Using Smart Meter Data to Improve the Accuracy of Intraday Load Forecasting Considering Customer Behavior Similarities

Franklin L. Quilumba, *Member, IEEE*, Wei-Jen Lee, *Fellow, IEEE*, Heng Huang, *Member, IEEE*, David Y. Wang, *Senior Member, IEEE*, and Robert L. Szabados, *Member, IEEE*

***Abstract*—With the deployment of advanced metering infrastructure (AMI), an avalanche of new energy-use information became available. Better understanding of the actual power consumption patterns of customers is critical for improving load forecasting and efficient deployment of smart grid technologies to enhance operation, energy management, and planning of electric power systems. Unlike traditional aggregated system-level load forecasting, the AMI data introduces a fresh perspective to the way load forecasting is performed, ranging from very short-term load forecasting to long-term load forecasting at the system level, regional level, feeder level, or even down to the consumer level. This paper addresses the efforts involved in improving the system level intraday load forecasting by applying clustering to identify groups of customers with similar load consumption patterns from smart meters prior to performing load forecasting.**

***Index Terms*—Advanced metering infrastructure (AMI), *k*-means clustering, load forecasting, load patterns, load profiles, neural network-based load forecasting, smart meters.**

I. Introduction

ELECTRICITY is an essential part of modern life. In our homes, we use it for lighting, running appliances and electronics, and often for heating and cooling. Most consumers do not think much about their electricity until a power outage occurs or when they get a high electricity bill. Today, many utilities are deploying smart meters as the first step of moving toward smart grids. However, smart meters generate data at a rate and volume that outpaces the capabilities of many traditional systems. As a result, much of the data is collected, but not analyzed. One of the most important challenges for electric utilities in the smart grid era is to use the smart meter data beyond its core function of measuring the interval data for customer billing. This will open the way for unprecedented opportunities that give electric utilities the ability to better manage their power grid and enable consumers to better control their consumption.

Smart meters are vital components of smart grid technology, capable of capturing customer consumption at frequent intervals (and possibly other parameters) using communication networks. With the adoption of advanced metering infrastructure (AMI), an avalanche of new energy-use information became available. One of the most promising applications for this newly available data is to improve load forecasting accuracy.

Conventionally, load forecasting is conducted using system level data with little or even no information regarding power consumption profiles at lower levels (e.g., regional level, substation level, transformer level, feeder level, or household level). Load forecasting with high-voltage level data is well-established among researchers in both academia and industry, as reflected by [1]–[4]. On the other hand, load forecasting utilizing smart meter data is still limited due to the previous lack of household level load data [5], [6].

Since smart meters are located at an endpoint, they can be aggregated in a variety of ways based on a set of criteria or common characteristics. For instance, the meters that are located at a distinguishing geographical location (e.g., neighborhood); connected to a particular feeder, transformer, substation, or whole distribution system; or belong to a specific service class can be aggregated together. Virtually any grouping is possible, and it is here where clustering may help researchers discover patterns in load consumption prior to forecasting as an aid for improving accuracy.

This paper presents the efforts involved in utilizing the AMI data to improve the load forecasting accuracy at the system level. We propose a three-step process to accomplish this. First, clustering approaches are used to group customers (smart meters) based on similarities in consumption behavior. Then, load forecasting is conducted for each group. Finally, the results from each group are aggregated to obtain the forecast at the system level. To demonstrate the applicability of our proposed approach, real-world smart meter data for residential customers of two different electric utility companies are used in this paper.

II. Literature Review and Problem Relevance

Load forecasting is crucial to power systems planning and operations. This section provides a brief literature review on

Manuscript received January 28, 2014; revised May 2, 2014, July 24, 2014, and September 13, 2014; accepted October 9, 2014. Date of publication November 3, 2014; date of current version February 16, 2015. Paper no. TSG-00065-2014.

F. L. Quilumba is with the Electrical Energy Department, National Polytechnic School, Quito EC170102, Ecuador (e-mail: quigufrale@ieee.org).

W.-J. Lee and H. Huang are with the Energy Systems Research Center, University of Texas at Arlington, Arlington, TX 76019 USA.

D. Y. Wang and R. L. Szabados are with the Consolidated Edison Company of New York, Inc., New York, NY 10003 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2014.2364233

load forecasting with the focus on the methods that utilize hierarchical information for very-short and short-term load forecasting.

A. Brief Overview of Load Forecasting

Load forecasting can be categorized by very short-term load forecasting (VSTLF), short-term load forecasting (STLF), medium-term load forecasting, and long-term load forecasting (LTLF), among which the forecasting horizon cut-offs are one day, two weeks, and three years, respectively [5], [7]–[9].

Although technical literature presented a wide range of methodologies and models to improve the accuracy of load forecasting, most of them are based upon aggregated power consumption data at the system (top) level with little to no information regarding power consumption profiles of different customer classes. The AMI data introduces a fresh perspective to the way load forecasting is performed ranging from VSTLF to LTLF at the system level, regional level, feeder level, and even consumer level.

Different techniques have been developed for VSTLF with system (top) level data. For instance, Taylor [3] presented and compared several univariate methods including a comparison with an additional approach based on weather forecasts. The author determined that for very short-term predictions, univariate models can be useful when the lead time falls between 10 and 30 min. In addition, beyond VSTLF, combining methods based on weather forecasts are promising to forecast load beyond an hour ahead. Guan *et al.* [10] used separate neural networks (NNs) applied to wavelet decomposed filtered load data. The results are combined to produce the final forecast consisting of forecast of loads 1 h into the future in 5-min steps in a moving window manner.

Many different techniques have been introduced for STLF with system (top) level data. A comprehensive review on this subject can be found in [2] and [4]. In general, time series models, traditional econometric models, artificial NN-based models, fuzzy logic-based models, nonparametric/semi-parametric regression-based models, hybrid models (i.e., combination of different models like neuro-fuzzy models, among others), and judgmental forecasting models have been applied with relatively high accuracy to STLF. For example, Fan and Hyndman [11] presented semi-parametric additive models to estimate the relationships between demand and the driver variables such as calendar variables, lagged actual demand observations, and historical and forecast temperature traces. These models allow nonlinear and nonparametric terms within the regression framework which can capture the complex nonlinear relationship between electricity demand and its drivers. This methodology was used to forecast the half-hourly electricity demand for up to seven days ahead.

State-of-the-art load forecasting techniques were presented in the hierarchical load forecasting track at the Global Energy Forecasting Competition 2012 [1]. In this competition, zonal level data were provided to mimic the forecasting job in the smart grid era where the forecasters have access to smart meter information. Recently, machine learning techniques [12], fuzzy logic [7], and exponentially weighted [13] approaches have been used for load forecasting or classification and performed relatively well.

B. Grouping Methods in Load Forecasting

With customers being spread across a large territory, the electricity consumption patterns may be different from one delivery point to another. Such hierarchical load information is readily available from smart meter data which allows developing models from the bottom or middle level of the hierarchy. Although this aspect is emphasized in [1], nobody did grouping in the GEFCom2012 competition.

The influence of the weather on electricity consumption is well known where temperature is usually the main variable driving the electricity demand. Therefore, improving temperature forecasting accuracy is an effective way to improve load forecasting accuracy. On this manner, Fan *et al.* [14] proposed combining forecasting from several alternative meteorological predictions to share the strength of the different temperature forecasts. This is particularly important in the smart grid era where the smart meters are deployed in large geographical areas and require multiple sources of weather information.

Two considerations, geographic hierarchy and weather station data, were studied in [15] as means of improving load forecast accuracy. They empirically performed various case studies and concluded that a single temperature cannot provide a good representation of weather across a large geographical area, and grouping similar regions together provided forecasts that were almost as good as summing the individual forecasts. Therefore, taking advantage of diverse weather stations and using lower level load information (grouping by regions) can enhance higher level load forecasting accuracy.

Silva *et al.* [6] investigated the positive impact of grouping on forecasting accuracy, and the effects that forecasting errors have on trading in an intraday local electricity market composed by consumers and prosumers. The impact of grouping on load forecasting accuracy was performed by randomly creating groups of different sizes from smart meter data and measuring the resulting accuracy with the mean absolute percentage error (MAPE). They concluded that the forecast accuracy increases with group size.

C. Load Forecasting With Smart Meter Data

In the published literature, researchers have focused their study on: 1) longitudinal and 2) cross-sectional grouping when dealing with forecasting load data. Longitudinal grouping refers to identifying time periods with similar consumption characteristics (load patterns) from historical data. This clustering method is frequently used in conjunction with the similar-day method for load forecasting. Cross-sectional grouping refers to the aggregation of customers (smart meters) with similar load patterns. Previous paper on cross-sectional grouping includes applications in the retail power market (e.g., tariff design) [16] and creation of new load profiles that can be applied to improve distribution network analysis [17].

Load forecasting with real-world smart meter data can be summarized as follows.

Marinescu *et al.* [18] examined six methods previously used to predict large-scale energy demand to forecast a load similar to that of a single transformer. Artificial NNs, auto-regressive, auto-regressive moving average, auto-regressive integrated moving average, fuzzy logic, and wavelet NNs were utilized for day-ahead and week-ahead electric load forecasting in a scenario with 90 houses and another with 230 houses. The authors concluded that at a small scale, the noise and chaotic behavior had a great impact on the forecast accuracy.

Chaouch [5] proposed forecasting functional time series applied to intraday household-level load curves using smart meter data via two methods: 1) functional wavelet-kernel (FWK); and 2) clustering-based FWK. Both approaches are identical, except the latter identifies a common pattern between days for each customer (following the idea of similar-day approaches) through an unsupervised classification method and then utilizes FWK to perform one day-ahead forecasting of each customer. It is also stated that household loads are very volatile, which makes household level forecasting difficult to solve.

Kwac *et al.* [19] utilized smart meter data for customer segmentation. First, daily usage patterns were decomposed into daily total usage and normalized daily load shape by means of a two-stage clustering: 1) an adaptive k-means; and 2) a hierarchical clustering. Then, an entropy analysis was performed to establish a method of customers' segmentation for demand response and energy efficiency targeting.

D. Problem Relevance and Contribution

Taking into consideration the above description, it is clear that despite the fact that load forecasting is a challenging task at any level and at any time horizon, it is especially difficult to forecast load at the household level with fine granularity data. Forecasting individual load consumption will require a volume of high-resolution data that necessitates extensive computing resources. Moreover, to apply any load forecasting technique to each household will require a detailed investigation of the influences on individual load consumption behavior in order to adequately capture complex dynamics.

Two general interests for using smart meter data in load forecasting are revealed: 1) forecasting individual household loads; and 2) aggregating loads and constructing a single forecasting model for the system load. However, we propose a different approach that evaluates smart meter data (lower level) to forecast high-level load, e.g., system load and considers individual volatile loads grouped by consumption behavior.

III. Methods

A. Data Collection

Real-world smart meter data for residential customers of two different electric utility companies, from the United States and Ireland, are used in this paper.

Starting in 2009, the Consolidated Edison Company of New York, Inc. (Con Edison) initiated a smart grid project aimed at deploying a wide range of grid-related technologies that include automation, monitoring, and two-way communications to make the electric grid function more efficiently and to enable the integration of renewable resources and energy-efficient technologies [20]. We used 21 months of 15-min load data from February 2012 to October 2013. The first 12 months of the data were used for estimating the model parameters, and observations from the last nine months were used for model evaluation.

The Commission for Energy Regulation [21] publicly released anonymized full data sets of the recent "*Electricity Smart Metering Customer Behavior Trials*" in Ireland recorded from July 14, 2009 to December 31, 2010 and with over 5000 Irish homes and businesses. The data were obtained via the Irish Social Science Data Archive (ISSDA) [22] on October 2013. The smart meters recorded electricity consumption at a resolution of 30 min. Although there are three available service classes from the data set: 1) residential; 2) small-to-medium enterprises (commercial); and 3) other, only residential customers were considered for this paper, and no distinction was made for any tariff program. We used 17 months of half-hourly load data from August 2009 to December 2010. Similarly, the first 12 months of the data were used for estimating the model parameters, and observations from the last five months were used for model evaluation.

Missing interval data were encountered and treated individually for each smart meter through data preprocessing. Missing interval data could be associated with transmission problems at either the receiving or sending end. These records must be completed, corrected, or eliminated to ensure the accuracy of the ultimate predictive model. In this paper, missing values in a load profile were imputed by the mean of the points in the vicinity of the missing ones [23].

In addition to these data sets, temperature data at both locations for their respective time periods were obtained from wunderground.com.

With all these data, we generated point forecasts using a rolling forecast for time horizons varying from 15 min or 30 min for up to one day-ahead predictions.

B. Smart Meter Load Data Grouping Based on Consumption Behavior Similarities

Figs. 1 and 2 show a daily "system" load profile composed by the aggregation of all residential customers and a single residential customer load profile on July 17, 2012 and December 25, 2009 for the two datasets, respectively. It can be observed in both cases that household consumption across the day is very different from that of an aggregated system load profile. Each household will have an individual daily load curve, though each will be different because each home has different appliances and is occupied by people with different schedules and usage preferences.

Figs. 3 and 4 show six different residential customer profiles (at random) on April 28, 2012 and August 12, 2009 for both datasets, respectively. It can be seen that the load pattern consumption changes between every customer. Furthermore, load consumption can change on a daily basis for even a single customer as shown in Figs. 5 and 6.

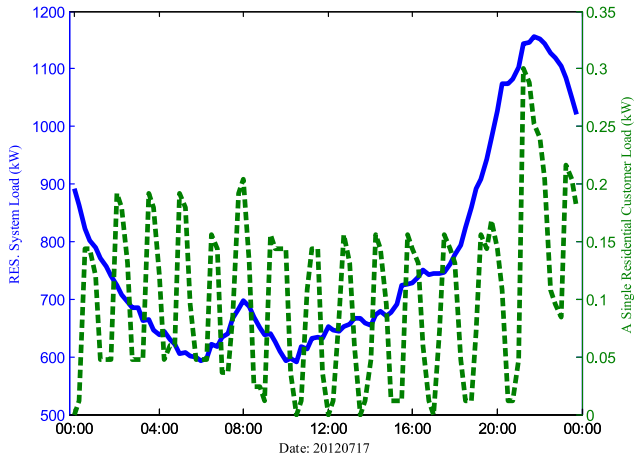

Fig. 1. Daily load profile for residential customers for residential system demand and a single residential customer across a 24-h period on July 17, 2012.

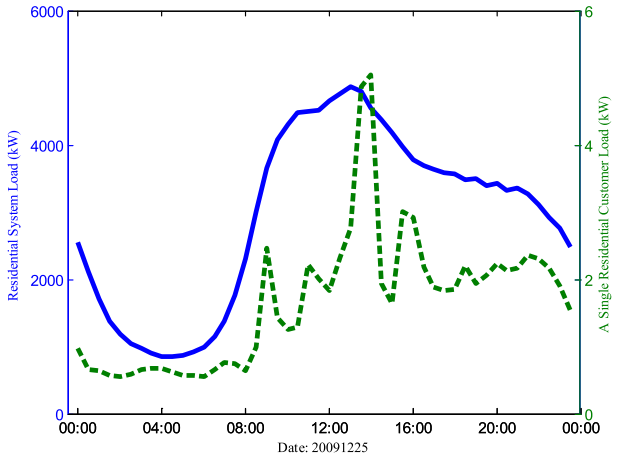

Fig. 2. Daily load profile for residential customers for residential system demand and a single residential customer across a 24-h period on December 25, 2009.

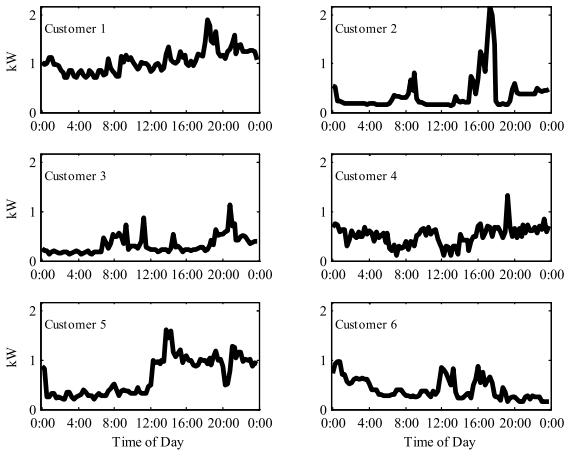

Fig. 3. Daily load profiles for six residential customers chosen at random illustrating variation between household consumers on April 28, 2012.

It is clear that load consumption differs in both magnitude and time of use and is dependent on lifestyle, weather, and many other uncontrollable factors. Thus, we seek to cluster

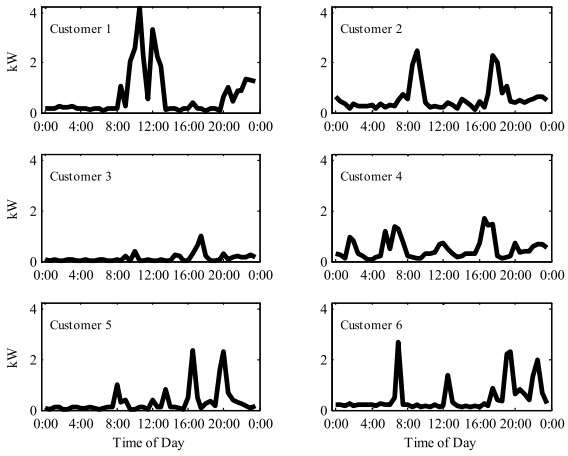

Fig. 4. Daily load profiles for six residential customers chosen at random illustrating variation between household consumers on August 12, 2009.

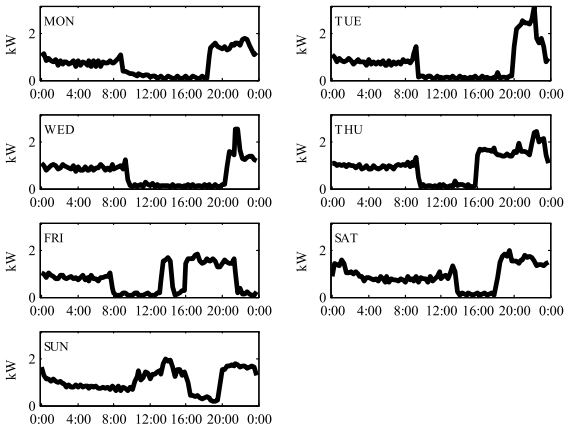

Fig. 5. Daily load profiles for a single customer chosen at random over a weekly period for dataset 1.

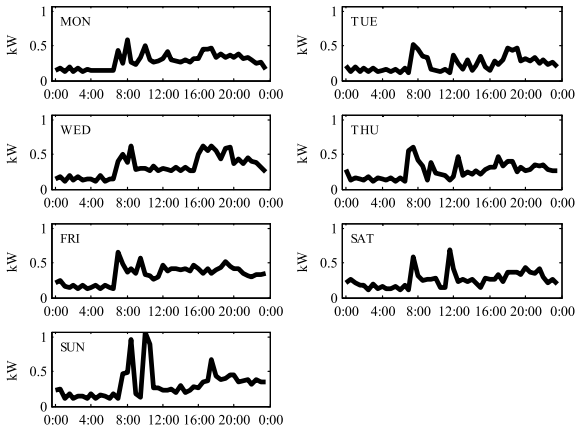

Fig. 6. Daily load profiles for a single customer chosen at random over a weekly period for dataset 2.

load customers in a meaningful way by taking into consideration these inherent daily and intradaily variations [24] that can/will improve current practices on load forecasting.

1) Data Clustering: Data clustering is concerned with exploring data sets to assess whether they can be summarized

meaningfully in terms of a relatively small number of groups. Given a representation of m objects, the goal is to find K groups based on a measure of similarity such that the similarities between objects in the same group are high while the similarities between objects in different groups are low [25]. An ideal cluster can be defined as a set of points that is compact and isolated.

2) Clustering Algorithm: Numerous clustering algorithms have been developed, and determining the best one depends greatly on the nature of the dataset and what constitutes meaningful clusters in an application [25]. After careful consideration, k-means, a robust and widely used clustering technique, was adopted for clustering load profiles.

a) k-means clustering algorithm [26]: Let D be a data set with m instances, and let $C_1, C_2, \ldots, C_k$ be the k disjoint clusters of D. The error function is defined as

$$E = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} d\left(\mathbf{x}, \mu\left(C_i\right)\right)$$

where $\mu\left(C_i\right)$ is the centroid of cluster C_i. $d\left(\mathbf{x}, \mu\left(C_i\right)\right)$ denotes the distance between $\mathbf{x}$ data point and $\mu\left(C_i\right)$, a typical choice of distance measure is the Euclidean distance $d_{\text{euc}}(\cdot, \cdot)$ [27], defined as

$$d_{\text{euc}}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2\right)^{\frac{1}{2}}$$

for any data points $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ and $\mathbf{x}_j = \left(x_{j1}, x_{j2}, \ldots, x_{jm}\right)$.

k-means clustering algorithm

1: Choose k cluster centers to coincide with k randomly-chosen patterns or k randomly defined points inside the hypervolume containing the pattern set.
2: Assign each pattern to the closest cluster center.
3: Recompute the cluster centers using the current cluster memberships.
4: If a convergence criterion is not met, go to step 2. Typical convergence criteria are: no (or minimal) reassignment of patterns to new cluster centers or minimal decrease in squared error.

Typically, k-means is run independently for different values of k and the partition that appears the most meaningful to the domain expert is selected. Different initializations can lead to different final clusters because k-means only converges to a local minimum. One way to overcome the local minima is to run the k-means algorithm with multiple initial partitions and choose the partition with the smallest squared error [25].

3) Clustering Implementation for Load Pattern Grouping: Several different cases (empirically) were assessed to group the smart meter data based on the load pattern behavior. While many cases can be suggested and examined, the following is a description of the methodology implemented in our paper that performed well for load pattern grouping in both data sets.

Start by selecting the season of the year where the load peak occurs during the whole year in the training set, and do the following for each of the m customers.

1) Divide each day into five segments corresponding to main intraday consumption behavior patterns.
2) Obtain an average consumption at each day of the week. This represents the average consumption pattern of a single customer for every day of a typical week.
3) Normalize the load in a range of 0–1 to emphasize grouping the customers according to who contributes to the total consumption at a certain time of the day, also known as coincident demand. Another benefit is that a more equally distributed number of customers at each cluster are obtained.

The data points can be arranged in the m-by-n data matrix D, where m is the number of meters, and n is the number of features. Therefore, the dimension of D is equal to 35 (five segments per day for the seven days of a typical week). At this moment, k-means is applied with k ranging from 1 to 12, and with 1000 repetitions to overcome local minima.

Now, rather than using any clustering validity index to decide on a suitable number of clusters, we pursued a different venue. Since our ultimate goal is to improve load forecasting by grouping customers based on their consumption behavior, we decided to evaluate the number of clusters based on forecasting performance. In this paper, we use the MAPE to measure the forecasting performance, and therefore MAPE was used to determine how many clusters are adequate. The MAPE is defined as the ratio between absolute forecast errors and the actual values

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|.$$

The idea is to find the optimal number of clusters by minimizing the aggregated load forecasting error for the testing data set [28]. The forecasting error generally decrease until the number of optimal clusters is reached. After that, the forecasting error increases as the number of clusters increases.

C. Smart Meter Data to Enhance the Performance of Load Forecasting

After each smart meter was assigned to a specific cluster, we summed the smart meter interval data in the group to obtain the partial system load forecast. We then forecast this partial system load at each group. Finally, the partial system load forecasts were summed to obtain the total system load forecast. At this point, we evaluated the aggregated load forecasting accuracy. Since our interest is to utilize AMI data for load forecasting, this paper focuses on sub-hourly forecasting with different time horizons up to one day ahead.

In this paper, we used NNs because NN models are widely used by many utilities in practice [4]. Moreover, since NNs can approximate any continuous function, they can be seen as a multivariate nonlinear method that can model complex nonlinear relationships. In addition, NNs are data-driven methods, and are therefore well-suited for use with smart meter data.

D. NN-Based Load Forecasting

When designing a NN-based forecasting model, the first step is to select an appropriate architecture. Although there are many types of NNs, most NN-based load forecasting models utilize a feed-forward multilayer perceptron (MLP) with satisfactory results in terms of accuracy [29]. In a typical MLP, neurons are organized in layers: one input layer, one or more hidden layers, and one output layer. Once the MLP-NN architecture is selected, one must decide the number of input nodes, the number of hidden layers and the type of activation function, and the number of output nodes.

Considering that our purpose is to forecast the load with a horizon of one day ahead at a resolution equal to the meters' resolution, there are essentially two ways on doing so.

1) Use the NN model to forecast one step ahead. Note that for leading times larger than the time interval considered, the one step ahead forecasts can be iteratively used as inputs in order to generate multistep predictions [3], [13].
2) Use the NN model to forecast multisteps ahead either by using a system of NNs with one for each time interval or by a large NN with many outputs corresponding to each interval for a full day-ahead forecast [29].

In this paper, we explore method one. Therefore, our NN-based forecasting model's architecture is an MLP with one hidden layer with a hyperbolic tangent activation function and one linear output neuron.

The parameters of this NN are the weights associated with the connections from the input nodes to the hidden layer, and the weights for the connections from the hidden layer to the output node. These are found by "training the NN," and its purpose is to find the weighting matrices that minimize a loss function [29].

The Levenberg–Marquardt approach was used to train the model. This approach is suitable for training medium-size NNs with low mean square error. One of the key problems in NN application is to select the number of hidden neurons in the hidden layer which affects the learning process and forecasting capability of the network. We adopted the approach similar as in [14] to overcome this problem. The method starts by choosing a small number of hidden neurons and gradually increases this number. At each stage, the model is trained, and a forecast error from the testing set is recorded for comparison. The process stops when the error decreases to an acceptable threshold or when no significant improvement is observed as the number of hidden neuron increases. Another issue that occurs during NN training is called overfitting, where the NN loses its generalization ability. We tackled this problem by using regularization. Through detailed analysis, we determined that using 20 hidden neurons and a regularization parameter of 0.9 performed well. Moreover, because the estimation of the weighting matrices is sensitive to the choice of initial values, each model was estimated 20 times from random initial values. Then, the best model was determined by minimizing the out-of-sample MAPE.

Once all the possible input variables were identified, we began with the full model containing all the variables. The predictive capacity of each variable was tested independently by removing each term from the model while retaining all other terms. Omitted variables that led to a decrease in MAPE were excluded from the model for subsequent tests [11]. The variables that we used in our model are the following.

1) Smart meter data variables:
   a) interval (sub-hourly) load variables;
   b) lagged loads at a sub-hourly resolution: last 3 h same day; last 3 h day before (plus same hour day before); last 3 h previous week (plus same hour previous week).
2) Calendar variables:
   a) day of the week variables;
   b) holiday variables;
   c) Monthly variables.
3) Weather variables:
   a) *temperature variables:* Temperature was interpolated between neighboring values to obtain measurements at a resolution similar as the smart meter data (e.g., 15 or 30 min). Only historical values were considered and no temperature forecasts were used as input for the LF.

IV. Forecasting Results

In this paper, we study the application of clustering by load consumption based on smart meter data at the household level to improve the performance of load forecasting at the system level. Our proposed method has been implemented with two different real smart meter datasets to demonstrate the effectiveness of our approach. As explained in the preceding sections, we constructed 78 independent NN-based forecasting models at each of the groups with k varying from 1 to 12. To determine the "optimal" number of clusters, we evaluated the MAPE at each interval of the day for each day of the testing period. Then, the mean MAPE was calculated for the whole testing period.

For dataset 1, we determined that three clusters give a reduction in MAPE of approximately 0.5% with respect to its counterpart with only 1 cluster for one day-ahead forecasting. Fig. 7 depicts this result on predicting the load at different lead times: 15 min ahead, 30 min ahead, 1 h ahead, 2 h ahead, . . . , 24 h ahead for six clusters. Although we calculated for 12 clusters, only six are depicted for the sake of clarity. Fig. 8 shows the average load profiles for each one of the three clusters during July 15, 2013 to July 21, 2013. It can be seen that the load profiles are different from all clusters.

For dataset 2, we determined that four clusters gave a reduction in MAPE of approximately 1.07% with respect to its counterpart with only 1 cluster at one day-ahead forecasting. Fig. 9 depicts this result for only six clusters for the sake of clarity. Fig. 10 shows the average load profiles for each one of the four clusters during December 20, 2010 to December 26, 2010. It is obvious that the load profiles are completely different from all clusters.

Experience in energy forecasting dictates that forecast accuracy increases with increasing group size, so the more meters grouped together, the greater the accuracy. This implies that if $k = 1$, one should expect to get the most accurate forecast result because grouping hides the stochastic behavior of all the

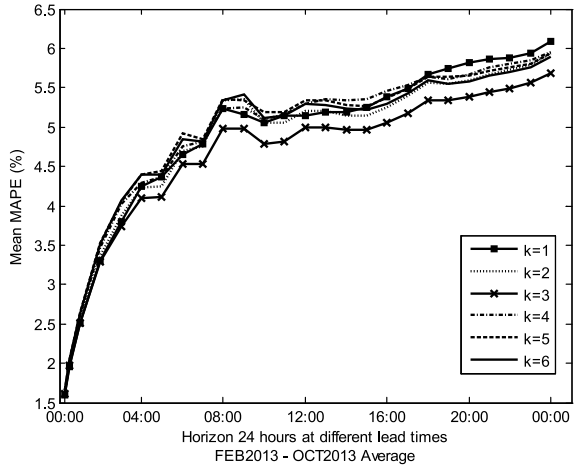

Fig. 7. MAPE results plotted against lead time for the nine-month out-of-sample period for lead times of 15 min ahead, 30 min ahead, 1 h ahead, 2 h ahead, . . . , 24 h.

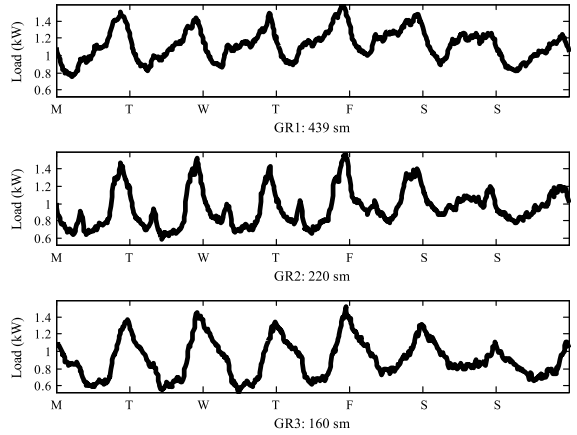

Fig. 8. Load profiles of the three groups of meters when $k = 3$, the optimal number of clusters in dataset 1.

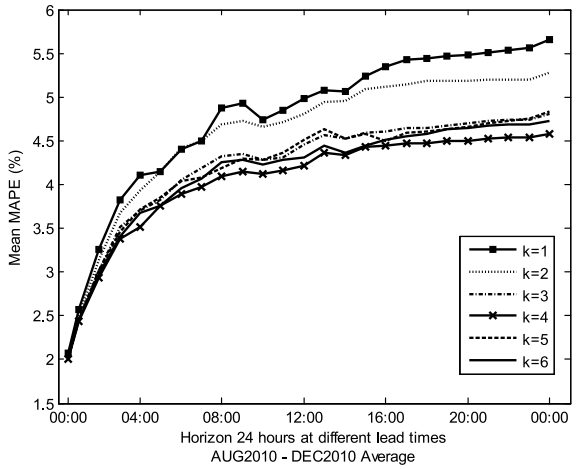

Fig. 9. MAPE results plotted against lead time for the five-month out-of-sample period for lead times of 30 min ahead, 1 h ahead, 2 h ahead, . . . , 24 h.

smart meters. However, we have proven with our approach that forecast errors can be reduced by effectively grouping different customers based on consumption behavior, forecasting the load at each group, and then aggregating the forecasted load to derive the system level load forecast.

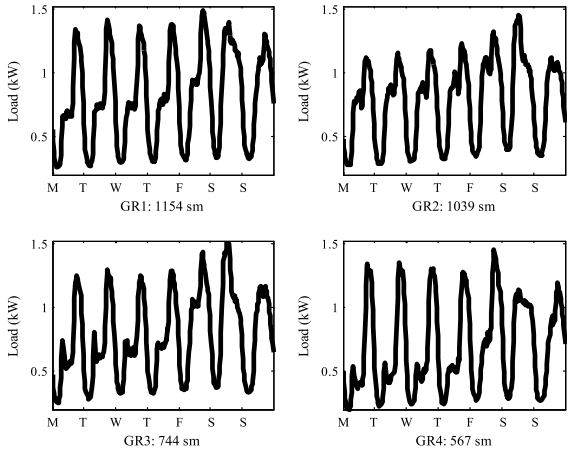

Fig. 10. Load profiles of the four groups of meters when $k = 4$, the optimal number of clusters in dataset 2.

V. Conclusion

The smart grid network paradigm relies on the exploitation of smart meter data to improve customer experience, utility operations, and advanced power management. We demonstrated the application of clustering to group customers by load consumption similarities as an aid to improve system level load forecasting. We showed how household load data from smart meters could be used to improve the load forecasting of the entire system by combining the forecasts from each group. The applicability of our approach was demonstrated on two different datasets from real-world residential smart meter data. Although this paper is based upon residential loads, it is expected that the clustering approach will yield similar results for both commercial and industrial customers.

References

[1] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *Int. J. Forecasting*, vol. 30, no. 2, pp. 357–363, 2014.

[2] T. Hong, "Short term electric load forecasting," Ph.D. dissertation, Oper. Res. Elect. Eng., North Carolina State Univ., Raleigh, NC, USA, 2010.

[3] J. W. Taylor, "An evaluation of methods for very short-term load forecasting using minute-by-minute British data," *Int. J. Forecasting*, vol. 24, no. 4, pp. 645–658, 2008.

[4] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans. Power Syst.*, vol. 16, no. 1, pp. 44–55, Feb. 2001.

[5] M. Chaouch, "Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 411–419, Jan. 2014.

[6] P. G. D. Silva, D. Ilic, and S. Karnouskos, "The impact of smart grid prosumer grouping on forecasting accuracy and its benefits for local electricity market trading," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 402–410, Jan. 2014.

[7] T. Hong and P. Wang, "Fuzzy interaction regression for short term load forecasting," *Fuzzy Optim. Decis. Mak.*, vol. 13, pp. 91–103, Mar. 2014.

[8] Y. Goude, R. Nedellec, and N. Kong, "Local short and middle term electricity load forecasting with semi-parametric additive models," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 440–446, Jan. 2014.

[9] T. Hong, J. Wilson, and J. Xie, "Long term probabilistic load forecasting and normalization with hourly information," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 456–462, 2014.

[10] C. Guan, P. B. Luh, L. D. Michel, Y. Wang, and P. B. Friedland, "Very short-term load forecasting: Wavelet neural networks with data pre-filtering," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 30–41, Jan. 2013.
[11] S. Fan and R. J. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 134–141, Feb. 2012.
[12] S. Fan, L. Chen, and W.-J. Lee, "Machine learning based switching model for electricity load forecasting," *Energy Convers. Manage.*, vol. 49, no. 6, pp. 1331–1344, 2008.
[13] J. W. Taylor, "Short-term load forecasting with exponentially weighted methods," *IEEE Trans. Power Syst.*, vol. 27, no. 1, pp. 458–464, Feb. 2012.
[14] S. Fan, L. Chen, and W.-J. Lee, "Short-term load forecasting using comprehensive combination based on multimeteorological information," *IEEE Trans. Ind. Appl.*, vol. 45, no. 4, pp. 1460–1466, Jul./Aug. 2009.
[15] S.-H. Lai and T. Hong, "When one size no longer fits all—Electric load forecasting with a geographic hierarchy," *SAS*, 2013. [Online]. Available: http://www.sas.com/en_us/whitepapers/when-one-size-no-longer-fits-all-106226.html
[16] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, pp. 68–80, Jun. 2012.
[17] I. Benítez, A. Quijano, J.-L. Díez, and I. Delgado, "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers," *Int. J. Elect. Power Energy Syst.*, vol. 55, pp. 437–448, Feb. 2014.
[18] A. Marinescu, C. Harris, I. Dusparic, S. Clarke, and V. Cahill, "Residential electrical demand forecasting in very small scale: An evaluation of forecasting methods," in *Proc. 2013 2nd Int. Workshop Softw. Eng. Challenges Smart Grid (SE4SG)*, San Francisco, CA, USA, pp. 25–32.
[19] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Trans. Smart Grid*, vol. 5, no.1, pp. 420–430, Jan. 2014.
[20] (2009, Aug. 4). *Con Edison Launches Smart Grid Pilot Program in Queens*. [Online]. Available: http://www.coned.com/newsroom/news/pr20090804_2.asp
[21] CER. (2013, Nov. 10). *Smart Metering Trial Data Publication*. [Online]. Available: http://www.cer.ie/en/information-centre-reports-and-publications.aspx?article=5dd4bce4-ebd8-475e-b78d-da24e4ff7339
[22] *Data from the Commission for Energy Regulation*. (2013, Sep. 20). [Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/
[23] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "An overview of AMI data preprocessing to enhance the performance of load forecasting," in *Proc. 2014 IEEE Ind. Appl. Soc. Ann. Meeting (IAS)*, Vancouver, BC, Canada.
[24] F. McLoughlin, A. Duffy, and M. Conlon, "Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study," *Energy Build.*, vol. 48, pp. 240–248, May 2012.
[25] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, pp. 651–666, Jun. 2010.
[26] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia, PA, USA: SIAM, 2007.
[27] C. A. Ratanamahatana *et al.*, "Mining time series data," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. New York, NY, USA: Springer, 2010, pp. 1049–1077.
[28] S. Fan, K. Methaprayoon, and W.-J. Lee, "Multiregion load forecasting for system with large geographical area," *IEEE Trans. Ind. Appl.*, vol. 45, no. 4, pp. 1452–1459, Jul./Aug. 2009.
[29] *Day-Ahead/Hour-Ahead Forecasting for Demand Trading: A Guidebook*, EPRI, Palo Alto, CA, USA, 2001.

Franklin L. Quilumba (S'10–M'14) received the Diploma degree in electrical engineering from the National Polytechnic School [Escuela Politecnica Nacional (EPN)], Quito, Ecuador, in 2008, and the M.S. and Ph.D. degrees in electrical engineering from the University of Texas at Arlington (UTA), Arlington, TX, USA, in 2014.

He was a Post-Doctoral Research Associate with the Energy Systems Research Center, UTA. In 2014, he joined the faculty of the National Polytechnic School, where he is currently an Assistant Professor. His current research interests include power systems analysis, operation, stability and control, computer simulation of electric power systems, power load modeling, generation and transmission planning, demand response, and load forecasting.

Wei-Jen Lee (S'85–M'85–SM'97–F'07) received the B.S. and M.S. degrees from National Taiwan University, Taipei, Taiwan, and the Ph.D. degree from the University of Texas at Arlington, Arlington, TX, USA, in 1978, 1980, and 1985, respectively, all in electrical engineering.

In 1985, he joined the University of Texas at Arlington, where he is currently a Professor in the Electrical Engineering Department and the Director of the Energy Systems Research Center. His current research interests include power flow, transient and dynamic stability, voltage stability, short circuits, relay coordination, power quality analysis, renewable energy, and deregulation for utility companies.

Prof. Lee is a registered Professional Engineer with the State of Texas.

Heng Huang (M'07) received the B.S. degree in automation and the M.S. degree in instrumentation engineering from Shanghai Jiao Tong University, Shanghai, China, in 1997 and 2001, respectively, and the Ph.D. degree in computer science from Dartmouth College, Hanover, NH, USA, in 2006.

He joined the Computer Science and Engineering Department, University of Texas at Arlington, Arlington, TX, USA, as an Assistant Professor in 2007, where he is the Director of the Biomedical Computing and Scientific Visualization Laboratory. His current research interests include pattern recognition, data mining, computer vision, and bioinformatics.

David Y. Wang (S'90–M'90–SM'07) received the B.S. degree in electrical engineering from the Shanghai University of Engineering Science, Shanghai, China, in 1988; the M.S. degrees in electrical engineering and computer science from the New Jersey Institute of Technology, Newark, NJ, USA, and New York University, New York, NY, USA, in 1990 and 1998, respectively; and the Ph.D. degree in electrical engineering from Polytechnic University, Brooklyn, NY, USA, in 2006.

He is a Technical Expert with the Distribution Engineering Department, Con Edison, New York. He joined Con Edison's Research and Development Department in 1991, where he is currently responsible for principle distribution system design and analysis software.

Robert L. Szabados (S'86-M'88) received the B.S. degree in electrical engineering from Manhattan College, Bronx, NY, USA, in 1988.

Mr. Szabados worked for Ebasco Services/Raytheon Inc., from 1988 to 1999, in the Distribution Engineering Department. He joined the Distribution Engineering Network System Group, Con Edison, New York, NY, in 1999, where he is currently a Senior Engineer responsible for network system design.