A Framework for Baseline Load Estimation in Demand Response: Data Mining Approach

Saehong Park, Seunghyoung Ryu, Yohwan Choi, Hongseok Kim
Department of Electronic Engineering
Sogang University, Seoul, Korea
Email: {saehong, shryu, yohwanchoi, hongseok}@sogang.ac.kr

***Abstract*—In this paper we propose a framework of customer baseline load (CBL) estimation for demand response in Smart Grid. The introduction of demand response requires quantifying the amount of demand reduction. This process is called the *measurement and verification*. The proposed framework of CBL estimation is based on the unsupervised learning technique of data mining. Specifically we leverage both the self organizing map (SOM) and K-means clustering for accurate estimation. This two-level approach efficiently reduces the high dimension of the input vectors into two dimensional output using SOM, and then this output vectors can be efficiently clustered together by K-means clustering. Hence we can easily find the load pattern that is expected to be similar to the potential load pattern of the day of demand response (DR) event. To validate our method we perform large scale experiments where the building complex power consumption is monitored by 2,500 smart meters. Our experiments show that the proposed technique outperforms a series of the day matching methods. Specifically, we find that the root mean square error is reduced by 15–22% in average, and the mean absolute percentage error is reduced by 15–20% in average as well.**

I. Introduction

In the electricity industry, demand side management (DSM) refers to the programs that attempt to influence customers' electricity consumption patterns to match electricity supply. Demand response (DR) is one of the key components of DSM [1], [2]. DR is defined and used by the U.S Department of Energy (DoE) and subsequently adopted by the Federal Energy Regulatory Commission (FERC) stated as 'changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time' [3].

Recently, the independent system operators (ISOs) make efforts to implement DR programs to liberalize electricity markets and reduce extra generation cost. The performance of DR programs is measured by the amount of power reduction that customers achieve, and thus the measurement and verification (M&V) process is required to verify the amount of load reduction when DR is activated. The advanced metering infrastructure (AMI) is used to collect the electricity consumption data. Smart meters first collect the data from the end users and then relay the data to local data collection units (DCUs) periodically, e.g., 15, 30 or 60 minutes. DCUs in turn relay the data to the utility servers.

To measure the amount of load reduction, we need to know two factors: the customer baseline load (CBL) and the actual load. CBL refers to the amount of electricity the customers would consume if DR were not activated. The actual load is the amount of electricity the customers consume during the DR event period. Then, the amount of load reduction is computed by the difference between CBL and the actual load. The accurate calculation of CBL is crucial to the success of DR programs because, otherwise, it is unclear how much power is reduced, and it is hard to determine the proper monetary reward that should be paid to the customers. For example, if the computed baseline load is too low, the customers are less motivated to participate in the DR program. Hence, the accurate baseline load estimation prediction is required for both the utility company and the customers [4].

Two widely adopted methodologies for baseline load estimation are the day matching method and the regression analysis. The day matching method takes a short historical period and estimates the electricity usage of the DR event day by simply averaging the data during the short period such as 5, 8 or 10 days. For example, the New England ISO uses five equivalent days prior to the DR event day while the California ISO (CAISO) uses three equivalent day. The Korea Power Exchange (KPX) takes the four highest demand days among the five equivalent days. The regression analysis creates a model based on statistical regression methods [5].

Load prediction has been one of the most important research topics. Specifically, there are several works related to short-term load prediction considering the weather. In [6], [7], the relationship of electricity demand and the weather was developed by transfer function where the daily load forecasting is computed from the cooling and heating condition. Another method for load prediction is to apply the artificial neural network (ANN). Typically, 24 neurons are used to forecast the hourly loads of the next 24 hours [8].

However, the above-mentioned methods are mainly about load prediction for the benefit of electricity suppliers. By contrast, the establishment of baseline load is focused on the demand side. The accurate demand prediction can motivate the customers to participate in the DR programs and to receive monetary rewards from the utility company. In [9], CBL calculation using the exponential smoothing model with weather adjustment was proposed. They exponentially weighted the past data with multiplicative adjustment. Recently, the Lawrence Berkeley National Laboratory (LBNL) found that applying the morning adjustment factor can reduce the bias and improve the accuracy of CBL [5].

In this paper, we propose a novel CBL calculation framework using the data mining techniques that exploit the Kohonen networks model and the unsupervised learning algorithm. We summarize our key contributions as follows. First, we build a framework of CBL estimation based on the data mining approach. Historical data are used to analyze the customers

978-1-4799-4934-2/14/$31.00 ©2014 IEEE

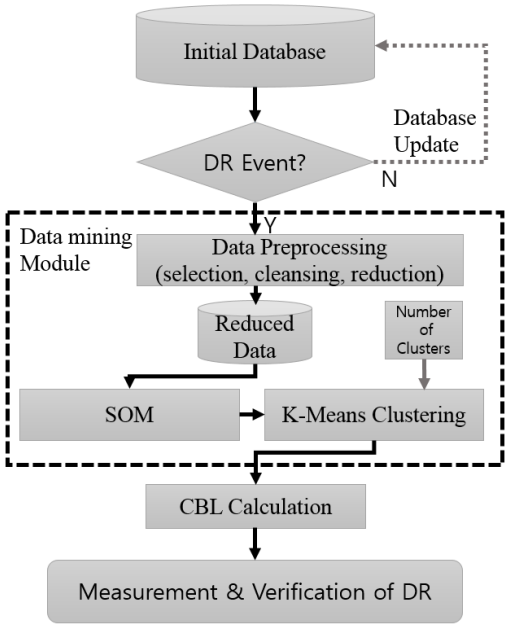

Fig. 1. Framework of the baseline load computation.

electricity consumption behaviors for DR. Second, we conduct the experiments on the real database of the large scale industrial building complex in Korea. This site has an area of 452,647m^2, and the advanced technologies for DR are applied. The number of deployed smart meters is approximately 2,500. Third, comparing the errors for each day indicates that our approach outperforms other algorithms. The results show that our data mining approach results in 15-22% less root mean square error, and 15-20% less mean absolute percentage error, respectively.

This paper is organized as follows. In Section II, we set up the data mining structure for CBL calculation and propose our algorithm. Section III describes a case study using the real measurement data from the commercial complex buildings to validate the proposed algorithm. In Section IV we conclude our work with remarks and future work.

II. Data Mining Based Approach

In this section, we describe the data mining framework to develop the CBL calculation as depicted in Fig. 1. The framework of CBL is based on a knowledge-discovery in databases (KDD) procedure supported by data mining techniques [10].

A. Data preprocessing

The data mining model for CBL calculation is based on the unsupervised learning techniques. Before setting up a model for our work, the data preprocessing needs to be done in advance. This is because some information can be distorted or missing during the period of collecting data and storing it to the utility server. This process includes three phases such as data selection, data cleansing and data reduction.

In data selection, more significant data is selected to process from the initial database. In our case, it is made according to the power level of the consumers. In data cleansing, we check the data inconsistency and find the outliers that distort the information about the customer facilities such as offices, factories or buildings. Inconsistent electricity consumptions are replaced by the average of the daily consumption. In addition, some missing values are also replaced by the daily average or removed at all according to the volume of missing data.

In data reduction, we first classify the load condition in terms of the seasons of the year and the type of weekday to reduce the data size because the electricity consumption depends on the day type, i.e., either week or weekend, and the seasonal weather. The electricity consumption pattern of each day is then characterized by a single curve called *load vector*. The *representative* daily load vector of the day $i \in \{1, \cdots, N\}$ is a vector $l_i = \{l_i(h)\}$ where $l_i(h)$ is the *normalized* values of the instant power consumed in the period of $h \in \{1, \cdots, H\}$ where H is the total number of periods in a day, e.g., $H = 24$.

B. Self-Organizing Map (SOM)

SOM was introduced by Kohonen [11]. SOM is one of the ANN architectures for the unsupervised classification of data into clusters [12]. SOM maps a multi-dimensional input space onto a topology-preserving output space with greatly reduced dimension where neighboring neurons (or units[1]) respond to similar input vectors. Each unit in the output space is assigned a weight vector with the same dimensionality of the input space, but SOM consists of two dimensional grid of map units. Each unit $s \in S$ is a represented by a vector $m_s = \{m_{s1}, \ldots, m_{sd}\}$, where d is the dimension of the input vector. Each unit is connected to the adjacent units by the neighboring relation. SOM is trained iteratively. At each training step, the input vector denoted by x is randomly chosen. The distance between x and all map units are computed. The best matching unit, denoted by b, is the map unit with the closest to x, i.e.,

$$b = \underset{s \in S}{\operatorname{argmin}} \|x - m_s\|. \tag{1}$$

Next, the map units are updated. The best matching unit of (1) and its close neighbors are moved such as

$$m_s(t+1) = m_s(t) + \alpha(t) h_{bs}(t)(x - m_s(t)) \tag{2}$$

where t is the iteration index, $\alpha(t)$ is the learning parameter and $h_{bs}(t)$ is the neighborhood relation function [12]. The magnitude of the adaptation is controlled via the learning parameter that decays over iterations. This process is repeated for all input vectors so that the different grid of map units harmonize with specific domain of the input variable which contains daily load diagram.

C. K-means clustering

K-means clustering partitions a data set into K clusters. This process is done without any prior knowledge about the data structure such as the number of groups, labels, and sample members. A widely adopted definition of optimal clustering is to form a partition that minimizes the distances within the intra-clusters and maximizes the distances among the inter-clusters. Given a set $\mathcal{Z}$ of SOM operation outcomes and an integer number K, the K-means algorithm searches for

[1]We use the neuron and the unit interchangeably.

a partition of $\mathcal{Z}$ into K clusters that minimizes the within groups sum of squared errors (WGSS). Mathematically, is given by [13] such as

$$\text{minimize} \quad P(W,\mathcal{C}) = \sum_{k=1}^{K}\sum_{i=1}^{N} w_{i,k} d^2(Z_i, C_k) \qquad (3)$$

$$\text{subject to} \quad \sum_{k=1}^{K} w_{i,k} = 1, \quad \forall i \in \{1,\ldots,N\}$$
$$w_{i,k} \in \{0,1\}, \quad \forall i \in \{1,\ldots,N\}$$
$$\forall k \in \{1,\ldots,K\}$$

where W is an $N \times K$ partition matrix, $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$ is a set of clusters in the same domain where $C_k, k = 1, \cdots, K$, denotes the centroid point in the cluster k, and $d^2(\cdot,\cdot)$ is the square of the Euclidean distance between two variables. The problem (3) is solved iteratively until no further change occurs in the cluster points.

D. The Proposed Algorithm

In this subsection, we propose an algorithm for calculating CBL. Basically this algorithm finds the most similar load patterns from the historical data and average them by hourly load.

There were many attempts to cluster the large data using SOM [12]. The primary benefit of two-level approach, i.e., first applying SOM and then using K-means clustering is the reduction of the computational cost. SOM efficiently transforms the large dimensional data space into two dimensional space. K-means algorithm is commonly used for clustering the data. However it is hard to handle noisy data and outliers. By combining these two models, we create more accurate solution dealing with large data sets.

We develop a framework that calculates the baseline load using the measured data, which is summarized in Algorithm 1. When DR event occurs, DR event time goes to the input variable. In the procedure, 12-hour load vectors prior to the DR start time for the day and the past days are formed from the database. Note that to improve the accuracy, we deliberately append the average temperature and the specified morning factor into the input vector. Specifically, as shown in Table I, X_1 to X_{12} are the hourly consumption before DR event occurs, X_{13} indicates the average temperature in the morning (8:00 AM-12:00 AM). We set the morning factor (X_{14}) as the average gradient of the electricity consumption curves because the gradient contains a useful information of finding the similar data. X_{15} is the working availability in weekdays.

In the next step, the grid of map units is initialized randomly. This is an $M \times N$ matrix that can represent the input data. Through the iteration steps of finding the best matching units (BMU) and updating the units as in (1) and (2), respectively, the output of SOM operation can represent the historical data in the 2-dimensional space $Z = (m, n) \in \mathcal{Z}$ where $m \in \{1, \cdots, M\}$ and $n \in \{1, \cdots, N\}$ represent the coordinate in $\mathcal{Z}$. After the SOM operation, the input vectors have their own SOM values and are assigned to one of the units in the grid. In other words, each input vector is assigned to the appropriate cluster. By minimizing the WGSS, this procedure produces K clusters as the output as described in (3). In the final step, we can find the best similar patterns of electricity consumption in the past. By averaging hourly load from this outcome, the predicted CBL is obtained.

Algorithm 1 Baseline Load Estimation Algorithm

```
procedure LOADPROFILE
    for ∀i ∈ {1, ··· , N} do
        X_i1, ··· , X_i12 ← 12-hour load
        X_i13 ← average temperature
        X_i14 ← morning factor
        X_i15 ← work availability
    end for
    repeat
        find the BMU of input vector
        update the units of grid
    until
    maximum number of iterations
    repeat
        assign data points to cluster and solve problem (3)
        update cluster centers
    until
    no cluster changes
end procedure
```

TABLE I. INPUT VECTOR FOR SOM OPERATION

$X_1 ... X_{12}$	12 hour consumption before DR activation
X_{13}	Average Temperature
X_{14}	Gradient of the load consumption (optional)
X_{15}	Work Availability (optional)

Note that the proposed algorithm has two key differences with the existing CBL methods. First, unlike the day matching methods, we find the most similar day given partial data. Second, we consider other parameters such as average temperature, the gradient of electricity consumption, work availability. By learning process, SOM output describes the data structure by matrix. Clustering output indicates the representative load of the database.

III. CASE STUDY: LARGE SCALE EXPERIMENTS

To validate the proposed CBL algorithm, we data-mine the data that have been collected by smart-meters in an industrial building in Seoul, Korea. This building is a component of large scale industrial building complex consisting 17 buildings for factory apartment, office, etc. The complex has an area of 452,647m^2 and several key technologies for DR have been applied such as AMI system (about 2,500 devices), BEMS, user feedback service (11 buildings, 1,100 units), parallel emergency generator system and market based electricity pricing simulation. We perform a large scale experiment to verify the proposed algorithm. Our database contains 12 months electricity consumption data. The instant power consumption for each day was collected at intervals of 15 minutes, and we get the hourly load by addition for convenience. Fig. 2 shows the typical consumption pattern of the industrial building monitored for the experiments. We note that this building reaches 1.3 MW peak value between 12:00 PM to 14:00 PM.

A. Data processing

A part of data usually have bad quality in the real database; the database has some problems including wrong information and missing data. In the pre-processing phase, these data

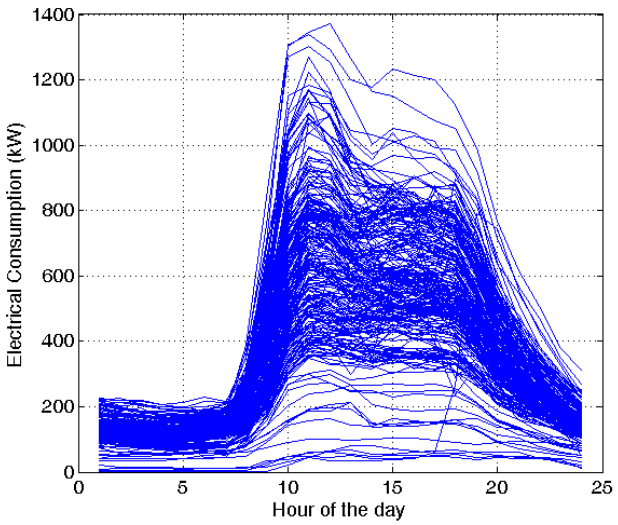

Fig. 2. Typical load curves of the industrial building.

have been corrected. The outlier and outages are detected and replaced by regression technique based on the similar days. With this procedure, the quality of the data has been improved with a minimal loss of information. These data are reduced and normalized using the procedure presented in Section II. After that, other parameters such as temperature, the morning factors, work availability are included into the input vector.

We use a 10×8 neuron architecture in SOM as can be seen in Fig. 3. The size of neurons are determined to maximize the so-called explained variance. Fig. 3 indicates the links among the neurons as well. It is called the unified distance matrix (U-matrix), which is a representation of a self-organizing map. The Euclidean distance between neighboring neurons is described in a gray scale image. U-matrix gives insights into the local distance structure of the high dimensional data set. It has become the standard tool for the display of the distance structures of the input data. For each unit in U-matrix, we assign the dark gray color to the unit that has a large number of close units. Similarly we assign the light gray color to the unit that has a small number of close units. The other shades of gray corresponds to the level of distance between units. The circles inside the unit hexagonal cells indicates how many data points are close to a particular unit. A larger circle means that there are more data representing the unit cell.

B. Determination of the number of clusters

The number of clusters is the input of the K-means clustering algorithm. In the unsupervised learning algorithm, choosing the proper number of group is critical. There are several methods for estimating the appropriate number of clusters. For example, the mean index adequacy (MIA) was proposed in [14]. This can be expressed as

$$\mathrm{MIA} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} d^2(r^{(k)}, C^{(k)})}. \tag{4}$$

We divide N input vectors into K clusters. Each cluster is formed by a subset $C^{(k)}$ of load vector, where $r^{(k)}$ is a representative load vector assigned to cluster k. MIA can

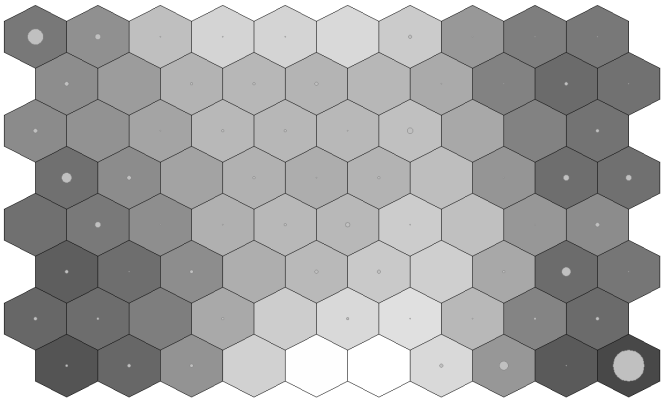

Fig. 3. Unified distance matrix of input load vector.

be obtained by averaging the square Euclidean distance (d^2) between this representative load vector and subset load vector.

The smaller values of MIA indicate more compact clusters, which means more accurate baseline load can be obtained. The K-means clustering algorithm was used to study the cluster tendency of the data based on the MIA measure. Through this process, the number k is obtained and presented in Fig. 4. We propose $k = 14$ is the best choice for clustering. MIA will be decreased when k is large. However, the number of data point in each cluster is also decreased, which makes it hard to build a model. We assume the minimum number of data in each cluster should be 5. As k is increasing, non-qualified clusters also increased in Fig. 4.

C. Numerical Results

In this subsection, we compare our proposed method with the day matching method. It is generally accepted that a period of approximately 10 days reasonably represents consumption for normal operations. Especially, LBNL suggested taking hourly usage in highest 3 out of 10 previous days which is called the baseline load profile 3 (BLP3) method [5]. Another approach is to select high 5 days out of 10 days, i.e., for a given time interval, baseline load is calculated as the average interval demand among the 5 highest energy usage days out of the prior 10 days.

Two criteria commonly used to evaluate the accuracy of load forecasting are the root mean square error (RMSE) and the mean absolute percentage error (MAPE) [15],

$$\mathrm{RMSE_i} = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(\hat{l}_i(h) - l_i(h))^2} \tag{5}$$

$$\mathrm{MAPE_i} = \frac{100}{H}\sum_{h=1}^{H}\frac{|\hat{l}_i(h) - l_i(h)|}{l_i(h)} \tag{6}$$

where H is the number of time intervals in ith day, $l_i(h)$ are real electricity consumption, and $\hat{l}_i(h)$ are estimated CBL.

Table II presents the final distribution of the days into 14 clusters after SOM and K-means clustering (2 clusters are null spaces). Each cluster corresponds to the different patterns of the customers. The representative load profiles are obtained by averaging load vectors in the same cluster and presented in Fig. 5. The created load profiles have distinct load shapes

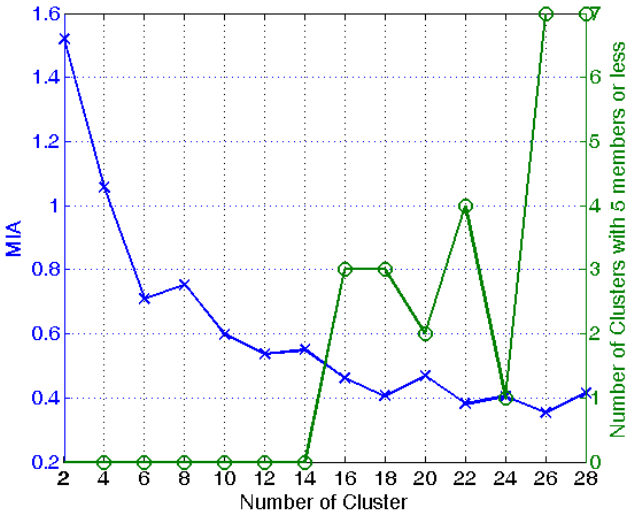

Fig. 4. MIA and non-qualified clusters with the number of clusters.

TABLE II. NUMBER OF DAYS WITHIN EACH CLUSTER

Cluster	0	1	2	3	4	5	6
Working Days	15	16	9	0	46	22	20
Cluster	7	8	9	10	11	12	13
Working Days	11	18	32	0	17	22	15

according to their consumption patterns. For the experiment, we test our method and other existing ones for each working day in a month. We assume the DR event day in summer because of the high peak load. DR events usually occur during the afternoon and the early evening and does not last longer than 6 hours. Most reports states that DR started between 1:00 PM and 2:00 PM in historically [16]. Hence, we assume the DR starts at 1:00 PM.

We calculate the RMSE, MAPE values for each day in August and plot them in Figs. 6-9. In RMSE evaluation, our method has 15% to 22% lower error than the day matching methods in average. In case of MAPE, our method has 15% to 20% lower error than the day matching methods in average. As can be seen, the day matching methods are sometimes better than our proposal in Fig. 6 and 7. This could happen when the customers consumption patterns are highly similar to the past a few days. In addition, the cumulative probability function (CDF) of the error is used for measuring the precision of the CBL. As can be seen in Figs. 8 and 9, we see that the proposed method outperforms the day matching methods.

IV. CONCLUSION AND FUTURE WORK

In this paper we proposed a framework of customer baseline load estimation for demand response. Unlike the conventional techniques such as the day matching methods, we leveraged the data mining techniques, SOM and K-means clustering to find the days that are expected to have the most similar load patterns to the day of DR event. In SOM operation we augmented the load vector with up to three other attributes such as the morning factor, the day type and the average temperature for better classification. Then, we applied K-means clustering to group the days that have

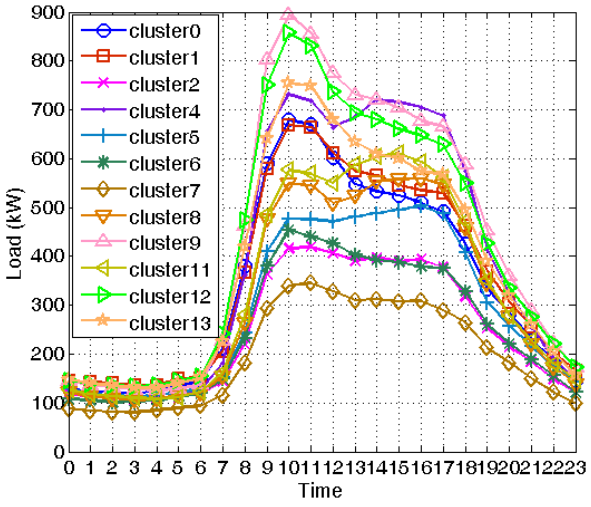

Fig. 5. Load profiles of historical day classes.

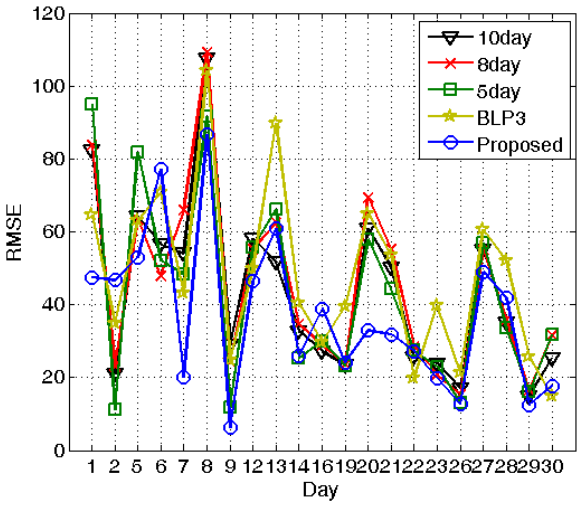

Fig. 6. Performance of the CBL models (RMSE).

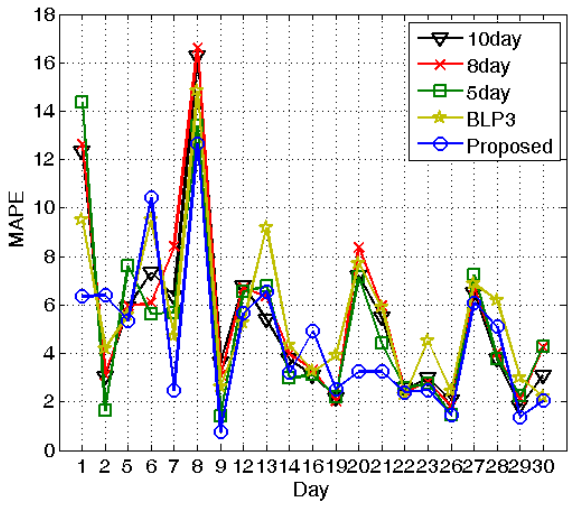

Fig. 7. Performance of the CBL models (MAPE).

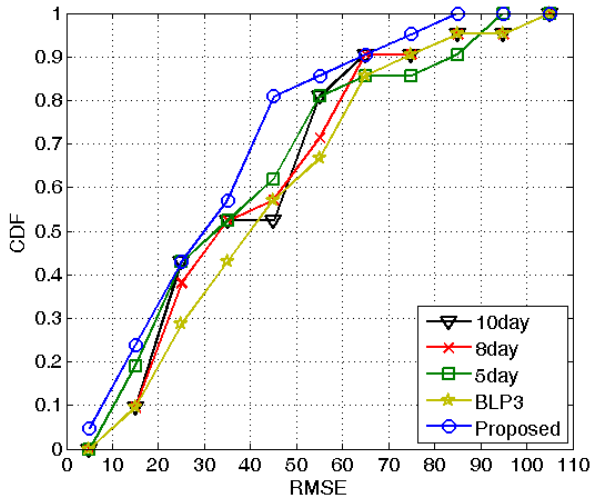

Fig. 8. Cumulative distribution function of RMSE.

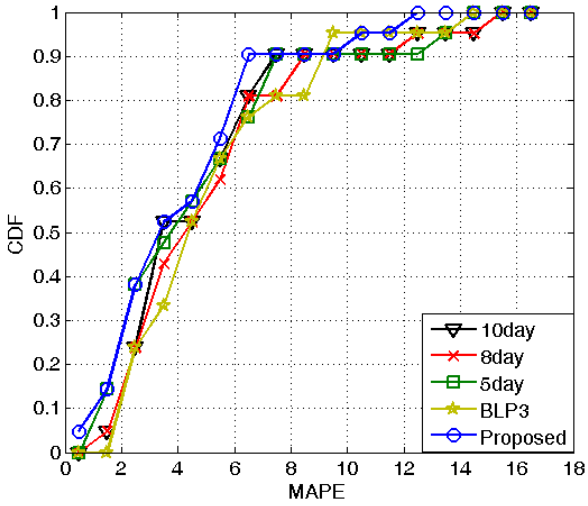

Fig. 9. Cumulative distribution function of MAPE.

similar load patterns. In determining the number of clusters, we exploited both the MIA and the number of small size clusters. To verify the proposed framework we performed large scale experiments where electric power consumptions are collected by 2,500 smart meters. The results show that our framework significantly improves the accuracy of the load estimation. Specifically, compared to the day matching methods, the root mean square error was reduced by 15-22% in average and the mean absolute percentage error was reduced by 15-20% in average as well.

The results of this paper can be further extended in several directions. First, accumulating database can provide more accurate CBL calculation. In this study, we have 12-month load vectors, but more data are expected to be available shortly. Second, different type of adjustments for CBL can reduce the error rate of the prediction. Although the weather based CBL adjustment is complex to verify, the adjustment based approach can be preferable. Third, it would be good to compare our framework with other CBL algorithms such as different learning approaches, the weather exponential smoothing models, etc.

ACKNOWLEDGMENT

The authors appreciate the support of Samsung C&T Corporation K-MEG team lead by Dr. Song S. Han and Dr. Wonyoung Yang. This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency)(NIPA-2013-(H0502-13-1060)).

REFERENCES

[1] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *Industrial Informatics, IEEE Transactions on*, vol. 7, no. 3, pp. 381–388, 2011.

[2] H. Kim, Y.-J. Kim, K. Yang, and M. Thottan, "Cloud-based demand response for smart grid: Architecture and distributed algorithms," in *Smart Grid Communications (SmartGridComm), 2011 IEEE International Conference on*. IEEE, 2011, pp. 398–403.

[3] M. L. Goldberg and G. K. Agnew, "Measurement and verification for demand response," 02 2013.

[4] A. L. R. Committee, "Demand Response Measurement & Verification," 03 2009.

[5] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, "Estimating demand response load impacts: evaluation of baseline load models for non-residential buildings in california," 2008.

[6] A. Pardo, V. Meneu, and E. Valor, "Temperature and seasonality influences on spanish electricity load," *Energy Economics*, vol. 24, no. 1, pp. 55–70, 2002.

[7] E. Valor, V. Meneu, and V. Caselles, "Daily air temperature and electricity load in spain," *Journal of applied Meteorology*, vol. 40, no. 8, pp. 1413–1421, 2001.

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

[9] Y.-M. Wi, J.-H. Kim, S.-K. Joo, J.-B. Park, and J.-C. Oh, "Customer baseline load (CBL) calculation using exponential smoothing model with weather adjustment," in *Transmission & Distribution Conference & Exposition: Asia and Pacific, 2009*. IEEE, 2009, pp. 1–4.

[10] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Advances in knowledge discovery and data mining." Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996, ch. From Data Mining to Knowledge Discovery: An Overview, pp. 1–34. [Online]. Available: http://dl.acm.org/citation.cfm?id=257938.257942

[11] T. Kohonen, *Self-organizing maps*. Springer, 2001, vol. 30.

[12] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *Neural Networks, IEEE Transactions on*, vol. 11, no. 3, pp. 586–600, 2000.

[13] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[14] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *Power Systems, IEEE Transactions on*, vol. 18, no. 1, pp. 381–387, 2003.

[15] G. Li, C.-C. Liu, C. Mattson, and J. Lawarree, "Day-ahead electricity price forecasting in a grid environment," *Power Systems, IEEE Transactions on*, vol. 22, no. 1, pp. 266–274, 2007.

[16] D. Hurley, P. Peterson, and M. Whited, "Demand response as a power system resource," 2013.