Asian Journal of Information Technology 13 (7): 375-381, 2014
ISSN: 1682-3915
© Medwell Journals, 2014

Mining Residential Electricity Consumption Patterns to Generate Tailored Baselines

[1]M. Sheeba Santha Kumari, [1]A.P. Shanthi and [2]V. Uma Maheswari
[1]Department of Computer Science and Engineering, [2]Department of Information,
Anna University, CEG Campus, Guindy, 600025 Chennai, Tamil Nadu, India

Abstract: Residential electric power consumption plays an important role in economical decision making process. It is beneficial to have residential consumers who are better aware of their consumption pattern so that they are more responsible in usage of power. Normally, they rely on their long term bill and do not have any insight into their pattern of consumption which can hinder efforts to reduce electricity consumption. Emergence of smart grid with advanced metering devices and data mining technique facilitates power consumers to perform efficient power management. Behaviour modification initiatives and tailor made suggestions can be generated by examining the consumption patterns, enabling consumers to control load, participate in demand response programs and to help suppliers fix time dependent tariff rates. This study intends to generate typical load patterns of a household highlighting the usage pattern using a Conceptual Hierarchical Clustering Method. A real data set is used for the study.

Key words: Data mining, pattern recognition, conceptual clustering, residential power consumption, tailored baselines, cluster validation

INTRODUCTION

Power consumers should get greater wakefulness and responsibility to save power resources as electricity is one of the most pertinent concerns that current societies are facing. With never-ending population growth, expansion of urbanization, technological booming, changing lifestyles and industrial intensification, consumption has increased (Hubacek *et al*., 2007). In particular, the consumption is rapid in residential and commercial buildings than transportation or industrial sectors which lead to blackouts and brownouts (Mittal *et al*., 2013).

Consumers are unaware of their usage of electricity consumption as they are metered out of sight with limited information left. Fortunately, through smart interval meters (Mittal *et al*., 2013), detailed electricity consumption of buildings is obtainable. In developed countries, these smart meters have come into place which enables data exploration. Smart interval meters should fill in place of old meters that facilitate bulk data storage in all buildings. Traditional power grid is getting transformed to modernized smart grid (Amin and Wollenberg, 2005) through advancement of information and communication technologies which is necessary for efficient power management. Thus, the development of power aware system is a prerequisite that brings a profound insight to consumers about their pattern of consumption and how their everyday activities impacts electricity usage. Data mining, machine learning and AI technologies are an effective means of dealing with such problems that can assist in analysis and finding intelligent ways to improve efficiency (Ramos and Liu, 2011).

Mathieu *et al*. (2011) has used a DRA based visual analysis to explore demand and the behavioral pattern of electricity usage in the university buildings. Magnitude (low and high consumption) and dynamics (shape) of consumption of each building group is examined. Occupancy and electricity consumption is analyzed based on the detection of internet usage in a university campus using linear regression (Martani *et al*., 2012). With the introduction of electricity markets, Demand Response (DR) is being implemented in many places (Balijepalli *et al*., 2011). The focus of Moran *et al*. (2013) is to discover DR Programs in commercial and industrial buildings to control load in peak hours and reduce electricity bills using linear regression model. Dent *et al*. (2012) has applied K-means to residential data to explore low and high variability in behavior, low and high usage in peak period. Pattern recognition methodology has been

Corresponding Author: M. Sheeba Santha Kumari, Department of Computer Science and Engineering, Anna University, CEG Campus, Guindy, 600025 Chennai, Tamil Nadu, India

widely used to study the electricity behavior of customers and evaluation of load profiles (Tsekouras *et al*., 2008; Loughlin *et al*., 2012; Dent *et al*., 2011). This analysis and prediction of consumption behavior plays a vital role in managing the balance between production and demand helping to reduce energy losses and pollution generation (Richardson *et al*., 2010).

METHODOLOGY

Clustering is concerned with the task of discovering similar objects or common patterns given unlabelled data. Classit (Gennari *et al*., 1989) is a concept based hierarchical clustering influenced from cobweb proposed by Douglas H. Fisher. The CLASSIT algorithm is faster than traditional hierarchical clustering since it dynamically builds a dendogram by processing one data point at a time instead of following divisive or agglomerative approaches (Fisher, 1987). The system initializes its hierarchy to a single node called root based on the values of the first instance. Classification and learning are intertwined with incremental adding of objects into the tree one by one, by performing hill-climbing search. For each instance, it determines the child of root that best hosts object considering four possible operations namely create new concept, insert into existing concept, merge or split two concepts (Gennari *et al*., 1989). The best host to fit in the object is found for each instance based on an evaluation function called category utility explained in study.

Category utility: Category Utility (CU) is defined as the increase in the expected number of attribute values that can be correctly guessed:

$$P(C_k)\sum_i \sum_j P\left(A_i = V_{ij}|C_k\right)^2$$

given a partition (C_1, C_2, ..., C_n) over the expected number of guesses with no such knowledge:

$$\sum_i \sum_j P\left(A_i = V_{ij}\right)^2$$

(Fisher, 1987; Han and Kamler, 2006). Category utility for categorical attributes is given in Eq. 1 where, k, i, j, n represents the classes, attributes, its values and number of categories, respectively:

$$CU(C_1, C_2, ..., C_k) = \frac{\sum_{k=1}^{n} P(C_k)\left[\begin{array}{l}\sum_i \sum_j P\left(A_i = V_{ij} \mid C_k\right)^2 - \\ \sum_i \sum_j P(A_i = V_{ij})^2\end{array}\right]}{n} \qquad (1)$$

Intra-class similarity is reflected by conditional probabilities of the form $P(A_i = V_{ij}|C_k)$ where, $A_i = V_{ij}$ is an attribute-value pair and C_k is a class. The larger this probability, the greater the proportion of class members sharing the value and the more predictable the value is of class members. Inter-class similarity is a function of $P(C_k/A_i = V_{ij})$. The larger this probability, the fewer the objects in contrasting classes that share this value and the more predictive the value is of the class.

The same concept is employed when calculating category utility for numerical attributes in CLASSIT. As applied to continuous domain it assumes normal distribution and uses probability density function. The formula is given as:

$$CU\left(C_1, C_2, ..., C_n\right) = \frac{\sum_{k=1}^{n} P\left(C_k\right) \sum_{i}^{I} \left(\frac{1}{\sigma_{ik}} - \frac{1}{\sigma_{ip}}\right)}{n} \qquad (2)$$

Where:

σ_{ik} = The standard deviation of a particular class

σ_{ip} = The standard deviation at the parent node independent of the class membership

Control system parameters: Control parameters in CLASSIT are acuity and cut off. They have a great influence on the structure of the tree, indirectly controlling the number of clusters in the tree. Acuity is set to be the minimal standard deviation of a cluster attribute. It is needed in order to disallow the cases where standard deviation is zero which leads to infinite values in the category utility formula. Cut off is set to be minimal category utility. It is the minimum increase of CU to add a new node to the hierarchy otherwise the new node is cut off.

Extraction of typical consumption patterns: The CLASSIT System is applied to the data set of a residential household, taken from the UCI machine learning repository (Hebrail and Berard, 1997). It contains a total of 2075259 instances with each instance representing a minute averaged reading. The dataset covers a total span of 47 months with attributes date, time, global active power, global reactive power, global intensity, voltage, sub meter reading 1, 2 and 3. Sub meter reading 1 and 2 corresponds to kitchen appliances and laundry room, respectively. Sub meter reading 3 corresponds to electric water-heater and air-conditioner. Instances for particular day is averaged to get the 'minute average reading' for that day. The clustering is done based on the daily consumption data obtained for all 4 years.

Choose optimal values for control parameters: To achieve the best success rate, optimal values of parameters Acuity and Cutoff should be chosen. Some instances are deemed sufficiently similar to others and retaining this instance is not useful. Cutoff governs the similarity threshold which is used to suppress the growth. Over fitting occurs for low values of acuity, since this encourages larger number of singleton classes to be formed and thus to idiosyncratic predictions. Thus, to overcome under fitting and over fitting effects, it is necessary to fine-tune these parameters and determine the quality of the clusters generated. A good Clustering Method has high intra-class similarity or cohesion and low inter-class similarity or separation (Han and Kamler, 2006; Zhao, 2012). Internal cluster validity indicators combining measures of cohesion and separation to assess the quality of the clusters are discussed.

Cluster Dispersion Indicator (CDI): CDI (Chicco *et al*., 2003) is defined as the mean infra-set distance between the members in the same cluster and inversely on the infra-set distance between the clusters. It can be estimated as follows:

$$\text{CDI}(\text{K}) = \frac{1}{\hat{\text{d}}(\text{C})}\sqrt{\frac{1}{\text{K}}\sum_{k=1}^{\text{K}}\hat{\text{d}}^2(\text{L}^{(k)})} \tag{3}$$

Where:

C = The set of cluster centres

$\text{L}^{(k)}$ = The set of members of the kth cluster. A lower value for CDI suggests a better clustering solution

Calinski and Harabasz index (CH index): The CH index (Calinski and Harabasz, 1974) is calculated using the following equation:

$$\text{CH index} = \frac{\text{SSB}(\text{M}-1)}{\text{SSW}(\text{N}-\text{M})} \tag{4}$$

It evaluates the cluster validity based on the average between and within cluster sum of squares (SSB and SSW). The maximum value gives the best clustering solution.

WB index: It provides a measure of the ratio of the within cluster scatter to the between cluster separation (Zhao *et al*., 2009). The effect of SSW is emphasized by multiplying it with the number of clusters. The minimum value gives the best clustering solution.

It is calculated as M×SSW/SSB where, M denotes the number of clusters. With extreme values of Acuity and Cutoff, few or many disjuncts were formed resulting in under fitting and over fitting, respectively. Based on the values of cluster validity indices calculated as seen in Table 1, Acuity value of 1.5 and Cutoff of 0.017 was determined as best.

Table 1: Cluster validity indices calculation

Number of clusters	CDI	CH index	WB index
2	4.37	0.001	1.541
3	1.90	0.005	0.890
4	1.32	0.012	0.821
5	1.00	0.021	0.768
6	0.61	0.053	0.443
7	0.49	0.083	0.418
8	0.44	0.104	0.424
9	0.46	0.100	0.568
10	0.47	0.097	0.734

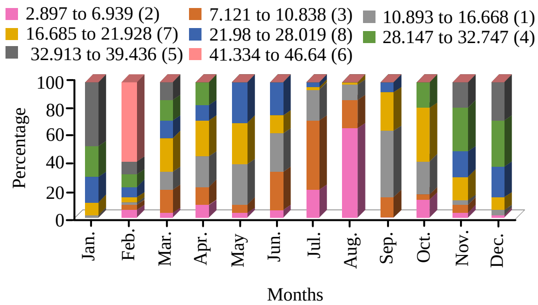

Fig. 1: Distribution of months in eight clusters. *Bullets denotes global active power (cluster ID)

Description of clusters: The clusters obtained are described below explaining the power consumption of days having similar patterns. Figure 1 presents the overall monthly analysis of all clusters and the Global Active Power or GAP (Wh) is the main criteria based on which all the studies are performed.

The final tree with 8 clusters is shown in Fig. 2 with lower CDI and second lower WB index values and had maximum CH index value. The root represents the most general concept that covers the entire observations. Each cluster or concept is represented in a rectangle box representing a class. Each cluster has a discrimination that assists to form daily routines and baselines.

Cluster 1: Cluster 1 includes 375 days with mean consumption of 13.93. The cluster is more frequent in September and spread among the months in spring season as shown in Fig. 1. The cluster epitomizes summer week day and early autumn behaviour (Fig. 3).

Cluster 2: Cluster 2 includes 75 days taken for the study. It indicates a load curve of extreme low consumption with a mean value of 5.22. It represents typical summer season asseen in Fig. 4a. It is particularly significant of electricity consumption in the august month which is due to vacation (Fig. 1). It indicates a warm temperate day where there is more sunshine and no space heating or water heating is required.

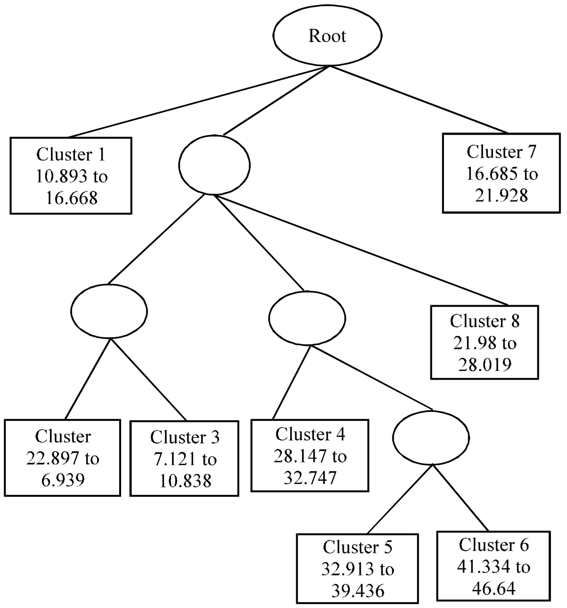

Fig. 2: Final hierarchical tree

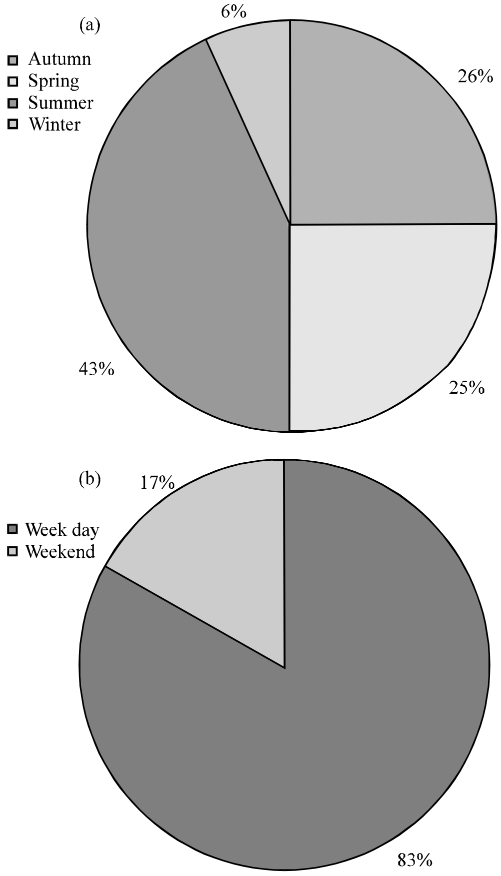

Fig. 3: a) Seasonal and b) Weekends/week days distribution of cluster 1

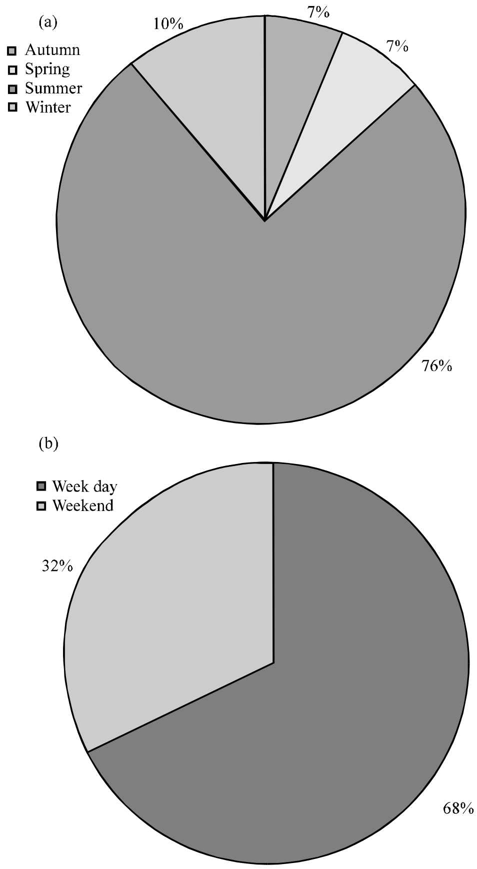

Fig. 4: a) Seasonal and b) Weekends/week days distribution of cluster 2

Cluster 3: The 84 days compose cluster 3 whose consumption is slightly higher than cluster 2 with a mean value of 9.29. It signifies summer season where distribution is more in June and July months (Fig. 1 and 5a). Most of working days from Monday to Friday form cluster 1, 2 and 3. Cluster 2 and 3 characterizes standby, off modes of appliances and absence of occupants.

Cluster 4: The 50 days characterize cluster 4 with a very high consumption with mean value of 30.21. It represents late Autumn and Winter season. It constitutes equal weekends and week days (Fig. 6). Consumption is high because of space and water heating due to extreme cold climate.

Cluster 5: Cluster 5 includes 32 days with a mean consumption of 35.8. It is particularly significant of electricity consumption in the Winter weekends (Fig. 7).

Cluster 6: Four February days fall in cluster 6 with extreme consumption. The days are not public holidays but all the four are weekends.

Cluster 7: Cluster 7 composes 429 days representing medium consumption with a mean value of 19.14. The distribution across months is wide excluding the Summer months with week days proportionally higher to weekends (Fig. 1 and 8). There is higher proportion of public holidays making up this cluster with 28.57%.

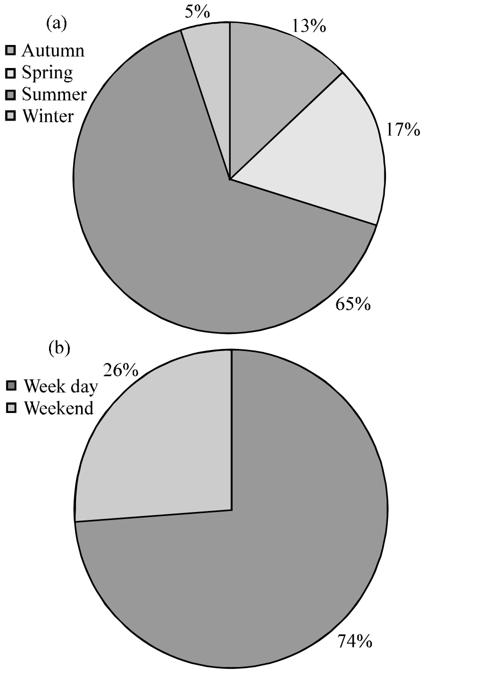

Fig. 5: a) Seasonal and b) weekends/week days distribution of cluster 3

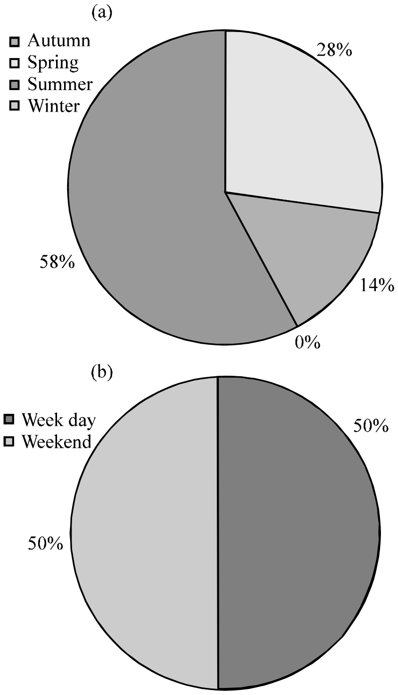

Fig. 6: a) Seasonal and b) Weekends/week days distribution of cluster 4

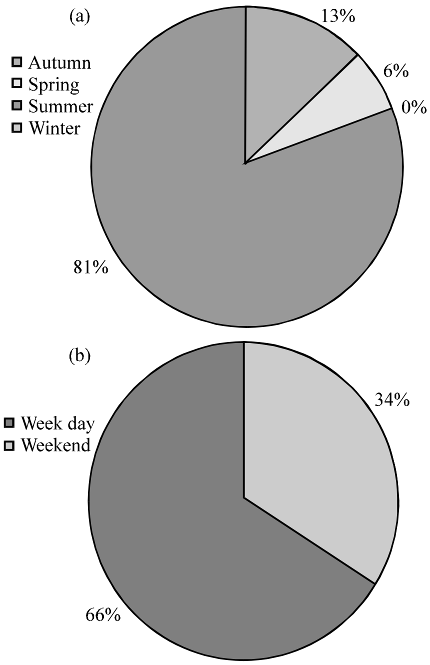

Fig. 7: a) Seasonal and b) Weekends/week days distribution of cluster 5

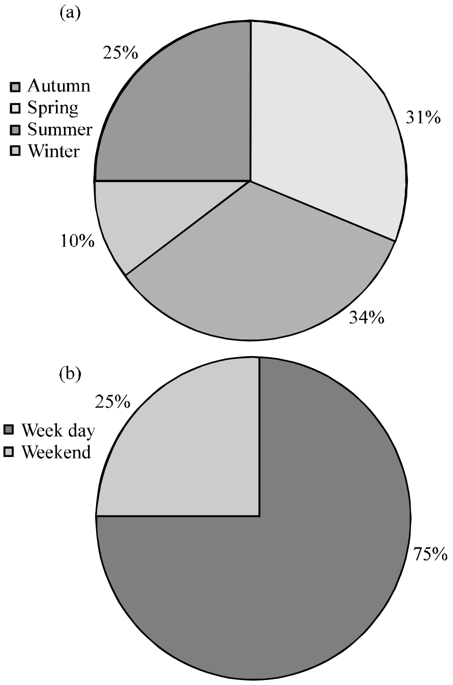

Fig. 8: a) Seasonal and b) Weekends/week days distribution of cluster 7

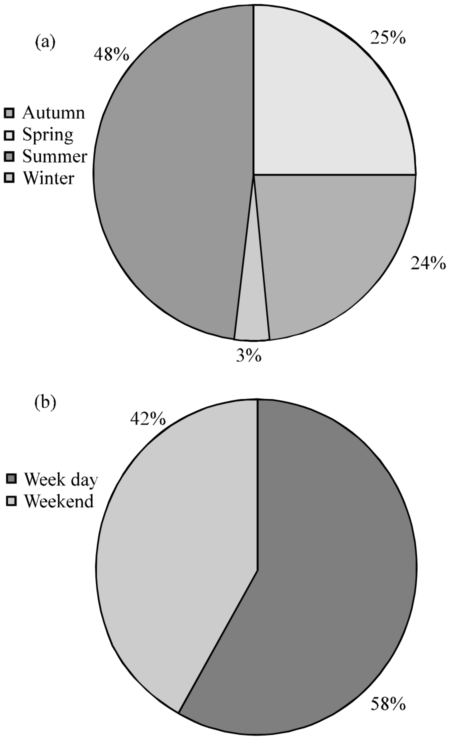

Fig. 9: a) Seasonal and b) Weekends/week days distribution of cluster 8

Cluster 8: The 227 days compose cluster 6 representing one fourth of Autumn and Spring and 50% of Winter (Fig. 9). The days included in this cluster are well distributed among weekend and week days. It characterizes mild cooler week days and weekends.

DISCUSSION

The analysis based on the season, weekends or week days and public holidays shows there is high demand in Winter weekends. Public holidays of the household do not have much impact on the consumption. The baselines obtained through the study are: majority of the instances that fall in cluster 4-6 corresponds to Winter weekends. The average readings of the three sub meters in these clusters are 2.84, 3.27 and 9.76. These readings are higher than those of the other clusters. This indicates that kitchen and laundry room appliances have highest usage during this period apart from heating appliances. This is an indication that people prefer to stay indoors during winter weekends; cluster 1-3 represents instances corresponding to week days in summer when there is low demand. The average readings of the three sub meters in these clusters are 0.46, 1.08 and 3.92. These are the least among all the clusters. It is an indication that kitchen and laundry room appliances have the least usage during this period. It is logically meaningful that people prefer going outdoors during this period and light Summer clothing generates less load on the laundry appliances. Based on the study, power distributors can take steps to generate more power during winter weekend periods that requires high demand. It also encourages performing electricity waste elimination on low demand days. Authorities can take appropriate decisions to reduce the residential power consumption during the peak period. For instance, to reduce the load on kitchen appliances, special deals may be offered by the restaurants to encourage home delivery of food. This research can be enhanced further by using the hourly and quarterly consumption of electricity to study in-depth usage and behaviour. This would enable the authorities to suggest appropriate period for the usage of different appliances in different localities.

CONCLUSION

The classit clustering has highlighted different classes using which baselines are formed. Recognition of the patterns by interpreting the days is done based on the loading characteristics. Generation of tailored baselines that reflect the condition of the household enables to develop feedback mechanism. The consumption patterns and the baselines obtained are useful for power service companies and to accomplish automatic service personalization as they reflect the lifestyle of the family members in the households of a particular locality. It is highly beneficial for the consumers to be aware of their electricity usage patterns to optimize on their energy costs.

REFERENCES

Amin, S.M. and B.F. Wollenberg, 2005. Toward a smart grid: Power delivery for the 21st century. IEEE Power Energy Magazine, 3: 34-41.

Balijepalli, V.S.K.M., V. Pradhan, S.A. Khaparde and R.M. Shereef, 2011. Review of demand response under smart grid paradigm. Proceedings of the IEEE PES Innovative Smart Grid Technologies-India, December 1-3, 2011, Kollam, Kerala, pp: 236-243.

Calinski, T. and J. Harabasz, 1974. A dendrite method for cluster analysis. Commun. Stat., 3: 1-27.

Chicco, G., R. Napoli, P. Postolache, M. Scutariu and C. Toader, 2003. Customer characterization options for improving the tariff offer. IEEE Trans. Power Syst., 18: 381-387.

Dent, I., T. Craig, U. Aickelin and T. Rodden, 2012. An approach for assessing clustering of households by electricity usage. Proceedings of the 12th Annual Workshop on Computational Intelligence, September 5-7, 2012, Edinburgh, Scotland.

Dent, I., U. Aickelin and T. Rodden, 2011. Application of a clustering framework to UK domestic electricity data. Proceedings of the 11th Annual Workshop on Computational Intelligence, September 7-9, 2011, University of Manchester, Manchester.

Fisher, D.H., 1987. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2: 139-172.

Gennari, J.H., P. Langley and D. Fisher, 1989. Models of incremental concept formation. Artificial Intell., 40: 11-61.

Han, J. and M. Kamler, 2006. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, New York.

Hebrail, G. and A. Berard, 1997. Individual household electric power consumption data set. University of California, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption.

Hubacek, K., D. Guan and A. Barua, 2007. Changing lifestyles and consumption patterns in developing countries: A scenario analysis for China and India. Futures, 39: 1084-1096.

Loughlin, F., A. Duffy and M. Conlon, 2012. Analysing domestic electricity smart metering data using self organising maps. Proceedings of the Workshop on Integration of Renewables into the Distribution Grid, May 29-30, 2012, Lisbon, pp: 1-4.

Martani, C., D. Lee, P. Robinson, R. Britter and C. Ratti, 2012. ENERNET: Studying the dynamic relationship between building occupancy and energy consumption. Energy Build., 47: 584-591.

Mathieu, J.L., P.N. Price, S. Kiliccote and M.A. Piette, 2011. Quantifying changes in building electricity use, with application to demand response. IEEE Trans. Smart Grid, 2: 507-518.

Mittal, S., Y. Maheshwari, A. Singh, A. Varshney and S. Agarwal, 2013. Smart grid-a pragmatic vision for blackout free future. Int. J. Sci. Eng. Res., 4: 614-620.

Moran, A., J.J. Fuertes, M.A., Prada, S. Alonso, P. Barrientos, I. Diaz and M. Dominguez, 2013. Analysis of electricity consumption profiles in public buildings with dimensionality reduction techniques. Eng. Appl. Artificial Intell., 26: 1872-1880.

Ramos, C. and C. Liu, 2011. Intelligent systems in power systems and energy markets. IEEE Intell. Syst., 26: 38-45.

Richardson, I., M. Thomson, D. Infield and C. Clifford, 2010. Domestic electricity use: A high-resolution energy demand model. Energy Build., 42: 1878-1887.

Tsekouras, G.J., P.B. Kotoulas, C.D. Tsirekis, E.N.Dialynas and N.D. Hatziargyriou, 2008. A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. Electric Power Syst. Res., 78: 1494-1510.

Zhao, Q., 2012. Cluster validity in clustering methods. Ph.D. Thesis, University of Eastern Finland, Finland.

Zhao, Q., M. Xu and P. Franti, 2009. Sum-of-squares based cluster validity index and significance analysis. Proceedings of the 9th International Conference on Adaptive and Natural Computing Algorithms, April 23-25, 2009, Kuopio, Finland, pp: 313-322.