Correlation Between Predictability Index and Error Performance in Customer Baseline Load (CBL) Calculation

Saeed Mohajeryami, Roozbeh Karandeh, and Valentina Cecchi
Department of Electrical and Computer Engineering
University of North Carolina at Charlotte
Charlotte, USA
Email: {smohajer, rkarande, vcecchi}@uncc.edu

***Abstract*—This paper explores the correlation between the low-frequency content of customers' consumption load, measured by a proposed index called predictability index, and the accuracy of Customer Baseline Load (CBL) estimation methods. The customer's consumption signal is transformed from time-domain to frequency-domain to separate its high- and low-frequency components. After reconstructing the time-domain equivalent of both high- and low-frequency signals, the predictability index for all customers is calculated. This proposed index is employed for the purpose of clustering the customers into different bins by a k-means clustering algorithm. The CBL for customers of each bin is estimated by two methods: CAISO and Randomized Controlled Trial (RCT). The data employed in this study belongs to Australian Energy Market Operation (AEMO), and are hourly consumption of 189 customers for the time span of a year (2012). In this paper, the correlation between average predictability index of each bin and its normalized average error is calculated. It is found that there is a strong correlation between the predictability index and the error performance of the CBL estimation methods.**

Index Terms*—Customer Baseline Load (CBL); Demand Response (DR) programs; Discrete Fourier Transform (DFT); k-means clustering; Predictability index; CAISO, Randomized Controlled Trial (RCT)

I. Introduction

In the current state of the electricity market, in a day-ahead market, the power plants offer their electricity generation as a supply and utilities bid for it. Utilities in this market can be regarded as the demand side. As a matter of fact, they are agents who purchase electricity in the wholesale market with real-time pricing and sell it to retail customers with a fixed tariff. Therefore, actual customers are shielded from the price fluctuations of the wholesale market. This isolation can significantly affect the efficiency of the market [1]. In facts, the efficiency of the free market depends on the elasticity of the supply and demand [2], [3]. In the current wholesale market, the demand is somewhat inelastic to the sudden changes in the supply side, which can cause a lot of troubles including sharp price spikes during supply shortages. For instance, many economists believe that the inelasticity of the demand side was one of the major contributing factors of the California energy crisis [4].

One of the proposed solutions to the aforesaid problem is the adoption of Demand Response (DR) programs. There are numerous papers in the literature describing merits of these programs. They are intended to break the isolation between the supply side (generation companies) and the demand side (residential, commercial and industrial customers). Although there is a consensus about the merits of these programs, there is still debate about the implementation of such programs, chief among them is Evaluation, Measurement, and Verification (EM&V) of the customers' load reduction. In order to perform the payment settlement, which is a critical part of all DR programs, it is essential that the load reduction is accurately measured and verified. In order to achieve this, Customer Baseline Load (CBL) must be reliably estimated. The CBL is the amount of electricity that customers would have consumed in the absence of the DR curtailment call. If the CBL is calculated accurately, then the real load reduction could be measured as a difference between actual consumption and the CBL.

In recent years, the technological infrastructure on the distribution system side has become mature enough to allow DR programs to offer their services to residential customers. For example, high penetration of smart meters at the residential level has provided high-quality, high-resolution consumption data [5]. The availability of the residential data has created unprecedented opportunities for load aggregators to offer DR programs to customers. Moreover, it has provided an opportunity for researchers to examine the strength of existing CBL methods for customers.

The authors in [6], examined the different CBL estimation methods to determine which CBL is more effective for residential customers. This work also has examined the performance of conventional estimation methods for residential customers. They found that the CBL calculation is more challenging for these customers compared to industrial and commercial customers. The authors in [7], [8] demonstrated that the historically-used CBL calculation methods, devised according to the nature of large industrial and commercial customers, are incapable of estimating an accurate residential customers' CBL. Moreover, the authors in [9]–[12] went beyond the accuracy of CBL calculation methods and elaborated how the application of CBL, in their current form, can create an incentive for gaming. Therefore, an ideal CBL estimation method must be able to provide an acceptable accuracy and eliminate the incentives for gaming.

In order to improve the CBL accuracy for residential customers, it is critical to study the nature of residential customers' consumption habits. One way to study the nature of the consumption patterns is to treat the consumption load as a time-domain signal and employ signal processing techniques. In this paper, the residential customers' hourly consumption data is decomposed into its underlying frequency-domain counterparts and divided into two high- and low-frequency components. These two components are then transferred back into the time-domain to create the time-domain counterparts of the high- and low-frequency components.

Moreover, it is demonstrated that the share of the high-frequency components of the consumption signal correlates with the accuracy of CBL calculation methods. In other words, if a researcher finds a way to decrease the share of the high-frequency components of a consumption signal, it is possible to claim that the method is able to increase the accuracy of the CBL estimation. In order to investigate this issue, in this paper, an index is proposed, called predictability index (P_{index}), which is designed to show the share of the predictable part of a signal. The index tries to reflect the share of low- frequency components of a signal, which is assumed to be easier to predict. Then, it is utilized to cluster customers into different bins. Next, the CBL for customers of each bin is calculated as well as the normalized average accuracy of each bin. Afterward, the correlation between the average P_{index} of each bin, and its normalized average is calculated. If the correlation between the P_{index} and the accuracy metric value is strong, it indicates that the proposed P_{index} could be applied for EM&V of the CBL estimation methods. The index reveals a lot of information about the signal, and also is a good indicator of the error performance of CBL estimation methods.

What distinguishes this paper from previous literature is that it proposes a new approach for EM&V of the proposed estimation methods in the literature. Moreover, an index is proposed that can be employed as a feature for customer clustering, which could be very instrumental in improving the accuracy error of CBL estimation methods. As mentioned before, the index reveals a lot of information about the consumption load and indicates how well an estimation method can perform.

The rest of the paper is organized as follows. The steps required for decomposing the consumption signal into its underlying high- and low-frequency components in the time-domain are introduced and explained in section II. The proposed predictability index is introduced and defined in section III. In section IV, *k*-means clustering is introduced and explained how it is implemented technique in this work. Section V introduces two CBL calculation methods of California ISO (CAISO) and RCT. Moreover, in this section, the dataset employed in this paper and the accuracy metric of Mean Absolute Error (MAE) are described. The results of the correlation analysis and discussion about the results are presented in section VI. Finally, the paper concludes in section VII.

II. Frequency-Domain Analysis

In this section, the elements required for the frequency-domain analysis are elaborated. The steps required for decomposing a signal into its underlying components is illustrated in Fig.1.

A. DFT Introduction

The discrete Fourier transform is a tool to transform a finite sequence of data (*N* samples separated by a sampling period) into coefficients of sinusoids, which are ordered by a complex-valued function of frequencies. In other words, the DFT could be regarded as a frequency-domain representative of the original time-domain input sequence. After decomposing a time-domain signal into its underlying components, then it is possible to search for any recurring pattern or periodicity in the original time-domain data. Moreover, DFT determines the magnitude of each periodic component, thereby showing the relative strength of the components. In this section, hourly electricity consumption of each customer for the course of one year is treated as a time signal, and by utilizing the DFT and two filters, they are divided into high- and low-frequency signals. The equations for DFT and inverse DFT are shown in (1) and (2).

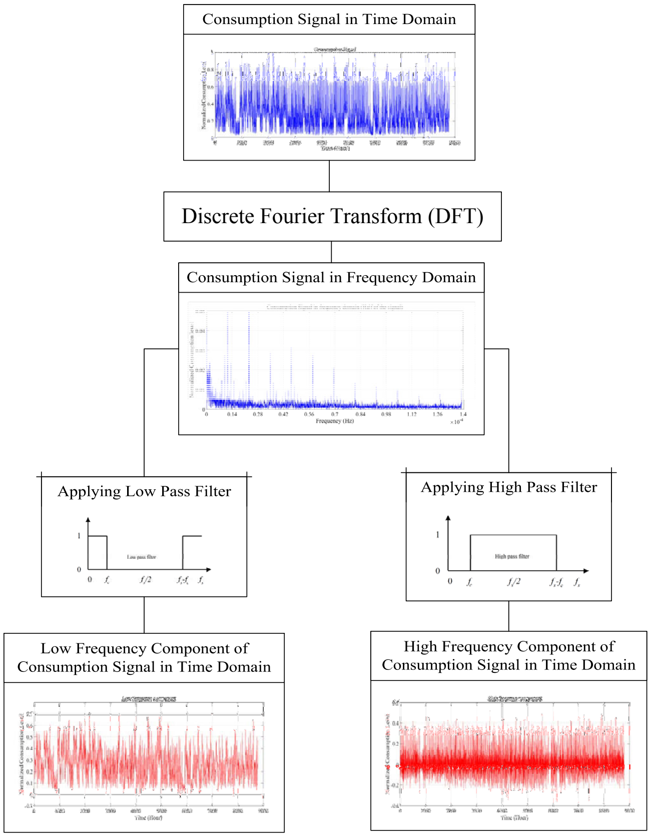
Consumption Signal in Time Domain
Discrete Fourier Transform (DFT)
Consumption Signal in Frequency Domain
Frequency (Hz)
Applying Low Pass Filter
Low pass filter
1
0
0
f_c
$f_s/2$
f_s-f_c
f_s
Applying High Pass Filter
High pass filter
1
0
0
f_c
$f_s/2$
f_s-f_c
f_s
Low Frequency Component of Consumption Signal in Time Domain
High Frequency Component of Consumption Signal in Time Domain
Time (Hour)
Time (Hour)

Fig. 1: Implementation of DFT and separating time-domain high and low frequency components

$$X_k = \sum_{t=0}^{N-1} s[t].e^{\frac{-i2\pi kt}{N}}, k = 0, 1, ..., N-1 \qquad (1)$$

$$X[t] = \frac{1}{N}\sum_{t=0}^{N-1} X_k.e^{\frac{+i2\pi kt}{N}}, t = 0, 1, ..., N-1 \qquad (2)$$

where *N* is the number of samples in the signal. The outcome of DFT is another signal with *N* components in which each component has a different frequency, and these frequencies are listed in monotonically increasing order. An important note about the outcome of DFT is that the real parts in the outcome signal are mirrored over half of the data points. Therefore, only the information of half the signal is relevant; the other half is a repetition of the first. The previous sentence is mathematically expressed in (3).

$$X_k = X^*_{N-k} \quad 1 < k < N-1 \qquad (3)$$

where operator (*) refers to the conjugation operator. In the DFT output signal, the frequency resolution can be calculated as (4).

$$f_r = \frac{1}{NT_s} \qquad (4)$$

where f_r refers to the frequency resolution and T_s refers to the time resolution in seconds. In this paper, the time resolution is 3600 seconds (1 hour) and the frequency resolution, given the 8784 sample points, is 31.6 nHz.

B. Filters

For the purpose of decomposing the consumption signal into its underlying components, two filters of high and low pass frequency are created around a cut-off frequency (f_c). These two filters separate a frequency-domain signal into high- and low-frequency components [13]. The cut-off frequency is a frequency that is assumed to be able to separate the predictable and unpredictable components of the consumption signal. Fig. 2 illustrates the filters. f_c is the cut-off frequency, and f_s is the frequency of the sampling period, which is 277.7 μHz in this study.

The cut-off frequency is different for customers in different sectors; however, in this paper, the cut-off frequency of 23.1 μHz (the equivalent of 12 hours in the time domain) is selected. The rationale behind selecting 12 hours is that frequency of almost all the spontaneous day-time activities' are under 12 hours.

C. Reconstruction

By applying the filters to the frequency-domain consumption signal, two high and low frequency signals could be obtained. Then, an inverse DFT can be applied to these frequency-domain signals to reconstruct their high and low frequency counterparts in the time-domain.

III. Predictability Index

In this section, an index for predictability is introduced. Afterward, the index is calculated for all customers in different clustering bins, and then, the results are presented. This index subtracts the share of high frequency signal from the whole signal as shown in (5).

$$P_{index} = 1 - \frac{\sum_{i=1}^{N} abs(c_i^{hf})}{\sum_{i=1}^{N} c_i} \qquad (5)$$

Where c_i is the consumption value of *i*-th sample point in the original consumption signal in the time-domain, and c_i^{hf} refers to *i*-th data point in the high-frequency component of the original signal in the time-domain. It is worth mentioning that since the P_{index} uses an absolute value of the data points in the high-frequency components, this index can be negative.

The rationale behind calling this index a predictability index is as follows. A signal, to be predictable, needs to follow a pattern and can not be fully random. Random elements in signals can not be forecasted and predicted, and most of the random activities of customers have very high frequency. By removing the share of those high-frequency components, the rest of the signal can be considered as easy to predict. Therefore, since this index reflects the share of the predictable part of the signal, it is called predictability index.

IV. Clustering

In this section, the customers are clustered based on their P_index values and event day average hourly consumption. In order to cluster the customers, *k*-means clustering algorithm is utilized.

A. k-means clustering

k-means clustering has originated from signal processing and has been extensively employed for creating different clusters in data analytics. It partitions the raw data into *k* clusters in which all observations are assigned to a specific cluster based on their proximity to mean of the cluster. In fact, *k*-means clustering is a tool for finding similar groups in a dataset [14].

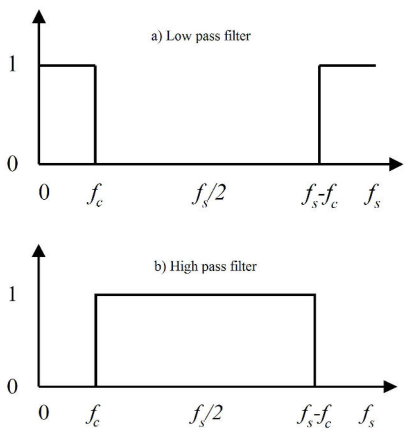

Fig. 2: (a) Low pass filter (b) High pass filter

In order to run a k-means algorithm, a few initial points must be randomly assigned. These points are called cluster centroids. The number of these points is typically determined by a criterion called elbow criterion. k-means is an iterative algorithm, and its two major functions are: 1) assigning a cluster, and 2) moving centroid to minimize an objective function.

Given a set of observations $(x_1, x_2, ..., x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k $(\preceq n)$ clusters $\mathbf{S} = \{s_1, s_2, ..., s_k\}$ so as to minimize the within-cluster sum of squares (WSS) (sum of distance functions of each point in the cluster to the k center). In other words, its objective is to solve the optimization relation of the equation (6).

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{6}$$

where μ_i is the mean of points in S_i.

B. Implementation

In this paper, by using the "elbow criterion" illustrated in Fig. 3, five cluster bins (k=5) are selected. As is shown in the figure, k=5 is a knee point at which the line slope, which is equivalent to the change in the value WSS/TSS ratio (TSS is the total sum of squares), starts to decrease.

Although there are many other ways to cluster the customers [2], k-means clustering is proven to be more efficient. The results of the clustering are shown in Fig. 4. The customers of each bin are shown with different signs. Customers in bin 1-5 are shown with magenta, green, red, blue, and black dots, respectively. Fig. 5 illustrates the load density of each bin. As is shown in the figure, as the P_{index} increases, the distribution becomes more widespread due to higher hourly average.

V. CBL Calculation Methods

In this section, two methods of CBL calculation methods are described. Moreover, the data set employed by this paper will be introduced. Finally, error metric of accuracy Mean Absolute Error (MAE) are introduced and explained.

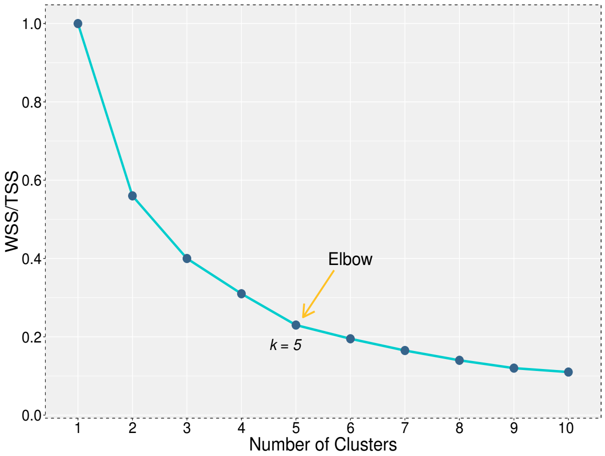

Fig. 3: Elbow criterion analysis

A. CAISO CBL Method

The CAISO CBL method is from the family of averaging methods. Averaging methods are the most popular methods for estimating CBLs, especially HighXofY methods. In this paper, CAISO (High10of10, a.k.a Last10of10) is employed for the purpose of CBL estimation. As discussed earlier, CBL is the amount of load that is estimated to be consumed by customers in the absence of a DR curtailment signal. In this method, Y days of non-event, non-holiday weekdays and weekends prior to a DR-event day are selected. Then, X days with maximum average consumption are selected out of the Y days. The baseline is defined for each hour of the event day as the average hourly load of these X days. The California ISO uses this method with X=10 and Y=10. In this paper, this method is selected with one modification. Since residential customers' consumption on weekends are observed to be similar to weekdays, the weekends are included in the process of the CBL estimation.

B. Randomized Controlled Trial (RCT)

RCT method is very popular and trustworthy as an evaluation method to the extent that, as mentioned earlier, many scholars regard it as the gold standard of evaluation methods. However, this method has some issues that can plague its performance [16]. Nevertheless, the discussion about these issues is beyond the scope of this paper.

The RCT method starts by random assignment of customers into two groups: control and treatment groups. The control group would serve as a basis for the calculation of the CBL for the treatment group [15], [16].

C. Dataset

The data used in this study belongs to Australian Energy Market Operation (AEMO) [17]. It is collected for 200 residential customers. However, in this paper, due to an issue concerning missing data, only consumption data of 189 customers are employed. The data is collected in the leap year of 2012 (366 days). Furthermore, in this paper, Jan. 30 is selected as an event day. This day is selected as an event day because it has the highest consumption in the month of January.

D. Error Metrics

In this paper, the accuracy is employed as a means to evaluate error performance. The metric used for the accuracy evaluation is Mean Absolute Error (MAE), which represents the hourly difference between the estimated CBL and the actual consumption.

Let C be the set of all 189 customers, D be the set of all days in the data set, and T be the set of hourly time-slots in a day. MAE for measuring baseline accuracy is defined as shown in (7). From this equation, it is understood that the lower the MAE, the higher the accuracy.

$$\alpha = \frac{\sum_{i \in C} \sum_{d \in D} \sum_{t \in T} |b_i(d,t) - l_i(d,t)|}{|C| \,.\, |D| \,.\, |T|} \tag{7}$$

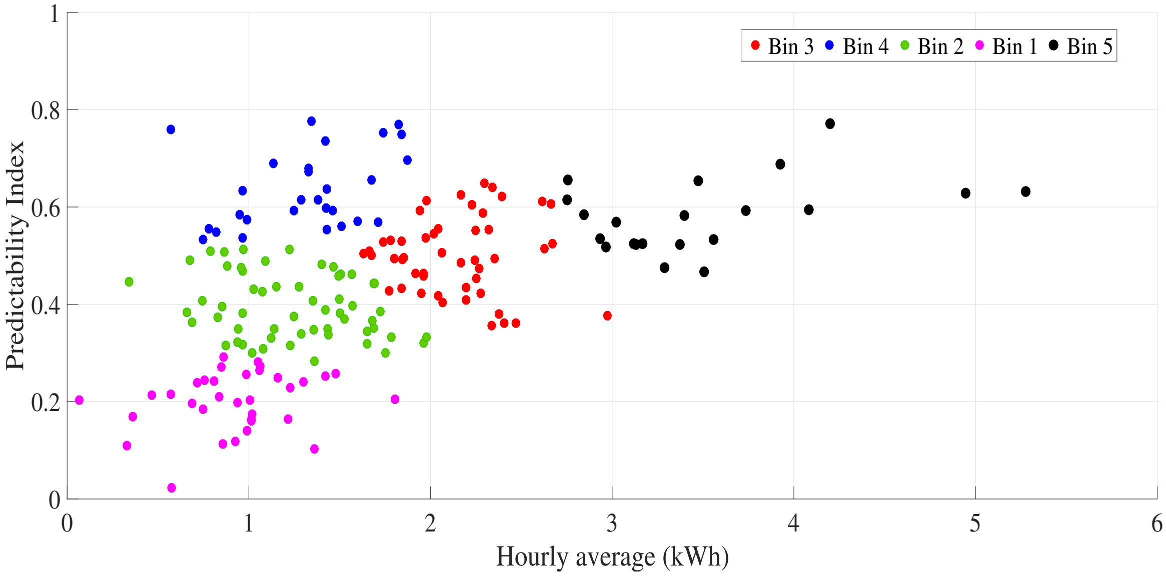

Fig. 4: The relationship between average consumption level and predictability index (different clustering bins are coded with different colors)

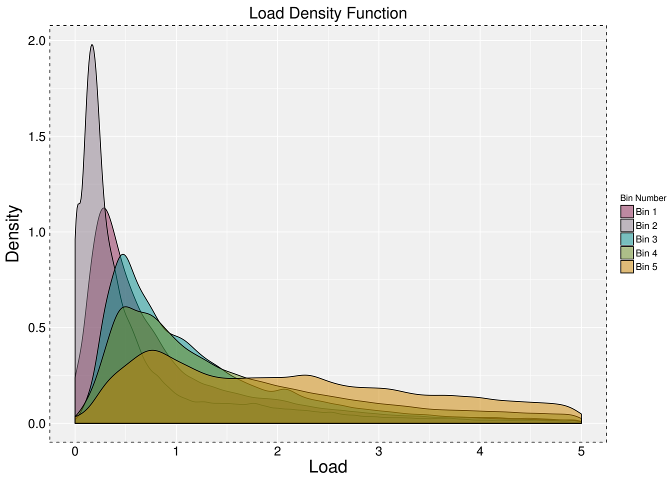

Fig. 5: Load density function in different bins

where b_i refers to the estimated baseline, and l_i is the actual consumption values. Many papers in this field calculate the accuracy just for event hours. However, since there is no DR event in the dataset, in this paper, the accuracy is calculated for the entire event day.

VI. Results and Discussion

In this section, the P_{index} is calculated for all customers. Then, the customers are clustered by k-means clustering into five bins. Later, the CBL for all customers in each bin is calculated, and the average accuracy MAE for each bin is computed. The information for the number of customers in each bin, the average P_{index} of each bin, the average value of the calculated CBLs' MAE for each bin, and the average hourly consumption in each bin are provided in columns 1-5 in Tables I and II.

Moreover, to show the relationship between the P_{index} and the accuracy, the MAE values should be normalized by the value of event-day average hourly consumption of each bin. As is shown in column 5, the average hourly consumption values are different in each bin; For this reason, just comparing the average MAE values of each bin (column 4), without considering the difference in the average hourly consumption is misleading. Therefore, for each bin, the MAE value in column 4 is divided by the event-day average hourly consumption value in column 5, and the normalized values are listed in the 6^{th} column in the above-mentioned Tables.

As is shown in Table I, the normalized values of MAE (last column) are decreasing as the values for average P_{index} are

TABLE I: THE ACCURACY RESULTS OF THE CAISO CBL CALCULATION METHOD

Bin#	#Cust.	Average P_index	Accuracy MAE	Event-day Average Consum.	MAE/Average
Bin 1	35	0.19	0.53	0.80	0.66
Bin 2	57	0.39	0.61	1.16	0.52
Bin 3	48	0.50	0.91	1.94	0.47
Bin 4	28	0.64	0.64	1.51	0.42
Bin 5	21	0.58	1.30	3.26	0.40
Correlation between P_index and MAE/Average					-0.98

TABLE II: THE ACCURACY RESULTS OF THE RCT CBL CALCULATION METHOD

Bin#	#Cust.	Average P_index	Accuracy MAE	Event-day Average Consum.	MAE/Average
Bin 1	35	0.19	0.75	0.80	0.94
Bin 2	57	0.39	0.87	1.16	0.74
Bin 3	48	0.50	1.16	1.94	0.60
Bin 4	28	0.64	1.00	1.51	0.66
Bin 5	21	0.58	2.18	3.26	0.67
Correlation between P_index and MAE/Average					-0.88

increasing. In other words, the higher the P_{index} value, the better the expected performance of CBL estimation methods. It is shown in this Table that the correlation between the P_{index} and the normalized MAE is -0.98. It demonstrates that there is a strong correlation between P_{index} value and the accuracy MAE of the CAISO CBL calculation method.

The same analysis is performed with the RCT method, and it is found that the correlation between the P_{index} and the normalized MAE is -0.88, which is still a fairly strong correlation.

Almost all papers on the subject of EM&V of CBL estimation methods use one form of accuracy metrics to compare different CBL calculation methods with together. Based on the results, the strong correlation between the P_{index} value and the metric of accuracy (i.e. MAE) indicates that the P_{index} could be utilized as an alternative or complement metric to the MAE. Moreover, the P_{index} could be used as a feature to demonstrate limitations of CBL calculation methods. If the P_{index} of a customer is low, no CBL calculation method can accurately estimate the CBL. On the other hand, if the P_{index} value is high, but a CBL method does not give a satisfactory performance, then it is probable that the CBL method has a problem, and a modification in the CBL calculation method may prove to be an effective way to improve the error performance.

VII. Conclusion

This paper examines the correlation between the content of high frequency component of a consumption signal and the accuracy of Customer Baseline Load (CBL) estimation methods. In order to carry out this study, the customer's consumption signal is transformed from time-domain to frequency-domain to separate its high and low frequency components. After reconstructing the time-domain equivalent of both high- and low-frequency signals, a proposed index called the predictability index is calculated for all customers. This index is used to cluster the customers into different bins. Then, the CBL for customers of each bin is calculated, and an average MAE (an accuracy metric) is computed for each bin. The correlation between the average P_{index} value of each bin and its normalized average MAE is calculated.

The key conclusions of this paper are the following:

- The P_{index} has a strong correlation with the accuracy of CBL calculation methods;
- The P_{index} could be employed as a feature for clustering purposes;
- The P_{index} could be utilized as a tool to determine the ceiling of the error performance of CBL estimation methods.

In future work, the authors plan to utilize more sophisticated methods of clustering to assign customers in different bins. Moreover, the correlation analysis in this paper is carried out for just one event day. The conclusion will be much more general if it is tested for multiple months to remove specific monthly characteristics. Also, other CBL calculation methods can be investigated to demonstrate the generality of the conclusion.

Acknowledgement

The authors would like to express their gratitude to Mr. Milad Doostan for kindly assisting with figures and providing comments to improve the manuscript.

References

[1] S. Mohagheghi, F. Yang and B. Falahati, "Impact of demand response on distribution system reliability," in 2011 IEEE Power and Energy Society General Meeting, San Diego, CA, 2011, pp. 1-7.

[2] A. Asadinejad, MG. Varzaneh , K. Tomsovic, C. Chen , R. Sawhney, Residential customers elasticity estimation and clustering based on their contribution at incentive based demand response, In IEEE Power and Energy Society (PES) General Meeting, July, 2016.

[3] A. Asadinejad, K. Tomsovic, and C. Chen, "Sensitivity of incentive based demand response program to residential customer elasticity," 2016 North American Power Symposium (NAPS), Denver, CO, 2016, pp. 1-6.

[4] A. Faruqui, and S. George, "Quantifying customer response to dynamic pricing." in *The Electricity Journal*, Vol 18, No. 4, pp. 53-63, 2005.

[5] R.J. Hamidi, S.H. Hosseinian, S.H.H. Sadeghi, and Z. Qu, "A Novel Approach to Utilize PLC to Detect Corroded and Eroded Segments of Power Transmission Lines," in *IEEE Transactions on Power Delivery*, vol. 30, no. 2, pp. 746-754, April 2015.

[6] T.K. Wijaya, M. Vasirani, and K. Aberer, When bias matters: An economic assessment of demand response baselines for residential customers, in *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1755 1763, 2014.

[7] S. Mohajeryami, M. Doostan, P. Schwarz, "The impact of Customer Baseline Load (CBL) calculation methods on Peak Time Rebate program offered to residential customers," in *Electric Power Systems Research (EPSR)*, Volume 137, August 2016

[8] S. Mohajeryami, M. Doostan , A. Asadinejad, and P. Schwarz, "Error Analysis of Customer Baseline Load (CBL) Calculation Methods for Residential Customers", in *IEEE Transactions on Industry Application*, 2016

[9] H.P. Chao and M. DePillis, "Incentive effects of paying demand response in wholesale electricity markets," in *Journal of Regulatory Economics*, Vol.43, No.3, pp. 265-283. 2013

[10] H. Pourbabak, T. Chen, W. Su, Consensus-based Distributed Control for Economic Operation of Distribution Grid with Multiple Consumers and Prosumers, in 2016 IEEE PES General Meeting Conference & Exposition, Boston, MA, 2016

[11] B. Falahati, Y. Fu and L. Wu, "Reliability Assessment of Smart Grid Considering Direct Cyber-Power Interdependencies," in *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1515-1524, Sept. 2012.

[12] B. Falahati and Y. Fu, "Reliability Assessment of Smart Grids Considering Indirect Cyber-Power Interdependencies," in *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1677-1685, July 2014.

[13] I.N. Moghaddam, B. Chowdhury, "Optimal Sizing of Hybrid Energy Storage Systems to Mitigate Wind Power Fluctuations," in 2016 IEEE PES General Meeting Conference & Exposition, Boston, MA, 2016

[14] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic." in *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol.63, No.2 2001

[15] "Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols", Techincal Report in *Electric Power Research Institute (EPRI)*, ID1020855, Apr. 2010

[16] P. Cappers, "Quantifying the Impacts of Timebased Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies: Protocols and Guidelines." in LBNL Paper LBNL-6301E, 2014

[17] Australian Energy Market Operation (AEMO) load profile data, available (online) at: http://www.aemo.com.au/Electricity/Data/Metering