A Time-Variant Load Model Based on Smart Meter Data Mining

Xiaochen Zhang, Santiago Grijalva, Matthew J. Reno
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, USA
xzhang322@gatech.edu; sgrijalva@ece.gatech.edu; matthew.reno@gatech.edu

***Abstract*— This paper proposes a novel time-variant load model based on data-mining of a historical smart meter database. As part of the ongoing smart grid transformation, smart meters have been widely installed producing massive amount of data and information yet unexplored. One of the critical needs for distribution system operations and planning applications is enhanced modeling of the load, in particular, its dependence on the voltage. Under the typical smart meter recording resolution (15-minutes), the load's P-V and Q-V properties are buried in the spontaneous load changes caused by random customer behaviors. To overcome this, the concept of load condition is introduced and data mining techniques such as Kullback-Leibler divergence and K-subspace are implemented. The proposed model is tested on a large database for the Georgia Tech campus, and the results demonstrate that the new model captures the time-variant property of the load on the building level without additional investment.**

***Index Terms*—Load modeling; data mining; databases; load management; parameter estimation**

I. INTRODUCTION

As one of the essential elements toward the future smart grid, smart meters have been widely installed in the developed world. It is the first time that utilities and system planners have access to measurements for customers at the building level with great time resolution. The massive historical database created by smart meters contains a wealth of information which has not been fully explored or exploited. One of the critical needs for enhanced distribution system operations and planning is a better load model. This paper proposes a new possibility of building a time-variant load model by implementing data mining techniques on smart meter historical database.

From a mathematic point of view, a load model is a formula of the relationship between bus voltage and power (real and reactive) [1]. Compared with the modeling of generators and the transmission system that have been studied in detailed, an accurate time-variant load model has been difficult to achieve due to the uncertainty of the load and the limitation of data available.

There are two popular approaches to establish a load model: measurement-based approach [2,3] and component-based approach [4,5]. The measurement-based approach determines the load model by recording the load responses directly through system voltage stage tests and actual system transients. Although accurate, the measurement-based approach is costly and unable to capture the time-variant properties of the load. The component-based approach estimates the system load's P-V and Q-V properties by aggregating typical load components to represent the ratio of each type of load in the system. This approach avoids costly system tests by taking surveys and building load profiles, but the accuracy of the approach strongly depends on the accuracy of the load components ratios and the specific models of the typical components.

The main contribution of this paper is exploring an alternative for load modeling using the historical data collected by the widely installed smart meters. The time-variant load model proposed in this paper has a wide range of applications. The most immediate application is conservation voltage reduction (CVR) [6], where the total energy consumption is controlled through voltage regulation according to loads' P-V and Q-V properties. Moreover, the dynamic nature of the new model provides valuable information for load forecasting and management; since the energy consumption pattern and power factor are collected for each customer, the new model allows a more flexible and detailed tariff strategy. Finally, when combined with the distribution network and GIS database, more accurate power flow analysis can be achieved by using the proposed voltage-dependent load model for both scenario study and time series simulation.

Generally, there are two major barriers for a data mining based approach. First, the load reading resolution for current smart meters ranges from 15 minutes to 1 hour. Data collected on such a resolution level cannot distinguish the effects from instantaneous load changes and from system voltage deviations, both of which are responsible for changes of real power and reactive power consumption of the load. Based on previous research by EPRI, this resolution of data is not enough to determine the load composition in detail to allow immediate implementation of the component-based load

The author acknowledges the support of the Power Systems Engineering Research Center (PSERC) under Project PSERC S49: Exploiting Emerging Data for Enhanced Load Modeling

model [7]. As a result, instead of seeking the decomposition of the load into specific load components, this paper introduces the concept of load condition and focuses on modeling the P-V and Q-V properties of the load as a whole. Second, the historical smart meter readings from the massive database need to be clustered, in order to prepare data for meaningful and high-quality load model parameter identification. As a result, multiple data mining techniques such as Kullback-Leibler divergence and K-subspace method are implemented to facilitate the modeling process.

The remainder of this paper is structured as follows: In section 2, the smart meter database on the Georgia Tech campus is introduced, and a novel time-variant load model is proposed; in section 3, data mining techniques are introduced, and load modeling process are explained in detail; in section 4, the new modeling method is implemented and evaluated on real data from the Georgia Tech campus. Section 5 concludes the paper and point out possible future research directions.

II. THE TIME-VARIANT LOAD MODEL

A. Smart Meter Data Collection

The data used in the study of this paper comes from a historical smart meter reading database collected by smart meters installed on the Georgia Tech campus. In order to enhance monitoring and reliability of the campus power network, smart meters were widely installed on Georgia Tech campus starting from in 2011. Currently, there are over 400 smart meters installed on the campus, covering each of the 200 buildings. Similar to most of the smart meters in the world, the data are recorded every 15 minutes including measurements of: real and reactive power (P, Q), power factor, voltage (V), and current for each phase. To illustrate the new modeling approach, different buildings are selected in this study, covering various load types such as commercial, residential and industrial loads.

B. ZIP Model

Traditionally, the voltage dependency of loads is expressed by exponential or polynomial models with constant coefficients. In this paper a time-variant load model is developed based on the traditional ZIP model [8], which is shown in (1) and (2).

$$P = P_0[p_1V^2 + p_2V + p_3] \quad (1)$$

$$Q = Q_0[q_1V^2 + q_2V + q_3] \quad (2)$$

Where, P and Q stands for the active and reactive power of the load, and $V=V'/V_0$ is the per unit voltage or the ratio between voltage V' and its nominal value V_0; P_0 and Q_0 are active and reactive power of the load at nominal voltage; In ZIP model, p_i and q_i represent the proportions of the corresponding components, which satisfy $\sum p_i = \sum q_i = 1$.

C. Load Condition Assumption

For the purpose of this paper, a *load composition* is the state of the total aggregate load, including total real/reactive power and the precise connected individual loads that represent this aggregate load value. Technically, each load composition can be modeled by a set of parameters using (1) and (2). However, due to the number of individual loads in a large building, the number of possible load compositions of which devices are connected is significantly large. In practice, a fixed and rigid ZIP model for a building is not accurate enough to model the dynamically changing nature of the load, because the load composition changes over seasons of the year, days of the week, and hours of the day.

In traditional measurement-based load modeling, data is collected at a very high frequency (1000Hz) before and after the voltage deviation [9]. As a result, the load composition is assumed to be fixed, and only voltage is responsible for the load's real and reactive power changes. However, for most smart meter databases, the data is logged at a resolution of several minutes or hourly, and load composition is subject to changes between different readings. In other words, voltage is no longer the only factor that influences the power consumption of the load. In this paper, an assumption about the load condition is made to justify that it is possible to filter out the instantaneous load changes, and build the P-V and Q-V model through data mining techniques.

To begin with, by definition, the number of load compositions is 2^n, where n is the number of appliances. Every smart meter reading for the load is measured under one of those load compositions. In this paper, load condition is defined as a group of load compositions sharing the similar P-V and Q-V properties. As a result, the smart meter readings can be clustered accordingly into several load conditions.

The energy consumption of customers can be separated into random behaviors (such as turning on a light or making a cup of tea) and routine behaviors (such as eating breakfast in the morning or turning on the heater in winter). It is assumed that routine-behavior loads are usually the dominant factor in energy consumption and are strongly correlated to time, such as seasons and working hours. In contrast, random-behavior loads can be interpreted as additional small loads on top of the energy consumption of routine behaviors. Compared with routine behaviors, random behaviors change more frequently and are responsible for the frequent instantaneous load changes.

Under this assumption, all load compositions within a load condition are considered to be different random behaviors on top of the same routine behavior. As the result, data mining techniques can be implemented to identify different load conditions by clustering all smart meter readings. When all data is clustered, a static ZIP model is built for every load condition using (1) and (2).

D. Time-Variant Model Structure and Data Label

In this paper the time-variant model consists of multiple static ZIP models, all of which are assigned with a label. The label contains information about the load type, time and load condition. In other words, the proposed model has a tree structure that branches through three layers: load type layer, time layer and load condition layer. All the smart meter readings in the database are also labeled and clustered correspondingly, TABLE I.

TABLE I. TIME-VARIANT LOAD MODEL AND DATA STRUCTURE

Model Struc.	First Layer	Second Layer			Third Layer
	Load Type	*Season*	*Day*	*Hour*	*Load Cond.*
Data Label	Commercial, Residential, Industrial	Spring, Summer, Fall, Winter	Weekday, Weekend, Holiday	Hr. Group 1, Hr. Group 2, ... Hr. Group K	Condition 1, Condition 2, … Condition K

On the first layer, all loads are classified into commercial, residential and industrial loads. Ideally, a data mining based load modeling method does not require a user to specify the load types as long as the load is equipped with smart meters. However, marking the data with load types can help us better understand the different time-variant properties among different types of load.

On the second layer, for each individual load, all the smart meter readings are marked with time labels. Different time labels are good indicators for customer routine behaviors.

On the third layer, smart meter readings with the same time label will further be clustered and marked with different load conditions. On this layer, the ZIP model parameters are identified using smart meter data of the same load condition label.

III. DATA MINING BASED MODELING

During the load modeling process, data mining and machine learning techniques are implemented. To be specific, KL (Kullback-Leibler) divergence is used to identify and merge different time labels into hour groups on the second layer in TABLE I; K-subspace method is used to cluster data into different load conditions on the third layer in TABLE I.

A. Time Label Identification and KL Divergence

Since customer routine behaviors have a strong correlation with time. The dynamic model is marked by different season of the year, different day type (weekday, weekend, and holiday), and different hour of the day. All data collected by smart meters are marked with the corresponding time labels.

The basic time label unit is set to be one hour blocks. On the one hand, higher resolution of time labels can identify more detailed routine load behaviors. On the other hand, higher resolution time labels will leave fewer smart meter readings to each time label for regression. In order to overcome this issue, KL divergence is introduced. KL divergences of real power, reactive power, and voltage distributions of all pairs of time labels are evaluated. And different time labels with similar routine load behaviors are identified and merged.

KL divergence is a non-symmetric measure of the difference between two distributions. Let $P_1(x)$ and $P_2(x)$ be two distinct distributions, the KL divergence of the two distributions $KL(P_1(x), P_2(x))$ is given by (3) [10].

$$KL(P_1(x), P_2(x)) = \sum_{x \in X} \left[P_1(x) \cdot \log(P_1(x) / P_2(x)) \right] \quad (3)$$

A symmetric variant of KL divergence [11] given by (4) is used in this paper to quantify the divergence of load behaviors throughout different time labels. After computing KL divergence among all pairs of time labels, a KL divergence matrix can be constructed.

$$KL_{sym}(P_1(x), P_2(x)) = \left[KL(P_1(x), P_2(x)) + KL(P_2(x), P_1(x)) \right] / 2 \quad (4)$$

Fig. 1 shows the hourly weekday P-V plots for a commercial building on campus for the fall of 2012. KL divergence matrices are computed to merge those hours with highly consistent energy consumption patterns (consistent routine behaviors). Fig. 2 visualizes three normalized KL divergence matrices for three distributions respectively: real power, reactive power and voltage. Three specific KL divergence thresholds will be set for the P, Q and V KL divergence matrices to determine hours that can be merged. The final hour partition results are the intersection based on the three KL divergence matrices after their individual thresholds have been applied.

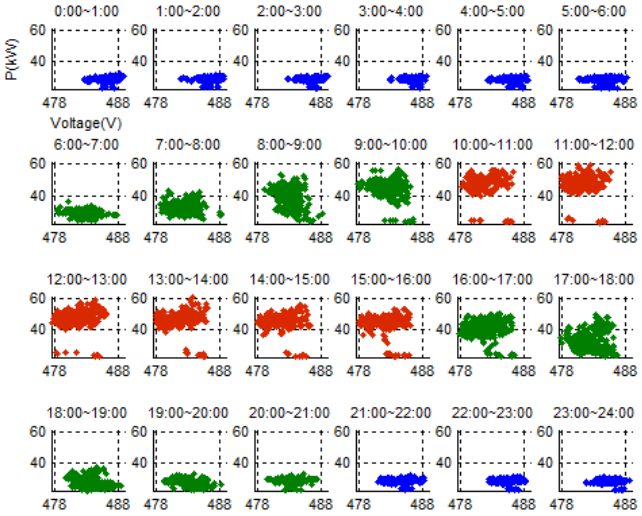

Fig. 1. P-V plots for each hour on weekdays for a commerical load.

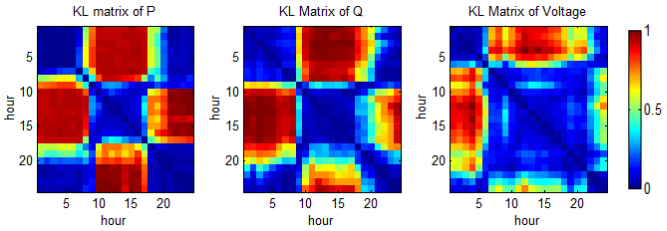

Fig. 2. Normalized KL divergence matrices for real power, reactive power, and voltage.

For the case in Fig. 1, 24 hours of the commercial load (Atlanta local time) are partitioned into working hours (red), off-working hours (blue), and hours in between (green). Since load behaviors within working hours and off-working hours are highly consistent, these hours are merged. As a result, the number of models on the second layer of TABLE I is reduced and the data for each time label increases correspondingly. Similarly, residential load and industrial load can be processed in the same way but not necessarily into the same hour partition results as the commercial loads.

B. *K-subspace Clustering*

In practice, multiple load conditions can exist under the same load type and time label. As a result, on the third layer of the model, smart meter readings are clustered into several load conditions so that each of the load conditions can be modeled by a static ZIP model.

Traditional K-means algorithm [12] clusters data based on their relative Euclidean distance to the nearest cluster center with an iterative process to adjust the centroid. The clusters' shapes are determined by the perpendicular lines between centroids. However, the smart meter readings of different load conditions are distributed in a very specific line-shaped pattern close to each other.

K-subspace method [13] allows the detection and clustering of line-shaped data by assigning each cluster C_k with a unit direction vector $\mathbf{a}_k$ and a center $\mathbf{c}_k$. The entire algorithm seeks to minimize the perpendicular distance of all the data points $\mathbf{x}_{k,i}$ to the line defined by $\mathbf{a}_k$ and $\mathbf{c}_k$ within each cluster, as shown in (5).

$$\min_{\mathbf{c}_k,\mathbf{a}_k} \sum_{i\in C_k} Dist\left(\mathbf{x}_i, C_k\right) = \sum_{i\in C_k} \left\| \mathbf{x}_i - \mathbf{c}_k - \alpha \mathbf{a}_k \right\| \tag{5}$$

Where, $\alpha = \left(\mathbf{x} - \mathbf{c}_k\right)^T \mathbf{a}_k$

Fig. 3 shows the Q-V and P-V plot of a commercial building during off-working hours on weekdays in the fall 2012. Comparing Q-V plot with P-V plot, it can be seen that reactive power are more sensitive to voltage deviations than active power. As a result, the load conditions are clustered using Q-V plot. In Fig. 3 the clustering results are marked with different colors, where the cluster number K is set to be 3.

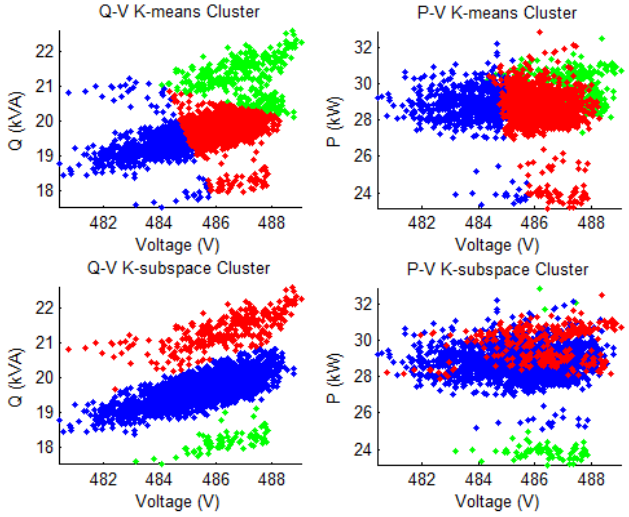

Fig. 3. Comparison between K-subspace method and K-means method.

The number of load conditions K is determined by conducting cluster evaluation. The distance from cluster C_i to cluster C_j is defined as (6). Similar to KL divergence matrix, a cluster distance matrix can be formulated in the same manner. In the cluster distance matrix, small $Dist(C_i,C_j)$ and $Dist(C_j,C_i)$ indicates cluster C_i and C_j are very close to each other and should be merged by reducing K. On the other hand, large $Dist(C_i,C_i)$ indicates larger K is required to identify all load conditions.

$$Dist(C_i, C_j) = \left|C_i\right|^{-1} \sum_{\mathbf{x}_i \in C_i} Dist\left(\mathbf{x}_i, C_j\right) \tag{6}$$

Fig. 4 shows how the number of clusters K (K=3) is determined for the Q-V plot shown in Fig. 4. A threshold h is set by experience to test the accuracy of K. In this case, h is set to be 0.1. The algorithm increases K until K equals 4 when $Dist(C_1,C_2)$ and $Dist(C_2,C_1)$ are both under the threshold h, which indicates the two clusters should be merged.

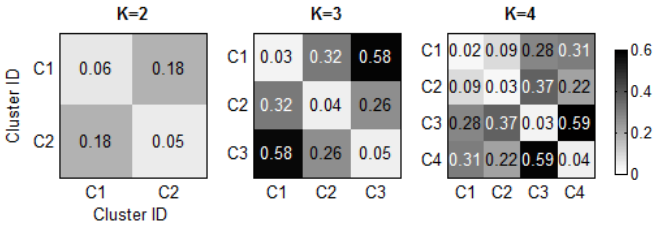

Fig. 4. Cluster Distance Matrices with Different K.

After clustering, all smart meter data has been labeled by load type, time and load condition. Data with the same label is grouped to represent a single load condition. Then regression is performed to identify the parameter of the corresponding load condition model using (1) and (2).

IV. Test Results

A. *Time label identification for different load types*

Various load types are studied to explore their differences in identifying the time label. In the study, a student residential hall and a family apartment are chosen as residential loads; an office building and a student center are chosen as commercial loads; and a chiller plant on campus is chosen as an industrial load. By using KL divergence matrices, their time label identification results for weekdays in fall are shown in TABLE II, where hours with consistent load behaviors are merged. Moreover, results shown in TABLE II also indicate that even under the same load type, different customers have their own power consumption pattern, such as the peak hours between student residential hall and family apartment. These customized properties can only be captured by performing smart meter data mining.

TABLE II. Time Label Identification Results (Weekdays, Fall)

Commercial Loads																							
Office Building	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Student Center	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Residential Loads																							
Residential Hall	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Family Apt.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Industrial Loads																							
Chiller Plant	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23

Note: ■ stands for working hours (peak hours); ■ for off-working hours (night hours); ■ for daytime hours specifically found in residential loads; ■ for hours that cannot be merged, and they are modeled on the hour basis.

B. Data Mining Based Load Model

One of the key advantages of the data mining based method is that a totally customized time-variant model can be built for every single customer equipped with a smart meter. To illustrate the idea, an office building has been chosen and modeled using the proposed data mining algorithm.

Once all of the smart meter data is clustered according to different times and load conditions, least square estimation is performed on each cluster to determine the parameters in (1) and (2). To suppress the noise from the data and avoid over fitting, an exponential model [14] is adopted in regression; p_i and q_i are computed using Taylor expansion at the nominal voltage point. TABLE III shows the partial regression results for an office building during the summer season.

TABLE III. PARTIAL MODELING RESULTS FOR SUMMER WEEKDAYS

Office Building	Working hours						Off-working hours					
	P(V)			Q(V)			P(V)			Q(V)		
Cond.	p_1	p_2	p_3	q_1	q_2	q_3	p_1	p_2	p_3	q_1	q_2	q_3
1	4.71	-5.82	2.10	18.05	-29.57	12.52	0.42	0.704	-0.12	10.86	-16.53	6.67
2	28.00	-63.01	36.00	26.63	-45.44	19.81	-0.01	0.03	0.98	12.17	-18.88	7.71
3	--	--	--	--	--	--	0.018	0.999	-0.017	11.24	-17.21	6.97

Note: P_0 and Q_0 are adopted as the load real and reactive power under the nominal voltage.

C. Model Evaluation

The new modeling method offers several advantages with respect to existing methods. First, the dynamically changing nature of the load is captured in the new model. On the short term, the load model considered the changes brought by seasons, day types and times of the day; on the long term, the customer's load composition changing is also considered, which can be more significant when comparing current load composition with the ones in late 90s [15]. Second, as long as smart meters are widely installed, there are no additional investments involved. There are no costly voltage stage tests, no load component tests, no public surveys, and no validation tests (the model is based on real system measurements). Third, the new model is fully customized for every customer equipped with a smart meter, thus more accurate than other aggregated models.

However, the proposed model has some limitations. As a result of the fact that reactive power is much more sensitive to voltage deviations than real power [14], more advanced data mining techniques are required to better capture the weak correlation between real power usage and system voltage to improve the P-V model accuracy. Since the model is based on real system measurements, which are usually around nominal operating point, the model lacks the information when voltage deviation is very large.

V. CONCLUSION

This paper demonstrates a novel data-mining method using historical smart meter databases to build a time-variant load model. We conclude from the implementation and results that the time-variant model is able to capture the dynamically changing nature of load. Due to the data mining algorithms, the proposed model can be built automatically and without additional investment involved. The time-variant load model can be implemented in various applications such as CVR, distribution system power flow analysis, load management, voltage control and tariff negotiations.

Further studies may include exploring the statistic information of the load conditions from the historical database and integrating the model into more advanced power system simulation and control applications.

REFERENCES

[1] IEEE Committee, Load representation for dynamic performance analysis. IEEE Transections on Power Systems, 1993, (2),472-482.

[2] T. Frantz, T. Gentile, S. Ihara, N. Simons, M. Waldron, "Load Behaviour Observed in LILCO and RG&E Systems", IEEE Trans., Vol. PAS-103, No. 4, April 1984.

[3] S.A.Y. Sabir, D.C Lee, "Dynamic Load Models Derived from data Acquired During System Transients," IEEE Trans., Vol. PAS-101, September 1982, pp 3365 to 3372.

[4] Vaahedi, E., et al. (1987). "Load Models for Large-Scale Stability Studies from End-User Consumption." Power Systems, IEEE Transactions on 2(4): 864-870.

[5] Price, W. W., et al. (1988). "Load modeling for power flow and transient stability computer studies." Power Systems, IEEE Transactions on 3(1): 180-187.

[6] Diaz-Aguilo, M.; Sandraz, J.; Macwan, R.; de Leon, F.; Czarkowski, D.; Comack, C.; Wang, D., "Field-Validated Load Model for the Analysis of CVR in Distribution Secondary Networks: Energy Conservation," Power Delivery, IEEE Transactions on , vol.28, no.4, pp.2428,2436, Oct. 2013

[7] EPRI report, 'End-Use Load Composition Estimation Using Smart Meter Data', 31-Dec-2010.

[8] IEEE Task Force Report, "Load Representation for Dynamic Performance Analysis," Paper 92WM126-3 PWRD, presented at the IEEE PES Winter Meeting, New York, January 26-30, 1992.

[9] Chiang, H.-D., et al. (1997). "Development of a dynamic ZIP-motor load model from on-line field measurements." International Journal of Electrical Power & Energy Systems 19(7): 459-468.

[10] Cover, T.M. and J.A. Thomas. "Elements of Information Theory," Wiley, 1991.

[11] Johnson, D.H. and S. Sinanovic. "Symmetrizing the Kullback-Leibler distance." IEEE Transactions on Information Theory

[12] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

[13] Dingding Wang (Sch. of Comput. Sci., Florida Int. Univ., Miami, FL, USA); Ding, C.; Tao Li Source: Machine Learning and Knowledge Discovery in Databases. Proceedings European Conference, ECML PKDD 2009, p 506-21, 2009

[14] Kundur, Power System Stability And Control, McGraw-Hill Education (India) Pvt Limited, 1994, pp. 272.

[15] Bokhari, A.; Alkan, A.; Dogan, R.; Diaz-Aguilo, M.; de Leon, F.; Czarkowski, D.; Zabar, Z.; Birenbaum, L.; Noel, A.; Uosef, R.E., "Experimental Determination of the ZIP Coefficients for Modern Residential, Commercial, and Industrial Loads," Power Delivery, IEEE Transactions on , vol.PP, no.99, pp.1,1