

Enhancing household-level load forecasts using daily load profile clustering

Edward Barbour

Civil and Environmental Engineering, Massachusetts
Institute of Technology
Cambridge, MA
Energy Technologies Area, Lawrence Berkeley National
Laboratory
Berkeley, CA
ebarbour@mit.edu

Marta González*

Department of City and Regional Planning, UC Berkeley
and
Berkeley, CA
Energy Technologies Area, Lawrence Berkeley National
Laboratory
Berkeley, CA
martag@berkeley.edu

ABSTRACT

Forecasting the electricity demand for individual households is important for both consumers and utilities due to the increasing decentralized nature of the electricity system. Particularly, utilities often have very little information about their consumers except for aggregate building level loads, without knowledge of interior details about the household appliance sets or occupants. In this paper, we explore the possibility of enhancing the day-ahead load forecasts for hundreds of individual households by clustering their daily load profile history to obtain each consumer's specific typical consumption patterns. The clustering method is based on load profile shape using the Earth Mover's Distance metric to calculate similarity between load profiles. The forecasting methods then predict the next day shape from the empirical probability of previous cluster transitions in the consumer's load history and estimate the magnitude either by using historical load relationships with temperature and forecast temperatures or previous day consumption levels. The generated forecasts are compared to a benchmark Multiple Linear Regression (MLR) day-ahead forecast and persistence forecasts for all individuals. While at the aggregate level the MLR method represents a significant improvement over persistence forecasts, on an individual level we find that the best forecasting model is specific to the individual. In particular, we find that the MLR model produces lower errors when consumers have a consistent daily temperature response and the cluster model with previous day magnitude produces lower errors for consumers whose consumption changes abruptly in magnitude for several days at a time. Our work adds to the state of knowledge surrounding individual household load forecasting and demonstrates the potential for cluster-based methodologies to enhance short term load forecasts.

*Corresponding author.

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

BuildSys '18, November 7–8, 2018, Shenzhen, China

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5951-1/18/11...\$15.00

<https://doi.org/10.1145/3276774.3276793>

CCS CONCEPTS

• **Theory of computation** → **Unsupervised learning and clustering**; • **Applied computing** → **Forecasting**; • **Hardware** → **Smart grid**; Energy metering; *Smart grid*; • **Information systems** → *Clustering*;

KEYWORDS

Smart meter data, clustering analysis, load forecasting, regression analysis

ACM Reference Format:

Edward Barbour and Marta González. 2018. Enhancing household-level load forecasts using daily load profile clustering. In *The 5th ACM International Conference on Systems for Built Environments (BuildSys '18)*, November 7–8, 2018, Shenzhen, China. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3276774.3276793>

1 INTRODUCTION

Traditionally, energy systems have operated on an aggregated demand-led basis, wherein power is generated in centralized large scale power stations and transmitted along the wires of the electrical network to consumers. Individual households and other small consumers are located at the ends of the low-voltage distribution network and their electricity is supplied from the centralized power plants, with electricity flowing radially outwards along the distribution network. However, with the increase in popularity of Distributed Energy Resources (DERs) which can be embedded in the end nodes of the distribution network, this situation is changing [29, 31]. Consumers located at the ends of the network are increasingly supplying power into the network and modifying their behavior according to price signals in the electricity market. To take advantage of these changes, the future smart grid seeks to operate the electricity system in an efficient and highly decentralized manner, primarily to facilitate a sustainable energy system. Local zones of the grid must be independently operated and finite resources (including fuel or transmission line capacity) must be correctly allocated. From a utility perspective, a failure to adapt to these changes has the potential to cause both significant operational and economic disruption [22, 32]. Therefore, accurate load forecasts on a distributed scale are crucial for ensuring reliable system and market operation, as well as for longer term planning activities [9, 30].

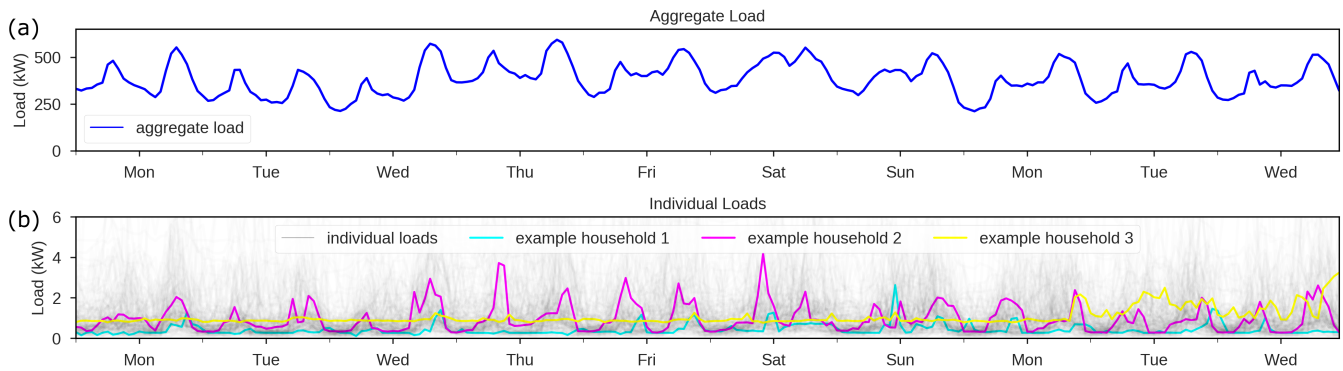


Figure 1: (a) Total Aggregate load for all households for 10 days starting on Monday 5th January for 10 days. (b) Example individual household loads over the same period.

Load forecasts have previously been studied on many different levels and timescales. Historically, utilities have focused on aggregated levels of hundreds of MW or more for both short and long term load forecasting. Short Term Load Forecasting (STLF) on this scale, which includes forecasting between 1 hour and 1 week ahead of real time, has been used to manage the unit commitment of the centralized utility generation fleets on a within-day level, while long term forecasts have helped with business and planning models [16]. Motivated by the increasing decentralization of grid operations, recent forecasting work has looked at more and more granular levels, including forecasting for zones [15], neighborhoods [17] and individual buildings, including households [30], commercial [6] and industrial buildings [36]. This work has been facilitated by the large (and growing) amount of (publicly) available smart meter data for individual buildings from many open data sources – for example [26], [2] and [18]. Largely in parallel with load forecasting, there has also been significant interest in smart meter data analytics [33]. This has included clustering daily electricity consumption patterns of vast amounts of individual consumers for many different applications, including identifying groups of consumers suitable for demand response [19] or energy efficiency measures [34], identifying AC setpoints [11] and identifying household features [1, 3] such as employment or physical characteristics like floor area and appliance types. However, the potential for combining load forecasting with load profile clustering has only recently been highlighted [28] and many aspects remain unexplored.

In this work, we propose a forecasting methodology based on clustering the daily load history from individual consumers, identifying typical daily load shapes specific to each consumer and their probability of occurrence on different days. We use these to estimate each consumer’s next day load. We compare the proposed model with a developed Multiple Linear Regression (MLR) model for forecasting day-ahead loads, which is successful at predicting the load on an aggregate level, and which is of a similar type to benchmark forecast models used by many utilities for zone level forecasting [14]. We use only building-level smart meter data, as this is much more widely available when compared with data sets relating to building interiors (*i.e.* sub-metered and appliance level usage data), and temperature data from a local weather station.

For 326 small electricity consumers, we find that our cluster-based models lead to a small reduction (around 2%) in the median MAPE (Mean Absolute Percentage Error) when compared with the developed MLR and a 24 hour persistence forecast for all consumers. In addition, the differences are much more significant for some individual consumers between the different forecasting models. However, no one model is able to generate the lowest MAPE for all consumers. Furthermore, in general the MAPEs for each model are highly correlated, suggesting that if a consumer is difficult to predict, all of the models struggle to accurately predict the load. We interpret our results to mean that cluster-enhanced forecasts may be useful for a large number of individual consumers, and while it is likely that the best forecasting models for many different households will include many different model types, cluster-based forecasting methods like the one presented could make a useful contribution.

2 RELATED WORK

The range of literature regarding load forecasting for individual buildings has employed a number of methodologies, including Multiple Linear Regression [5], Neural Networks [12, 30], Autoregressive Integrated Moving Average (ARIMA) [12, 30, 37] and Holt-Winters smoothing [37] amongst others. Many of these studies and others have exploited highly disaggregated data sources, including appliance level monitoring [12, 30, 38] and occupancy sensing to better forecast the building loads. While it has been shown that including these additional data sources relating to interior building activities can improve the forecasting accuracy [38], questions remain regarding whether this type of data could be widely available for entities seeking to forecast individual loads for many different individual buildings (*i.e.* grid controllers of utilities). Therefore, in this work we only use datasets which are likely to be available for this purpose in the near future, including building level hourly data and weather data.

Regarding cluster analyses, several methods have been proposed to group consumers based on their smart meter data [8], which include k -means [4, 19], agglomerative hierarchical clustering [19],

finite mixture models [13] and self organizing maps [3]. Additionally, different distance metrics have been used to specify the similarity between load profiles. While the euclidean distance between load profiles has been most commonly used, more recent work has employed shape based distance metrics for clustering load profiles, including Earth Mover's Distance [20], Dynamic Time Warping (DTW) distance [28, 35] and k -shape distance [35]. The advantage of these shape-based metrics is that the different hours of the day are not treated as orthogonal by the similarity measure. This is appropriate given that similar behaviors may involve an appliance usage at slightly different times, and these load profiles will be more likely to be grouped together under a shape-based metric than with the euclidean distance. Furthermore, it has been shown that shape-based metrics result in a lower number of clusters representing typical behaviors [28].

Only a handful of studies have combined cluster analysis and load forecasting. [25] demonstrates a small improvement in forecasts of the aggregate load of 5000 consumers by enhancing neural networks with clusters corresponding to common consumer usage patterns. [28] developed a DTW cluster based forecasting model which reduced the DTW error of the predicted loads using 22 days of data for 1057 households. Other work has linked forecasting with alternative unsupervised learning methods, such as in [37], where features identified by sparse coding were shown to improve load forecasts. Our work uses shape based clustering to identify the typical daily consumption patterns of individual consumers, using these and temperature data to forecast next day loads. We compare the forecasted loads to a day-ahead MLR forecasting model and, as well as illustrating the general error trends, examine how the models perform for each individual consumer.

3 DATA SETS

The demand data in our study is from the Pecan Street project, which provides 15-minute resolution electricity data for in excess of 1000 consumers on a voluntary basis [26] and which is freely available for academic purposes. The consumers are spread across many different locations, although they are most concentrated in Austin Texas, which is the site of the original project. We downloaded all the available hourly electrical usage data for Austin households in the period 1st January 2015 to 31st October 2015, and after filtering for missing and erroneous data we select 326 consumers with a sufficiently complete set of consumption data. We split the data into a training set and a test set, wherein the training set spans the period 1st January 2015 to 31st August 2015 and the testing set spans the rest of the data, *i.e.* 1st September 2015 to 31st October 2015. Figure 1 shows the aggregate load from all the consumers as well as some of the individual consumer demands.

We also use publicly available temperature data from a local weather station (the *TX_Austin_33_NW* station), downloaded from the National Oceanic and Atmospheric Administration (NOAA) website [10]. This weather station is situated on the retired Austin Municipal airport site on which the Pecan Street houses have been developed, and hence the distance between the weather station and the households is minimal. The temperature data had 61 hours in the year 2015 where data was missing. To fill in these periods (which were up to a maximum of 10 hours in length) we took the values

for temperature for the 10 hours preceding the missing values and the 10 hours immediately after and used cubic interpolation to fill in all the missing values.

4 AGGREGATE LOAD FORECASTING

We first develop a MLR model for predicting the aggregated load (the total demand for all the households) and compare it to a naive persistence model. When developing MLR models for STLF, it is often useful to do some exploratory data analysis to determine which variables and interactions should be utilized. These variables include quantitative variables, such as the temperature or previous load at a particular time lag, and qualitative variables such as day-of-the-week or period-of-the-day.

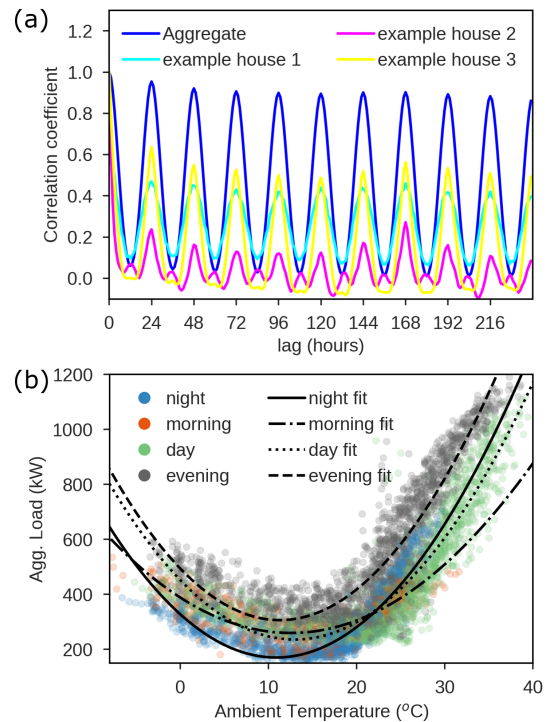


Figure 2: Exploring predictors of electricity loads. (a) Auto-correlation of the aggregate load as well as autocorrelations for three example households. (b) Scatter plot of Temperature versus Load. Data from different times of the day has been highlighted and fitted with distinct quadratic polynomials.

Figure 2a shows the autocorrelation function of the aggregate load (as well as the autocorrelation functions of three example individual households). It can be seen that the function is highly periodic, with the strongest periodicity occurring on a daily basis (24 hour lag). Following this the autocorrelation declines further from the period of interest, although it increases slightly at a lag of 168 hours (weekly lag). Therefore, we consider that in our model it is beneficial to include the value 24 hours previous as well as 168 hours previous as possible predictor variables. This fits with our

intuition, whereby we may expect that for weekdays Tuesday to Friday, the load 24 hours previous could be a good predictor of the load at time t , whereas for Mondays and weekend days the load one week previous may be a better predictor.

4.1 Persistence forecasting model

The strong autocorrelation with the value 24 hours previous indicates that a good choice of a persistence model is to assume the value of the load at time t will equal the value of the load 24 hours previously, as shown in Equation 1.

$$\hat{l}_t = l_{t-24} \quad (1)$$

Here, \hat{l}_t is the predicted load at time t and l_{t-24} is the actual load 24 hours before t .

4.2 MLR model for aggregate load

In general, it is very important to consider the effects of temperature when predicting electricity loads, as significant proportions of domestic electricity use goes towards heating and cooling. Therefore, we also include temperature as a predictor variable in our MLR model. The effect of temperature on load has been modeled as a piecewise linear relationship [24], as well as 2nd [7] and 3rd [14] ordered polynomials. Based on observations from Figure 2b we opt for a 2nd order polynomial. Additionally, the effect of temperature is known to be different at different times of the day, as people can respond differently to temperature depending on their activities, which vary during different times of the day [14]. Upon inspection of the scatter plot of temperature against load shown in Figure 2b, we observe that this indeed appears to be the case, and the plot indicates the presence of interaction effects between temperature and time of day. For example, there is a clear trend that the load is lower during the morning hours for the same temperature than that during the late evening hours. This could be due to the fact that in many of these households the occupancy in the morning is typically lower and, as a result, the occupants may have higher thermostat settings during these hours, for example, to save on electric bills or try and reduce their energy footprint. There may also be different interaction effects at different times of the year, however since we are predicting loads for months for which we have no prior data we do not include these interaction effects.

Including the interaction effects of day-of-the-week and time-of-day, the MLR load prediction model is expressed as:

$$\hat{l}_t = \sum_{i=1}^2 DoW_{t,i}(\beta_{0,i} + \beta_{1,i}l_{t-24} + \beta_{2,i}l_{t-168}) + \sum_{i=1}^4 HoD_{t,i}(\beta_{3,i} + \beta_{4,i}T_t + \beta_{5,i}T_t^2) \quad (2)$$

In Equation 2, the β variables are the regression coefficients. $DoW_{t,i}$ and $HoD_{t,i}$ are qualitative predictor variables corresponding to day of the week and time of day. They are defined as follows:

$$[DoW_{t,1}, DoW_{t,2}] = \begin{cases} [1, 0] & \text{IF day is Tues, Wed, Thurs, Fri} \\ [0, 1] & \text{otherwise} \end{cases} \quad (3)$$

$$[HoD_{t,1}, HoD_{t,2}, HoD_{t,3}, HoD_{t,4}] = \begin{cases} [1, 0, 0, 0] & \text{IF } t \in \{23, 0, 1, 2, 3, 4\} \\ [0, 1, 0, 0] & \text{IF } t \in \{5, 6, 7, 8\} \\ [0, 0, 1, 0] & \text{IF } t \in \{9, 10, 11, 12, 13, 14, 15\} \\ [0, 0, 0, 1] & \text{IF } t \in \{16, 17, 18, 19, 20, 21, 22\} \end{cases} \quad (4)$$

For the aggregate load forecasting models, the MAPEs are shown in Table 1. MAPE is defined as:

$$MAPE = \frac{100}{n} \sum_{t=1}^{t=n} \frac{\hat{l}_t - l_t}{l_t} \quad (5)$$

It is clear that the MLR model represents a significant improvement over the persistence forecast. However, as may be expected after studying Figure 3a, the persistence based forecast has relatively low error to begin with, which is due to the high degree of autocorrelation at a 24 hour lag. Figure 3a shows the evolution of the aggregate load over the prediction period, illustrating why the persistence model is so effective.

Table 1: MAPEs for aggregate load forecasting models

Model	Training Period	Testing Period
Persistence	11.10	10.71
MLR	8.53	9.32

5 INDIVIDUAL LOAD FORECASTING

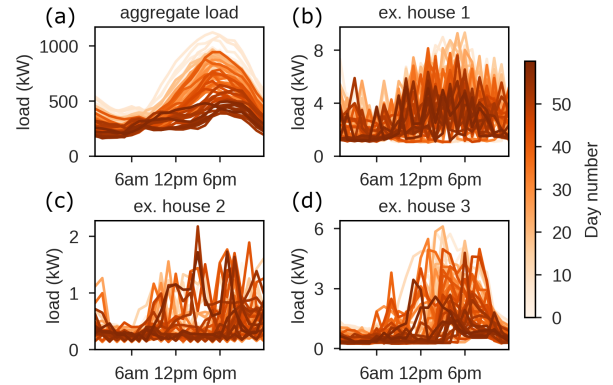


Figure 3: Load evolution. (a) The evolution of the aggregate load over the prediction period. Load evolution is also shown for (b) example household 1, (c) example household 2 and, (d) example household 3.

Figure 3 also shows the load evolution over the same period for three example households. While the aggregate load seems highly predictable, the load for individual households has a much less clear pattern. Nevertheless, as a first attempt at modeling the load for all the individual households, we employ the same persistence and

MLR models as have been used for the aggregate load. Therefore, for each household we develop a MLR model with the same features as illustrated in Equation 2. The training and testing periods are the same for as for the aggregate load case.

5.1 Clustering methodology

We now develop an alternate load forecasting model for each of the individual consumers by employing a clustering based approach. As previous work has demonstrated, different consumers have different typical lifestyles and behavior patterns which may be identified by segmenting daily load profiles of electric usage [19]. For the reasons outlined in the Related Work section, we opt for a shape-based distance metric over euclidean distance for specifying the similarity between load profiles, settling on the Earth Movers Distance (EMD) [27]. EMD essentially gives a measure of the amount of work that must be done to change one distribution into another, by specifying a ground distance between the distribution bins and expressing work as the quantity moved from each bin to each other bin integrated over the ground distance. Compared to euclidean distance, the EMD is more computationally expensive to calculate. Therefore, in [20], k -means and euclidean distance was first used to create a manageable initial sized dictionary of loads shapes, and then the dictionary entries were clustered with EMD, which allows for a ground-distance to be specified between the different hours of the day.

While previous work has focused on grouping the load profiles of multiple households together to identify typical behaviors of *consumer groups*, here we group the load profiles for each *individual consumer* separately, since our aim is to obtain representative typical behaviors specific to each individual consumer. This also has the advantage that it comprises a relatively small number of load shapes – we only have 243 days of training data for each consumer. Therefore, even though the EMD is computationally expensive to calculate, we simply employ a hierarchical agglomerative clustering with EMD as the distance metric directly on each household's normalized daily load profiles. The normalization of each daily load profile is expressed as follows:

$$e_{j,D,t} = \frac{l_{j,D,t}}{T_D}, \quad T_D = \sum_{t=1}^{24} l_{j,D,t} \quad (6)$$

where $l_{j,D,t}$ is the j^{th} consumer's load at time t during day D , and T_D is their total load for that day.

The EMD between two different normalized load profiles is then given by Equation 7:

$$d^{EMD}(e_{j,D_1}, e_{j,D_2}) = \frac{\min \sum_{t_1=1}^{24} \sum_{t_2=1}^{24} f_{t_1,t_2} d_{t_1,t_2}^G}{\sum_{t_1=1}^{24} \sum_{t_2=1}^{24} f_{t_1,t_2}} \quad (7)$$

Here, d^{EMD} is the EMD between profiles e_{j,D_1} and e_{j,D_2} which are both normalized daily load profiles of consumer j (on the days D_1 and D_2 respectively). f_{t_1,t_2} is the flow from load profile e_{j,D_1} at time t_1 to load profile e_{j,D_2} at time t_2 , and d_{t_1,t_2}^G is the ground distance between hours t_1 and t_2 . The minimum flow problem is then calculated subject to the following constraints:

$$f_{t_1,t_2} \geq 0 \quad (8)$$

$$\sum_{t_2=1}^{24} f_{t_1,t_2} \leq e_{j,D_1,t_1} \quad (9)$$

$$\sum_{t_1=1}^{24} f_{t_1,t_2} \leq e_{j,D_2,t_2} \quad (10)$$

Equation 8 ensures that the flows are always positive and Equations 9 and 10 ensure that the flows can never move more load than is available each time period. We solve for the EMD using an algorithm proposed in [23].

To perform the hierarchical clustering, we create a distance matrix for each consumer, M_j^{EMD} , which contains the pairwise EMDs between each normalized daily load profile for an individual consumer.

$$M_j^{EMD}(D_1, D_2) = d^{EMD}(e_{j,D_1}, e_{j,D_2}) \quad (11)$$

The distance matrix is symmetric. Finally we then use agglomerative hierarchical clustering with complete linkage as the clustering algorithm, using the distances specified in M_j^{EMD} . We cut the dendrogram at a maximum $EMD = 2$, which results in a maximum distance between two load profiles in the same cluster corresponding to translating a load profile by ± 2 hours.

5.2 Forecasting next day load using clusters

The aforementioned clustering process creates a number of different clusters for each individual consumer. Each consumer j then has a set of K_j different clusters, $\{C_1, C_2, \dots, C_{K_j}\}$ with the corresponding set of centroids $\{\zeta_1, \zeta_2, \dots, \zeta_{K_j}\}$, and the number K_j depends on the similarity between the household's daily load profiles. Figure 4 step 3 shows the set of clusters created for an example individual user. We find that certain clusters seem more likely to occur on different days of the week, in particular for many users certain clusters occur much more frequently on weekends than on weekdays or visa versa.

We consider that each cluster for each individual consumer corresponds to a potential state of consumption for that individual. The previous load data for each household can then be characterized as a sequence of cluster states, $e_j: (C_{D_1}, C_{D_2}, \dots, C_{D_N})$, corresponding to the normalized daily load profiles on days (D_1, D_2, \dots, D_N) and where $C_D \in \{C_1, C_2, \dots, C_{K_j}\}$.

Empirically, we then study the transition probability between different states $P(C_y|C_x)$. We construct separate transition probabilities corresponding to weekdays and weekends, given that our clustering procedure has identified different typical shapes for weekends and weekdays for many consumers. For both weekdays and weekends, we construct a transition matrix for each consumer M_j^{TRANS} which gives the probability of transitioning from one cluster on a given day to each other cluster on the next consecutive day of the same type (separating weekends and weekdays). The elements of $M_j^{TRANS}(C_x, C_y)$ are determined empirically by the fractional count of transitions that we observe in the consumer j 's load history from cluster C_x to cluster C_y .

To forecast the load profile for the prediction day, D_{N+1} , we then choose the cluster with the highest probability given the previous

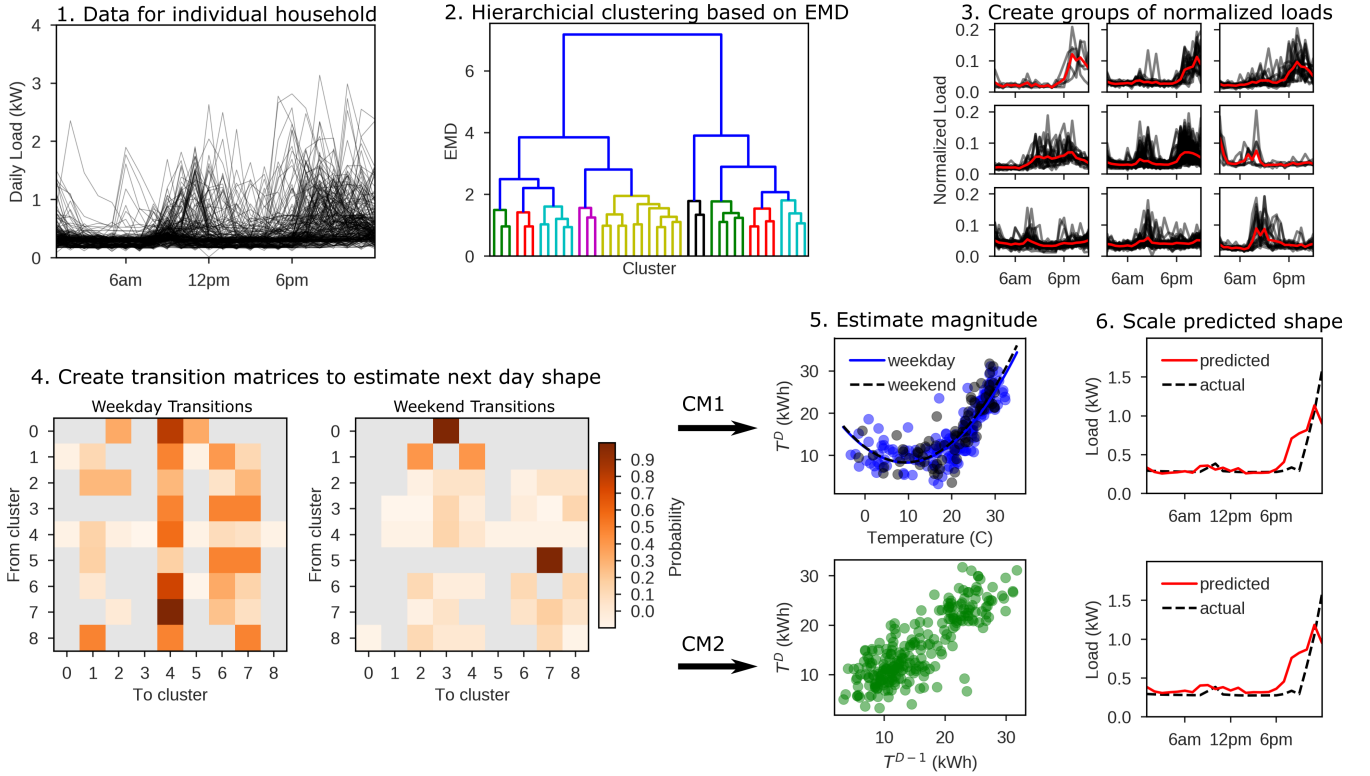


Figure 4: Illustrating the clustering and load forecasting processes. For each consumer the raw data is clustered using an agglomerative hierarchical process. Transition matrices are empirically created containing the probabilities of a particular next day load shape for weekdays and weekends respectively, and the load *shape* of the next day is predicted. CM1 scales the predicted shape using the historical relationship between the consumer’s daily load magnitude and temperature and CM2 scales the load shape according to the previous day load.

day cluster. We denote the chosen cluster as C^* . The normalized next day load is predicted as the centroid of C^* , ζ_{C^*} . Once the chosen cluster has been selected we update the transition probability matrices and the cluster centroids accordingly.

Finally, after predicting the load *shape* for the next 24 hours, we must predict the *magnitude* to scale the shape accordingly. Here we split the model into two, by considering two options for predicting the magnitude of the load shape.

5.2.1 Cluster Model 1 (CM1). First, we fit a 2^{nd} ordered polynomial to describe the relationship between temperature and total daily consumption for all previous days for each consumer. We then use the temperature for the prediction day D_{N+1} to predict the consumption magnitude, having separate models for weekdays and weekends. We denote this model “cluster model 1” or CM1 and it is described in the Equations 12, 13 and 14. The load magnitude for the predicted day is described by Equation 12, where $\gamma_{0,i,j}$, $\gamma_{1,j,i}$ and $\gamma_{2,j,i}$ are the coefficients of the polynomial fit between temperature and consumer j ’s historical daily load magnitudes. The index i iterates the variable $DoW_{D_{N+1},i}$ which is used to differentiate the relationship for weekdays and weekends.

$$|\hat{j}_{j,D_{N+1}}^{CM1}| = \sum_{i=1}^2 DoW_{D_{N+1},i} (\gamma_{0,j,i} + \gamma_{1,j,i} T_{D_{N+1}} + \gamma_{2,j,i} T_{D_{N+1}}^2) \quad (12)$$

$$[DoW_{D_{N+1},1}, DoW_{D_{N+1},2}] = \begin{cases} [1, 0] & \text{IF } D_{N+1} \text{ is weekday} \\ [0, 1] & \text{otherwise} \end{cases} \quad (13)$$

The final predicted load for the next day D_{N+1} in this model is then expressed as:

$$\hat{j}_{j,D_{N+1}}^{CM1} = |\hat{j}_{j,D_{N+1}}^{CM1}| \zeta_{C^*} \quad (14)$$

In Figure 4, the upper plots of steps 5 and 6 illustrate this process.

5.2.2 Cluster Model 2 (CM2). Second, because some households have less dependence on temperature, we create a second estimation of the prediction day magnitude by assuming it is equal to the magnitude of the previous day’s consumption, $|\hat{j}_{j,D_{N+1}}^{CM2}| = |\hat{j}_{j,D_N}^{CM2}|$. We denote this method “cluster model 2” or CM2. The final predicted load for the next day D_{N+1} is:

$$\hat{j}_{j,D_{N+1}}^{CM2} = |\hat{j}_{j,D}^{CM1}| \zeta_{C^*} \quad (15)$$

In Figure 4, the lower plots of step 5 and 6 illustrate this process.

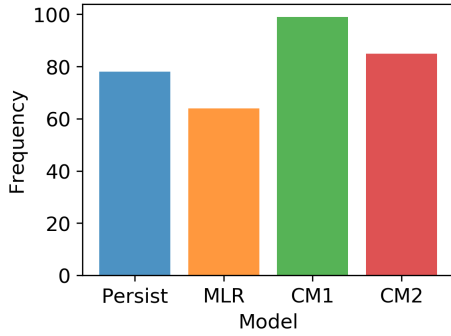


Figure 5: Number of consumers for which each model produces the lowest MAPE.

The steps of the overall clustering and load forecasting process is illustrated in Figure 4.

After the load has been predicted for a particular day, the load history for that consumer is updated with the actual consumption. The actual load profile for that day is classified to the closest cluster centroid and the transition matrix is updated accordingly. The load profile centroids are also updated to be used for forecasting the next day, D_{N+2} .

5.3 Results for the individual load forecasting

Table 2: MAPEs for individual load forecasting models

Model	Training Period	Testing Period
Persistence	56.09	52.18
MLR	55.02	53.52
CM1	56.13	53.28
CM2	52.71	51.30

Table 2 shows the median MAPEs for each forecasting model for the individual consumers. We can see that the median error is the highest for the MLR model, followed by CM1, then the persistence model and finally CM2 has the lowest median error. Interestingly, the errors in the training period are larger than those for the testing period, which is likely a result of many more unpredictable events which occurred over the longer time of the training period compared with the testing.

In terms of the best model for the individual consumers, we find that no model significantly outperforms all the others across all the individual consumers. Figure 5 illustrates this, highlighting the number of consumers who find each model to produce the lowest forecasting error in the testing period. While the total difference in the median MAPE for each model is less than 5%, there are some significant differences between the forecast models for the individual consumers. Figure 6 shows the distribution of the differences in the MAPEs for the individual consumers between the different forecasting models. This suggests that certain individuals may be better predicted by different forecasting models.

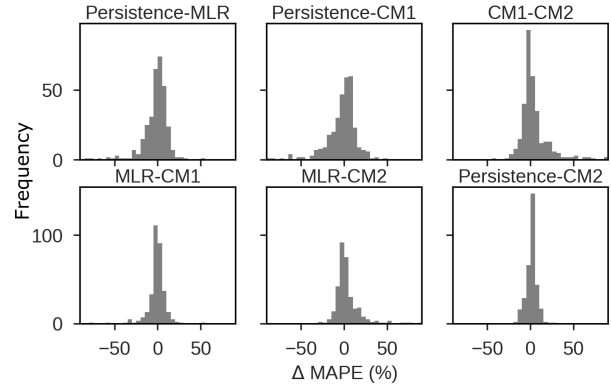


Figure 6: Differences in the MAPEs for all consumers between the different forecasting models. CL1 and CL2 are the abbreviations for cluster model 1 and cluster model 2 respectively.

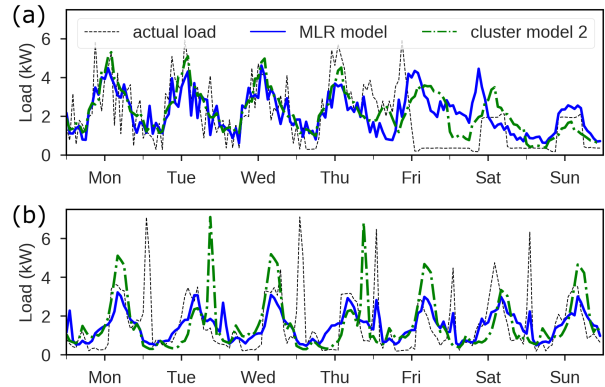


Figure 7: Comparing the MLR and CM2 for two consumers. (a) A consumer for whom the CM2 has the lowest MAPE (b) A consumer for whom the MLR has the lowest MAPE.

In general, the MAPEs for all forecast models are highly correlated, suggesting that if an individual consumer is difficult to predict, it is difficult to predict for all the models and will result in a high MAPE. In particular we find that the results of the MLR and CM1 are highly correlated as are the persistence model and the CM2. This is a result of explicitly including predictors for temperature in the MLR as well as using temperature to predict daily magnitude for the CM1. On the contrary, the temperature is only accounted for implicitly in the persistence and CM2 models, due to the fact that temperatures between one day and the next are highly correlated. Figure 7 shows one week of the predicted loads for two consumers, one for whom the CM2 forecast has the lowest MAPE (Figure 7a) and one for whom the MLR forecast produces the lowest MAPE (Figure 7b). For the first consumer, we see that on this particular week after Friday at midday, their load profile changes suddenly in shape and magnitude, and the error of the MLR forecast is higher due to the inconsistent relationship with temperature. For the second individual, we see that the cluster based forecast CM2 is

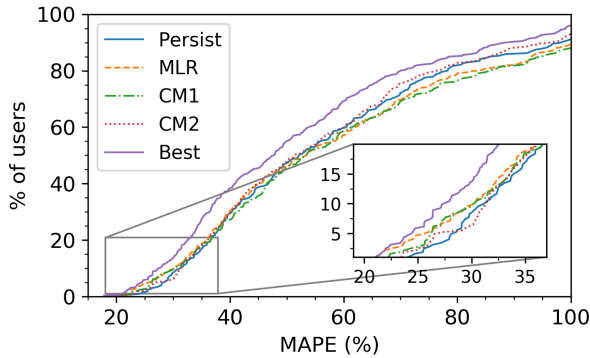


Figure 8: CDF plot of the MAPEs for individual consumers. The inset zooms in on the lower left region. The ‘Best’ line shows the MAPEs when the best performing model is selected for each consumer.

missing the times of the demand spikes for these days, and although the spike magnitudes are significantly less accurate with the MLR model, the timing has lower error, ultimately resulting in a lower MAPE.

Figure 8 shows a CDF plot of the forecast errors for each of the different forecasting models. We note that the smallest errors produced with any forecast occur with the MLR forecast, however it generally performs worse in comparison with the Persistence and CM2 forecasts as larger numbers of consumers are considered. This suggests that there may be a small number of consumers who have a strong predictable dependence on temperature and as a result are well predicted by the MLR model, while others (who are better forecast with the Persistence or CM2 models) exhibit behavior with significant burstiness over successive days and therefore temperature is a less important factor. This further highlights that there may be significant potential for improvement in forecasting large numbers of individual households by using different forecasting methods to predict the loads for different households. The hypothetical case where each consumer is predicted by the best performing model is also shown in Figure 8.

6 DISCUSSION

The ability to generate accurate day-ahead forecasts for highly distributed sections of the network is likely to be of significant importance in a future with an electricity system which operates in a highly decentralized manner. The goal of this paper was to explore whether clustering the load history of individual households could be used to accurately forecast their day ahead consumption and compare the results to a benchmark forecasting method. We developed a load shape forecasting method using a shape-based clustering technique with two options for predicting the daily consumption magnitude (for the CM1 and CM2 forecast models respectively) and compared these with a 24-hour persistence forecast model and a MLR model for each consumer. While both cluster models predicted the same load shape, the method for estimating the daily load magnitude differed, based on the historical dependence of a consumers load on temperature and their previous daily

consumption respectively. The result in terms of prediction was that CM1 produced lower MAPEs for consumers whose response to temperature was more consistent, whereas CM2 produced lower MAPEs for consumers who exhibited sudden large changes in daily consumption magnitudes which persisted for more than one day. The latter could possibly be explained by the presence of multiple day trips away from home. It is notable that the variation in the MAPEs for the individual consumers is high with all models, with the 5th – 95th MAPE percentiles spanning a minimum range of 28% – 108% for CM2 to a maximum range of 26% – 127% for CM1.

Our results illustrate that, in general, unpredictable consumers are less predictable with all models. However, we do observe that certain models produce significantly lower MAPEs for certain consumers. In particular, the *consistency* of the consumer’s response to temperature has a large effect on determining model suitability, as does whether they exhibit the tendency to have very low consumption days (which most likely correspond to vacations or work trips which are hard to predict).

6.1 Limitations and further work

In the individual forecasting models we have been careful not to “peek” into the future of the consumer loads, however, we have assumed that the hourly temperature is known 24 hours in advance. Temperature forecasts are generally considered to be accurate over this timescale [14].

It must also be conceded that the cluster forecasting models (CM1 and CM2) and the other models (MLR and persistence) have some incompatibility in the prediction periods. This is because the cluster models forecast the next day based on the entire load profile of the previous day, whereas the other models are functioning on a rolling 24 hour basis. While we do not believe that this negates the usefulness of the proposed cluster models, further work will seek to make explicitly comparable forecasts. This could involve clustering on subsections of the day or employing a rolling actual-load-to-cluster matching process.

Recent work on load forecasting has also explored combining forecasts [21]. Therefore, a promising avenue for future research is to seek to minimize predicted errors by combining load forecasts. This could be particularly successful if periods during which one technique is significantly better than the others can be robustly identified.

7 CONCLUSIONS

We have developed a method for short term load load forecasting based on clustering the daily load profiles in a consumers load history, and compared the results with a standard MLR forecasting technique.

Our results indicate that, in general, the clustering method offers a marginal improvement over the regression model, however the difference is quite significant for some individual consumers. In particular, consumer’s with large step changes in their daily behavior see lower forecasting errors with the clustering model and daily load scaled according to the previous day.

While clustering methods are a worthwhile addition to short term electric load forecasting, it seems unlikely that any one load

forecasting methodology will consistently produce the lowest error for a wide range of individual electric consumers.

Future work will focus on a deeper understanding of the differences between forecasts, introducing other forecasting methods and investigating whether forecasts can be systematically combined to take advantage of the strengths of individual forecasting methods.

ACKNOWLEDGMENTS

The research was supported in part by grants from the Center for Complex Engineering Systems, the MIT Energy initiative and the MISTI-Brazil exchange program. The authors gratefully thank Riccardo Di Clemente for several hours of useful discussion.

REFERENCES

- [1] Adrian Albert and Ram Rajagopal. 2013. Smart Meter Driven Segmentation : What Your Consumption Says About You. *IEEE Transactions on Power Systems* 28, 4 (2013), 4019–4030. <https://doi.org/10.1109/TPWRS.2013.2266122>
- [2] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific data* 2 (2015).
- [3] Christian Beckel, Leyna Sadamori, Thorsten Staake, and Silvia Santini. 2014. Revealing household characteristics from smart meter data. *Energy* 78 (2014), 397–410. <https://doi.org/10.1016/j.energy.2014.10.025>
- [4] Ignacio Benitez, Alfredo Quilano, Jose-Luis Diez, and Ignacio Delgado. 2014. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. *International Journal of Electrical Power & Energy Systems* 55 (2014), 437–448.
- [5] Antonio Bracale, Guido Carpinelli, Pasquale De Falco, and Tao Hong. 2017. Short-term industrial load forecasting: A case study in an Italian factory. In *Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), 2017 IEEE PES, IEEE*, 1–6.
- [6] Young Tae Chae, Raya Horesh, Youngdeok Hwang, and Young M Lee. 2016. Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. *Energy and Buildings* 111 (2016), 184–194.
- [7] Nathaniel Charlton and Colin Singleton. 2014. A refined parametric model for short term load forecasting. *International Journal of Forecasting* 30, 2 (2014), 364–368. <https://doi.org/10.1016/j.ijforecast.2013.07.003>
- [8] Gianfranco Chicco. 2012. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy* 42, 1 (2012), 68–80. <https://doi.org/10.1016/j.energy.2011.12.031>
- [9] Hamed Chitsaz, Hamid Shaker, Hamidreza Zareipour, David Wood, and Nima Amjadi. 2015. Short-term electricity load forecasting of buildings in microgrids. *Energy and Buildings* 99 (2015), 50–60.
- [10] Howard J Diamond, Thomas R Karl, Michael A Palecki, C Bruce Baker, Jesse E Bell, Ronald D Leeper, David R Easterling, Jay H Lawrimore, Tilden P Meyers, Michael R Helfert, et al. 2013. US Climate Reference Network after one decade of operations: Status and assessment. *Bulletin of the American Meteorological Society* 94, 4 (2013), 485–498.
- [11] Mark E H Dyson, Samuel D. Borgeson, Michaelangelo D. Tabone, and Duncan S. Callaway. 2014. Using smart meter data to estimate demand response potential, with application to solar energy integration. *Energy Policy* 73 (2014), 607–619. <https://doi.org/10.1016/j.enpol.2014.05.053>
- [12] Krzysztof Gajowniczek and Tomasz Ząbkowski. 2017. Electricity forecasting on the individual household level enhanced based on activity patterns. *PLoS one* 12, 4 (2017), e0174098.
- [13] Stephen Haben, Colin Singleton, and Peter Grindrod. 2016. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Transactions on Smart Grid* 7, 1 (2016), 136–144. <https://doi.org/10.1109/TSG.2015.2409786>
- [14] Tao Hong. 2010. *Short term electric load forecasting*. North Carolina State University.
- [15] Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman. 2016. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting* 32, 3 (2016), 896–913. <https://doi.org/10.1016/j.ijforecast.2016.02.001>
- [16] Tao Hong, Jason Wilson, and Jingrui Xie. 2014. Long term probabilistic load forecasting and normalization with hourly information. *IEEE Transactions on Smart Grid* 5, 1 (2014), 456–462.
- [17] Samuel Humeau, Tri Kurniawan Wijaya, Matteo Vasirani, and Karl Aberer. 2013. Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households. In *Sustainable Internet and ICT for Sustainability (SustainIT), 2013. IEEE*, 1–6.
- [18] Irish Social Science Data Archive. 2011. CER Smart Metering project. <http://www.ucd.ie/issda/data/commissionforenergyregulationcer/>
- [19] Jungsuk Kwac, June Flora, and Ram Rajagopal. 2014. Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid* 5, 1 (2014), 420–430. <https://doi.org/10.1109/TSG.2013.2278477>
- [20] Jungsuk Kwac, June Flora, and Ram Rajagopal. 2016. Lifestyle segmentation based on energy consumption data. *IEEE Transactions on Smart Grid* (2016).
- [21] Bidong Liu, Jakub Nowotarski, Tao Hong, and Rafal Weron. 2017. Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Transactions on Smart Grid* 8, 2 (2017), 730–737.
- [22] Yael Parag and Benjamin K Sovacool. 2016. Electricity market design for the prosumer era. *Nature Energy* March (2016), 16032. <https://doi.org/10.1038/nenergy.2016.32>
- [23] Ofir Pele and Michael Werman. 2009. Fast and robust earth mover’s distances. In *Computer vision, 2009 IEEE 12th international conference on. IEEE*, 460–467.
- [24] Mohammad Javad Abdolhosseini Qomi, Arash Noshadran, Jake M Sobstyl, Jameson Toole, Joseph Ferreira, Roland J-M Pellenq, Franz-Josef Ulm, and Marta C Gonzalez. 2016. Data analytics for simplifying thermal efficiency planning in cities. *Journal of The Royal Society Interface* 13, 117 (2016), 20150971.
- [25] Franklin L. Quilumba, Wei Jen Lee, Heng Huang, David Y. Wang, and Robert L. Szabados. 2015. Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities. *IEEE Transactions on Smart Grid* 6, 2 (2015), 911–918. <https://doi.org/10.1109/TSG.2014.2364233> arXiv:arXiv:1502.03908v1
- [26] Joshua D. Rhodes, Charles R. Upshaw, Chioke B. Harris, Colin M. Meehan, David A. Walling, Paul A. Navrátil, Ariane L. Beck, Kazunori Nagasawa, Robert L. Fares, Wesley J. Cole, Harsha Kumar, Roger D. Duncan, Chris L. Holcomb, Thomas F. Edgar, Alexis Kwasinski, and Michael E. Webber. 2014. Experimental and data collection methods for a large-scale smart grid deployment: Methods and first results. *Energy* 65 (2014), 462–471. <https://doi.org/10.1016/j.energy.2013.11.004>
- [27] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover’s distance as a metric for image retrieval. *International journal of computer vision* 40, 2 (2000), 99–121.
- [28] Thanchanok Teeraratkul, Daniel O’Neill, and Sanjay Lall. 2017. Shape-Based Approach to Household Electric Load Curve Clustering and Prediction. *IEEE Transactions on Smart Grid* (2017).
- [29] Reinaldo Tonkoski, Dave Turcotte, and Tarek HM El-Fouly. 2012. Impact of high PV penetration on voltage profiles in residential neighborhoods. *IEEE Transactions on Sustainable Energy* 3, 3 (2012), 518–527.
- [30] Andreas Veit, Christoph Goebel, Rohit Tidke, Christoph Doblender, and Hans-Arno Jacobsen. 2014. Household electricity demand forecasting: benchmarking state-of-the-art methods. In *Proceedings of the 5th international conference on Future energy systems. ACM*, 233–234.
- [31] Jan Von Appen, Martin Braun, Thomas Stetz, Konrad Diwold, and Dominik Geibel. 2013. Time in the sun: the challenge of high PV penetration in the German electric grid. *IEEE Power and Energy magazine* 11, 2 (2013), 55–64.
- [32] RA Walling, Robert Saint, Roger C Dugan, Jim Burke, and Ljubomir A Kojovic. 2008. Summary of distributed resources impact on power delivery systems. *IEEE Transactions on power delivery* 23, 3 (2008), 1636–1644.
- [33] Yi Wang, Qixin Chen, Tao Hong, and Chongqing Kang. 2018. Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid* (2018).
- [34] Sharon Xu, Edward Barbour, and Marta C González. 2017. Household Segmentation by Load Shape and Daily Consumption. (2017).
- [35] Junjing Yang, Chao Ning, Chirag Deb, Fan Zhang, David Cheong, Siew Eang Lee, Chandra Sekhar, and Kwok Wai Tham. 2017. k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy and Buildings* 146 (2017), 27–37.
- [36] Baran Yildiz, Jose I Bilbao, Jonathon Dore, and Alistair B Sproul. 2018. Short-term forecasting of individual household electricity loads with investigating impact of data resolution and forecast horizon. *Renewable Energy and Environmental Sustainability* 3 (2018), 3.
- [37] Chun-Nam Yu, Piotr Mirowski, and Tin Kam Ho. 2017. A sparse coding approach to household electricity demand forecasting in smart grids. *IEEE Transactions on Smart Grid* 8, 2 (2017), 738–748.
- [38] Holger Ziekow, Christoph Goebel, Jens Strucker, and Hans-Arno Jacobsen. 2013. The potential of smart home sensors in forecasting household electricity demand. In *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on. IEEE*, 229–234.