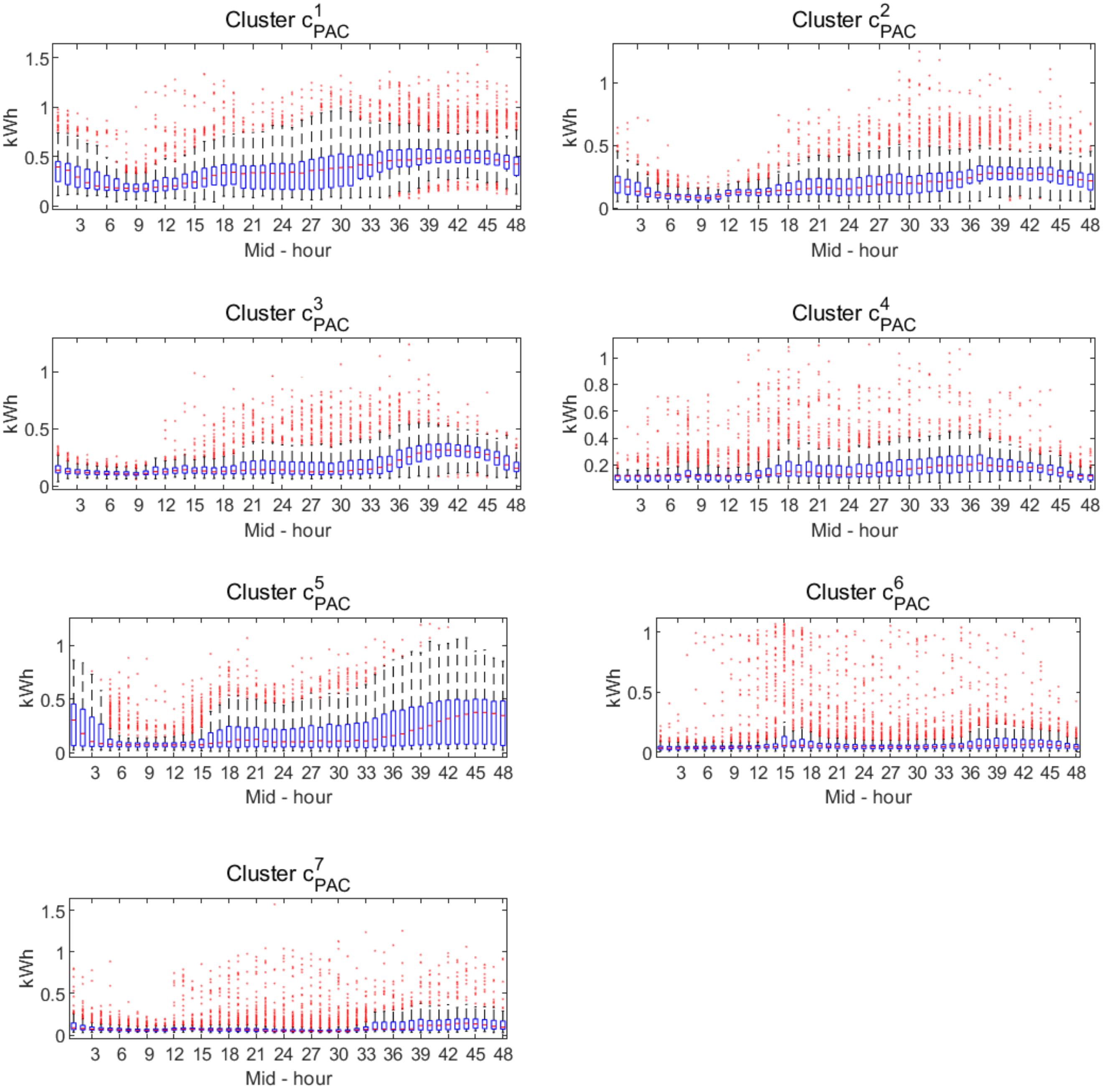

Figure 9. Prototype's half-hourly profile for clusters obtained with PAC coefficients and complete linkage.

[24] O. Y. Al-Jarrah, Y. Al-Hammadi, P. D. Yoo, and S. Muhaidat, "Multi-layered clustering for power consumption profiling in smart grids," *IEEE Access*, vol. 5, pp. 18 459–18 468, 2017.

[25] M. Chaouch, "Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 411–419, 2013.

[26] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 911–918, 2014.

[27] A. Al-Wakeel, J. Wu, and N. Jenkins, "K-means based load estimation of domestic smart meter measurements," *Applied energy*, vol. 194, pp. 333–342, 2017.

[28] K. Bandara, C. Bergmeir, and S. Smyl, "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," *arXiv preprint arXiv:1710.03222*, 2017.

[29] B. Yildiz, J. I. Bilbao, J. Dore, and A. Sproul, "Household electricity load forecasting using historical smart meter data with clustering and classification techniques," in *2018 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)*. IEEE, 2018, pp. 873–879.

[30] B. Lafuente-Rego and J. A. Vilar, “Clustering of time series using quantile autocovariances,” *Advances in Data Analysis and classification*, vol. 10, no. 3, pp. 391–415, 2016.

[31] P. J. Brockwell, R. A. Davis, and S. E. Fienberg, *Time series: theory and methods: theory and methods*. Springer Science & Business Media, 1991.

[32] O. Linton and Y.-J. Whang, “The quantilogram: With an application to evaluating directional predictability,” *Journal of Econometrics*, vol. 141, no. 1, pp. 250–282, 2007.

[33] T.-H. Li, “Quantile periodograms,” *Journal of the American Statistical Association*, vol. 107, no. 498, pp. 765–776, 2012.

[34] A. M. Alonso and D. Peña, “Clustering time series by linear dependency,” *Statistics and Computing*, vol. 29, no. 4, pp. 655–676, 2019.

[35] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[36] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*, ser. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[37] H. Ishwaran, “Variable importance in binary regression trees and forests,” *Electronic Journal of Statistics*, vol. 1, pp. 519–537, 2007.

[38] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, “Understanding variable importance in forests of randomized trees,” in *Proceedings of the 26th International Conference on Neural Information Processing Systems*. NIPS, 2013, pp. 431–439.

[39] [Online]. Available: https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

[40] [Online]. Available: https://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf

[41] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *Journal of intelligent information systems*, vol. 17, no. 2-3, pp. 107–145, 2001.

[42] E. Keogh and S. Kasetty, “On the need for time series data mining benchmarks: a survey and empirical demonstration,” *Data Mining and knowledge discovery*, vol. 7, no. 4, pp. 349–371, 2003.