CLASSIFICATION OF ELECTRIC GRID CUSTOMERS USING DATA MINING METHODS

For CSCI 6907 Big Data and Analytics

Summer 2016

7/18/16

Amit Talapatra
GWID: G49976025
amit_talapatra@gwmail.gwu.edu

Abstract

This research paper illustrates the role that data mining techniques and “Big Data” analytics can have in energy management problems faced by electric utilities. The paper identifies customer classification as one of the key needs of utilities that can be addressed through data mining. Several clustering-focused methods for customer classification are identified with examples and detailed explanations of the pros and cons of each. The methods covered include: k-means clustering, fuzzy c-means clustering, hierarchical clustering, self-organizing maps, and clustering validity assessment.

Table of Contents

1. Classification of Electric Grid Customers Using Data Mining Methods

1.1 Introduction

Discussions on energy management are increasingly in the public spotlight. This is an issue tied to climate change, the security of our power grid, the advancement of renewable technologies, the growth of computing power, and the increasing ability of the average consumer to generate their own energy independently or even sell it to the grid. I have worked as an engineer and analyst for 5 years implementing energy management programs for utilities, recommending energy efficiency projects for industrial facilities, and working with federal agencies that are responsible for guiding the development of advanced renewable energy technologies. As I transition my career towards data science, it is clear to me that data mining methods and "Big Data" analytics are playing a significant role in the development of a more efficient energy grid. For the purposes of this paper, data mining refers to the practice of extracting useful information from large datasets. "Big Data" is currently a fairly nebulous term, but for this paper, I use it to refer to the area of analytics methods that seek to solve problems that come into play when data scales to a size where it cannot be processed on just one machine. The methods discussed in this paper fit under the category of data mining techniques, and many of these methods are designed to process data at the "Big Data" scale.

Though there are many problems in energy management that can be addressed through data science, this paper focuses on customer classification and compares the clustering methods that can be used to address this problem for electric utilities. Customer classification is important because it lets utilities better predict future energy needs and design more efficient programs that effectively incentivize customers to invest in technologies that benefit the grid. For example, if a food processing facility is identified as an industrial customer with peak energy use in the mid-afternoon, utilities can develop a program that encourages that customer to shift their energy use to other parts of the day, or reduce it using energy efficient technologies. Data mining methods can enable utilities to identify likely candidates for these types of measures out of their millions of customers. The reduced peak demand means that the utility can reduce its power generation capacity, which may mean reduced operations and maintenance costs for its facilities.

The following sections illustrate the role that data mining may have in addressing this problem, identify some of the data mining techniques for customer classification that best suit this problem, and compare the pros and cons of each of these techniques.

1.2 Background

With distributed power generation and the advancement of new energy technologies, electric grids are getting more complex, and utilities can make use of data mining and "Big Data" analytics methodologies to better understand and manage the power needs of customers. If we use graph theory to understand an electric grid, customers can represent the vertexes or nodes. Each customer is a discrete unit with individual energy demand parameters. That is, each vertex (customer) requires energy flowing to it from the grid to meet its daily needs. Much of the available data comes in the form of energy demand profiles, as shown in Figure 1 These show energy use measured as demand (as kW) over a 24-hour period, and are valuable for classification because different types of residential, commercial, and industrial customers show vastly different demand profiles. In addition, electricity prices vary based on time-of-day, making demand profiles one of the most useful data types for understanding how customers impact the grid. Other useful data types for categorizing customer needs are building characteristics and survey responses.

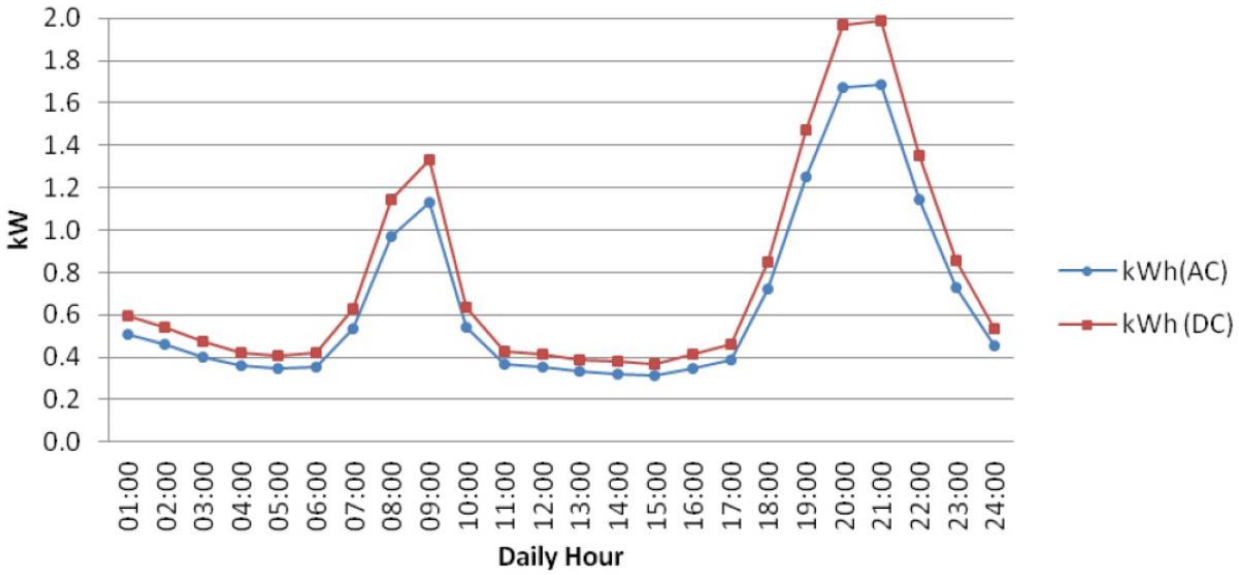

Figure 1. An example of a daily load profile for a residential unit (Arif, Oo and Ali 2013).

Energy demand as a function of time varies for each of the customers in the grid, and providing an efficient and reliable flow of energy is essential for managing a reliable electric grid. This problem is likely to be compounded in the future, as distributed generation (e.g. customer-owned solar generation), and distributed energy storage allows for energy to flow both *to* and *from* customers.

The flow of energy from customers is currently managed by two common types of tariff structures: net-metering and feed-in tariffs. Net-metering allows customers with distributed generation to transfer power back to the grid, offsetting the cost of power received from the grid and treating the flow of energy to and from the grid as a net sum. Feed-in-tariffs measure customer-generated electricity independently of electricity used from the grid, and typically come with a long-term price contracts that make the price paid for generated electricity independent of the current grid price. This incentivizes customers to invest in renewable energy generation by removing the risk of uncertainty in the price of electricity (Poullikkas 2013). As more customers move to renewable distributed generation, utilities have an incentive to better categorize customer needs which can allow for better tuning of the parameters of tariff structures. Properly designed tariff structures can both incentivize customers to invest in distributed generation, and be economically viable for utilities.

Although there are current structures in place to manage customer demand, the market share of distributed power generation technologies is growing, and the market is massive. Figure 2 shows the projected growth of renewable distributed energy capacity for the Eastern U.S. through 2030. Note the significant increase in the projected impact of solar power. This expanded capacity refers to an increase in distributed generation; meaning that more customers and energy producers will be sending energy back into the grid. This change in customer behavior will create new types of energy use patterns and the expanded role of distributed generation may mean that utilities have more options for ensuring that customers get the energy they need.

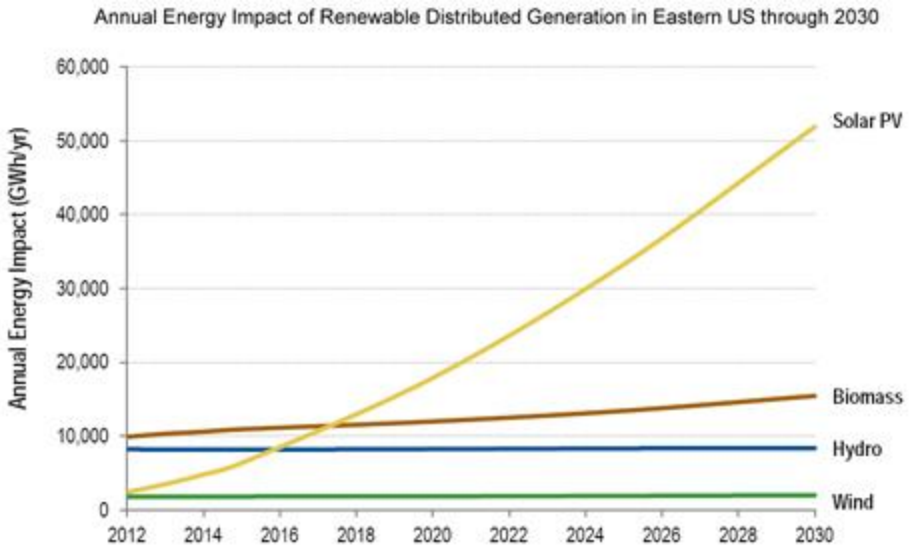

Figure 2. Annual energy impact of renewable distributed generation in the Eastern U.S. through 2030 (Projection by Navigant Research) (Pentland 2013).

Table 1. Number of electricity customers in the U.S. by type of energy provider (American Public Power Association 2016).

Number of Customers

	Full-Service Customers	Delivery-Only Customers	Total	% of Total
Publicly Owned Utilities	21,384,953	9,383	21,394,336	14.5%
Investor-Owned Utilities	88,111,658	13,040,013	101,151,671	68.4%
Cooperatives	18,903,950	16,051	18,920,001	12.8%
Federal Power Agencies	38,870	0	38,870	0.0%
Power Marketers	6,344,231	0	6,344,231	4.3%
TOTAL	**134,783,662**	**13,065,447**	**147,849,109**	**100.0%**

Table 1 shows the number of customers serviced by electric utilities in the United States. With customers in the millions, a growing capacity for distributed generation, and collection of both time-series demand data and customer characteristics, the scale of data described here fits under the category of "Big Data" and utilities can learn from data analytics methodologies to derive insights from this data. The data available to utilities is customer-centric, which is why customer classification using this data is one of the most important data analytics applications for utilities.

1.3 Clustering Methods

One family of methods that can be used to understand electric grid customers is clustering. Clustering allows us to group data observations relative to each other based on a particular algorithm. In the case of customer classification, it can help us create distinct categories of customers, but the value of these groupings requires an analyst to choose the right clustering method for the intended goal. When considering clustering methods, the analyst's primary goal is to create a set of customer classes that can be used to maximize the effect of the energy management program that they intend to design. However, with the growth of available data, and the availability of more advanced data processing methods, analysts must often weigh external factors such as speed and interpretability of the data mining methods that they have available.

1.3.1 K-Means Clustering

K-means clustering is a commonly used clustering algorithm which creates "cluster centers" or new data observations that represent a group of observations from the dataset as a central point. For classification, the k-nearest neighbor algorithm can then be used to find the nearest "cluster center" for each data observation, creating groups of observations defined by the similarity of their variables. In the case of customer classification these groups, or clusters, can represent distinct categories of customers, and utilities can use this data to develop programs that better suit their needs.

One use of K-means clustering for customer classification is presented in the Institute of Electrical and Electronics Engineers (IEEE) paper: "An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques" (Figueiredo, et al. 2005). In this example, a dataset of electrical load profiles (like the one shown in Figure 1), are run through a load profiling module followed by a classification module. Load profiling creates the set of consumer classes based on a dataset of many customer load profiles and defines rules that describe each customer type. The classification module then uses the C5.0 algorithm and the rule definitions created by load profiling to develop a decision tree that sorts customers into these classes. A diagram of the process is shown in Figure 3, with examples of the load profile classes generated in Figure 4. Note the location of K-means clustering as part of the load profiling module (Box B in Figure 4).

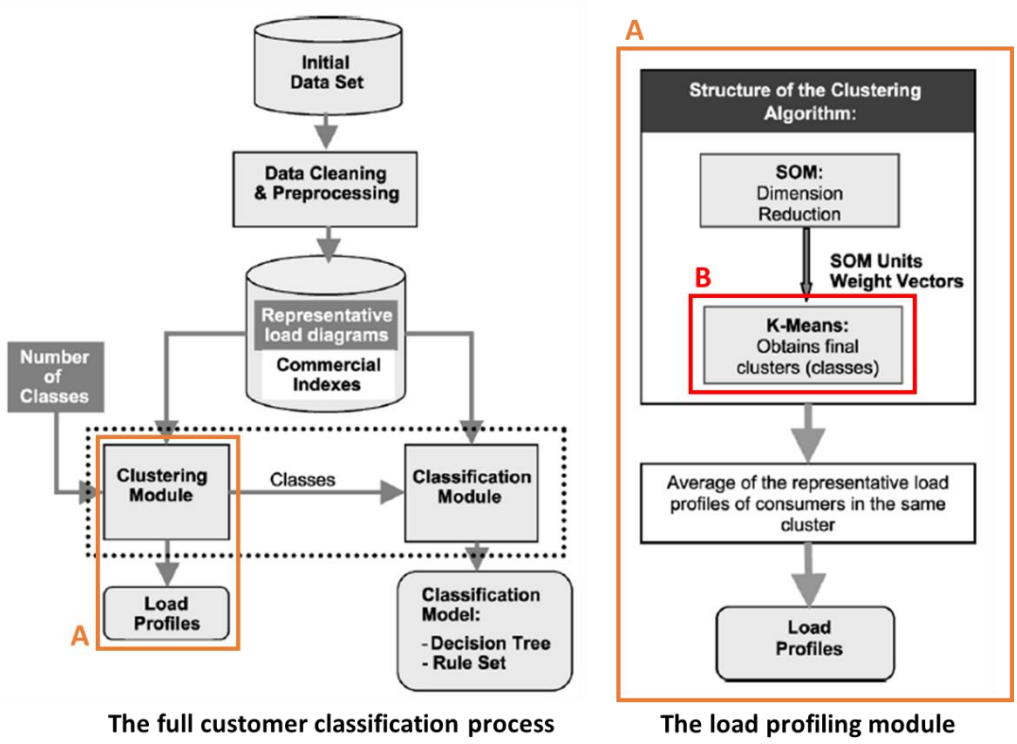

Figure 3. (A) The load profiling module. (B) The K-means clustering step (Figueiredo, et al. 2005).

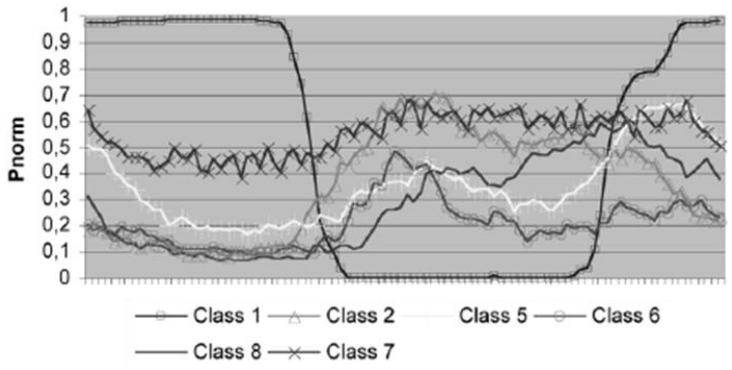

Figure 4. Example load profiles to represent the 8 classes defined by the classification method (Figueiredo, et al. 2005).

This example shows a typical role that k-means clustering can play in classification. K-means clustering is useful for load profile-based classification because it compares observation variables by distances (i.e. Euclidean distance) and produces higher variability in groupings when using a low number of variables (low-dimensional data). As load profiles measure two variables, this is a suitable application. K-means clustering is also preferred for having lower

time complexity than comparable methods, and therefore benefitting from fast computation speeds.

One limitation of k-means clustering is that it requires the user to provide another method to define "k". In many cases, this can be done manually, where the user runs the clustering algorithm for several different values of "k" and selects the result where the clusters best describe the data for their purposes. The examples cited in this report typically define ~10 customer types, and for smaller datasets, it may be practical for the user to take a manual approach to defining "k." For datasets with more variability in customer types, a calculated metric may have to be defined to automate the process to select "k" (see 1.3.5 Clustering Validity Assessment).

K-means clustering is also known to suffer from the "curse of dimensionality," as many clustering methods do to some degree (Steinbach, Ertoz and Kumar 2004). The "curse of dimensionality" refers to how some data analysis methods become ineffective when observations in the dataset contain a high number of distinct variables. For the purposes of customer classification, we treat the demand profile data from each customer as a unique observation, though in some cases other characteristics may be used such as building type, square footage, insulation type, building age, and occupancy level (Kim, Aravkin and Zondervan 2016). Dimensionality problems can occur when attempting to cluster along too many of these variables such that the algorithm can no longer cluster the observations in a useful way. Dimensionality reduction refers to the family of methods used to condense information from multiple variables and these methods can be used to reduce the impact of the "curse of dimensionality." An example of a dimension reduction step is the self-organizing map (SOM) covered in section 1.3.4. See this section for more detail.

1.3.2 Fuzzy C-Means Clustering

An alternative to k-means clustering used in utility customer classification is fuzzy c-means clustering (FCM). This is similar to the k-means method in that FCM also assigns observations to clusters. The difference is that FCM assigns all observations to all clusters and uses a metric to measure how similar each observation is to each cluster. This algorithm is demonstrated for utility customer classification in the IEEE-published paper "Application of Clustering Technique to Electricity Customer Classification for Load Forecasting" (Wang, Li

and Yang 2015). In this example, rather than demand profiles, researchers analyzed electricity consumption, recorded quarterly, to identify different classes of customers. The report compares fuzzy c-means clustering to k-means clustering using a clustering validity assessment index (this is discussed in further detail in section 1.3.5 Clustering Validity Assessment). In general, when comparing FCM to k-means clustering, both produce similar results, but FCM is computationally slower. However, FCM can approximate a solution faster in many cases, making it a useful alternative to k-means clustering (Ghosh and Dubey 2013).

Another IEEE-published paper, "Consumer Load Profiling Using Fuzzy Clustering and Statistical Approach" (Zakaria, Othman and Sohod 2006), compares a fuzzy clustering method against k-means for customer demand load profiles. In this example, the fuzzy clustering method was determined to be more accurate that k-means due to its flexibility in assessing cluster boundaries. Where k-means assigns each observation to a strictly-defined cluster, fuzzy clustering provides the user with more information about how similar each observation is to each potential type of cluster, which provides more information about the nature of the cluster relationships in the data set. These cases demonstrate how fuzzy c-means clustering can be advantageous when computation speed is not primary concern.

1.3.3 Hierarchical Clustering

The paper "Methods for Generating TLPs (Typical Load Profiles) for Smart Grid-Based Energy Programs" (Kim, Ko and Choi 2011) presents another clustering method that is applicable to the classification of customers in utility networks. This example, like most of the examples already covered in this report, also uses demand load profiles as the primary data type. Hierarchical clustering works by creating a cluster tree out of the data. When applied to demand load profiles, first the algorithm must assess similarity between the profiles. Next, the algorithm creates rules by which to divide the data set, progressively separating groups of profiles into branches defined by the rules used to separate them. This can be best illustrated by the diagram in Figure 5. This image depicts a dendrogram; a simplified form of the type of hierarchical cluster tree created by this type of clustering algorithm. Clusters are defined as the process moves further down the tree, and the branches in the lowest layer represent the final set of clusters.

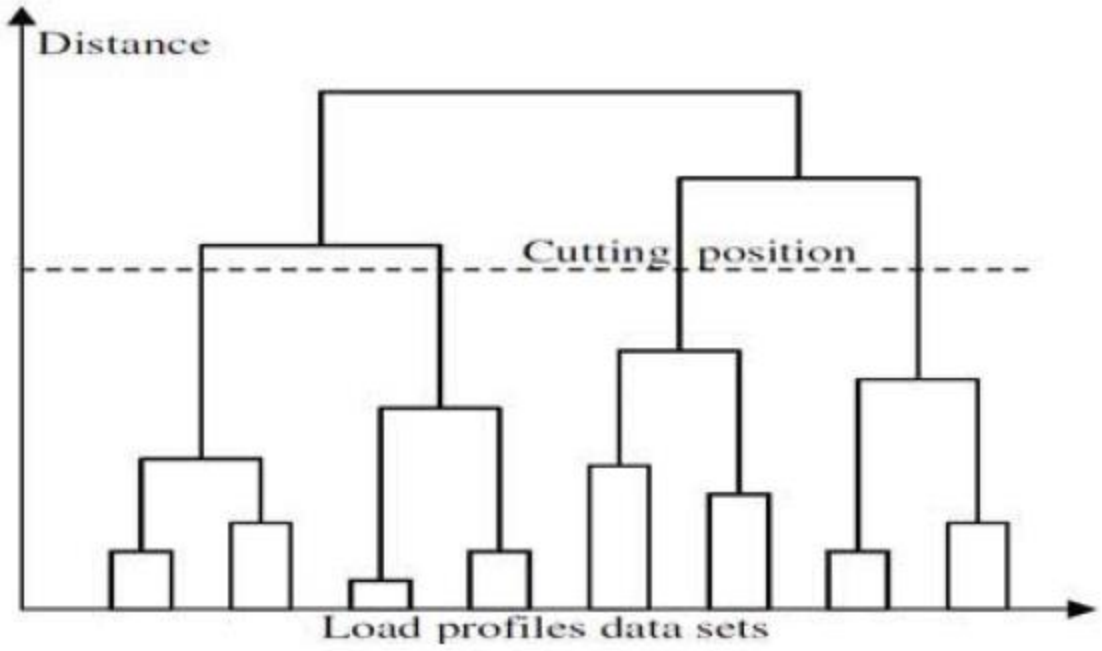

Figure 5. A dendrogram illustrating the methodology of hierarchical clustering for demand load profiles (Kim, Ko and Choi 2011).

The key advantage that this method has over k-means and fuzzy c-means clustering is that the number of clusters does not have to be pre-determined. Hierarchical clustering develops the cluster set starting with the full set of observations at the top of the cluster tree. Moving down the tree, the data branches at various "cutting positions," where subsets of observations are branched off based on a particular variable rule (e.g. customer load profiles with a max demand load of 1 kW in a 24-hour period). The final set of branches (with no further cutting positions) makes up the clusters.

Hierarchical clustering does not require user input on the number of clusters and can be represented by a dendrogram that allows for easy interpretability of its process. Despite these benefits, among the literature reviewed for this report, it is less common for customer classification than methods like k means or fuzzy c-means clustering. One of the reasons for this is that hierarchical clustering is often slower, with comparable accuracy to these other clustering methods (Singh and Singh 2012). An example of this is shown in the paper "Performance Evaluation of K-Means and Hierarchical Clustering in Terms of Accuracy and Running Time" (Singh and Singh 2012). This paper directly compares the effectiveness of k-means clustering and hierarchical clustering on two datasets with known clustering from the University of

California Irvine Machine Learning Repository: “Iris” and “Diabetes.” Descriptive information on each dataset is presented in Table 2 with the results of the clustering comparison in Table 3.

Table 2. Descriptive information on the "Iris" and "Diabetes" datasets (Singh and Singh 2012).

Datasets	Number of Attribute	Number of records/Instances
IRIS	05	150
DIABETES	09	768

Table 3. Metrics for comparing k-means clustering and hierarchical clustering on the "Iris" and "Diabetes" datasets (Singh and Singh 2012).

Datasets	k-means running time(sec)	Hierarchical clustering running time	k-means Accuracy %	Hierarchical clustering Accuracy %
IRIS	0.03	0.17	88.667	66
DIABETES	0.06	2.14	51.6927	65.1042

From these examples, we see that hierarchical clustering is slower than k-means, and even significantly slower in the case of the “Diabetes” dataset. Accuracy of k-means fluctuates more than hierarchical clustering depending on the dataset. This illustrates another reason why hierarchical clustering still has a purpose in clustering of datasets. K-means is a good first method to apply due to its speed, but if the time and resources are available, hierarchical clustering may even improve on k-means in some cases.

1.3.4 Self-Organizing Map

The self-organizing map (SOM) differs from the other clustering methods discussed here in that it is a type of artificial neural network. Self-organizing maps are particularly useful for the problem of customer classification because this requires simplifying a large amount of high dimensional data into a relatively small amount of discrete groups. Electric grid customer data can be considered to have a high amount of dimensionality when we’re looking at a large dataset where each customer is a single observation. For each of these observations, we may have either a demand load profile or an energy consumption profile, usually on a 15-minute interval basis. For a demand load profile over a 24-hour period, each of these data points can be considered to be a separate variable for the observation (customer). This complexity is compounded when we add identifying characteristics like building type, square footage, etc. to the customer. Self-

organizing maps can condense the demand load profile or energy consumption profile data into a low-dimensional map for faster characterization of customer types. For this reason, they are a good fit for processing energy use data from customers.

The paper "Application of Self-Organizing Maps for Classification and Filtering of Electrical Customer Load Patterns" (Verdu, et al. 2008) provides an example of this application. This study used a dataset consisting of residential, industrial, and university buildings. In total, 327 electrical demand profiles were included, ranging from 200 kW annual peak loads (for smaller facilities,) to customers with a 10 MW annual peak load. The demand profiles tracked load efficiency (how much of the power drawn from the grid was actually consumed,) and were also marked by date. This allowed for classification, not just by customer, but also based on use cases at particular times of the year. Some examples from this dataset are shown in Figure 6.

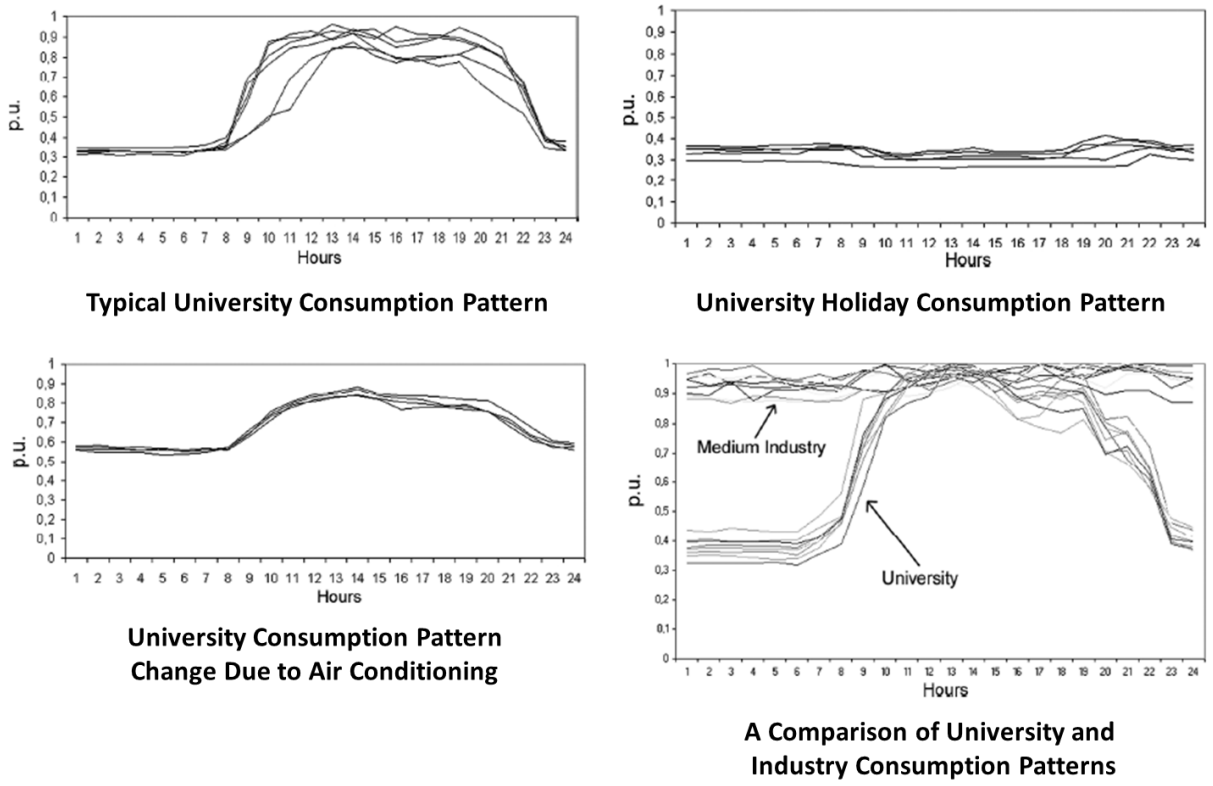

Figure 6. Examples of consumption patterns used in SOM analysis (profiles track load efficiency [p.u.] over a 24-hour period) (Verdu, et al. 2008).

A common visualization used for portraying the SOM method involves placing the observations on a hexagonal or square grid to show similarity between their energy consumption

profiles. This is shown in Figure 7. Twenty numbered labels were created representing different types of customer use cases and the results of SOM analysis on each customer's energy consumption determined their position on the grid. The borders drawn in Figure 7 represent zones of customers with similar characteristics in their energy consumption patterns. These borders can be drawn based on visual inspection, or through some other clustering method (e.g. k-means, etc.) (Verdu, et al. 2008). The paper concludes by suggesting that the customer segments defined by this SOM method could be used to evaluate the cost-effectiveness of demand-side energy management policies, and to evaluate the potential for incentivizing customers to install energy efficient technologies. The researchers also suggest that the customer segmentation produced by this SOM map could be further improved with the inclusion of an external parameter such as a customer's economic activity, or the quality and reliability of their energy supply.

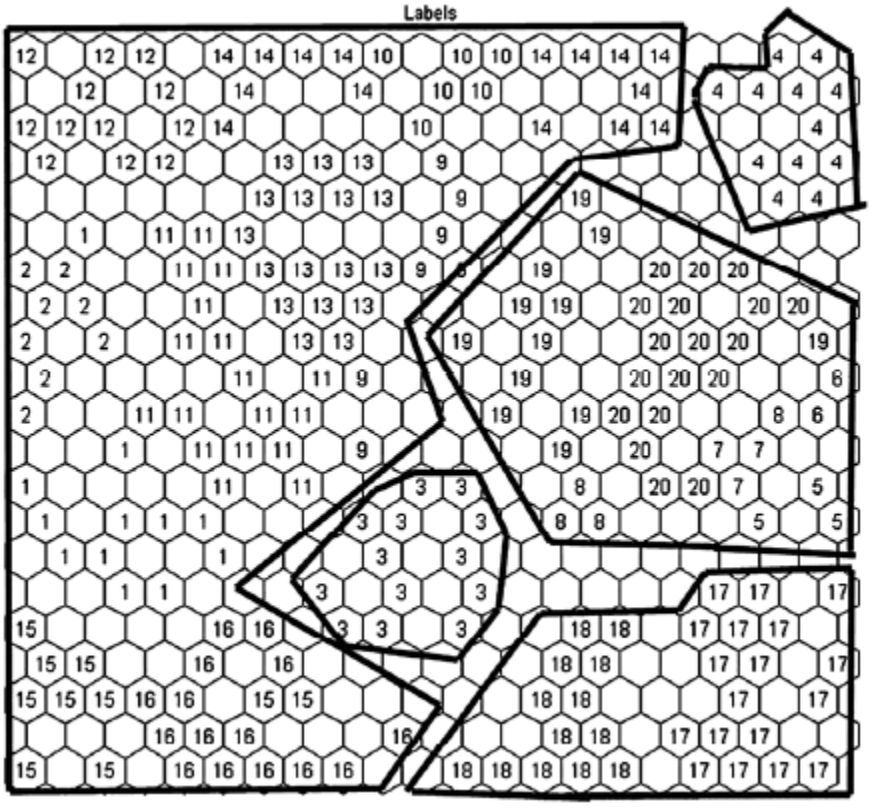

Figure 7. Customer labels marked on a hexagonal grid using SOM analysis. The black borders represent zones where customer energy consumption patterns show similar characteristics (Verdu, et al. 2008).

As shown in this example, SOM serves the purpose of dimension reduction as well as facilitating clustering. This was also the case in an example presented in section 1.3.1 K-Means Clustering. If we look back at the load profiling module from Figure 3 (depicted again in Figure 8, below), we see that a self-organizing map was used as a dimension reduction stage before the use of k-means for clustering. In this example, SOM was used to condense the customer's demand load profiles into bidimensional space (Figueiredo, et al. 2005). One reason for the use of SOM first is that it has good performance with very large datasets. Once the smaller, two-coordinate dataset is created, it becomes much more manageable for k-means clustering. In this way, multiple data mining methods can be used together to take advantage of their known strengths.

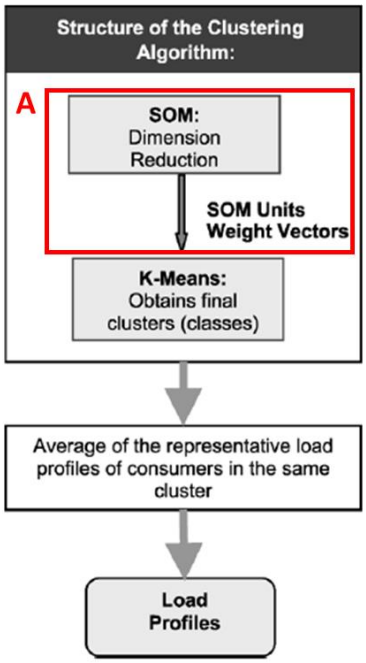

Figure 8. A load profiling module used for customer classification. Box A highlights the role of SOM as a dimension reduction step used in conjunction with k-means clustering (Figueiredo, et al. 2005).

1.3.5 Clustering Validity Assessment

As covered in this paper, there are many clustering methods that can be applied to customer classification in electric grids, and each has its own pros and cons that must be measured. In some cases, these comparisons would be qualitative. If the user is designing a utility program that requires specific classifications or parameters to be accounted for, they may

simply use the method that produces a result that is easiest to interpret in the context of the utility program. Alternatively, a strictly quantitative approach may be the best way to justify the chosen methodology.

The paper "Application of Clustering Technique to Electricity Customer Classification for Load Forecasting" (Wang, Li and Yang 2015) presents an example where k-means clustering and fuzzy c-means clustering (FCM) are applied to electricity consumption curves for a set of customers in southern China. As both of these methods require the user to provide the desired number of clusters, the researchers created a methodology for clustering validity assessment that compares the effectiveness of the clustering methods for a range of cluster numbers. The researchers used a distance-based correlation coefficient to approximate similarity within clusters, and calculated a value based on that which is referred to as the "index." They then calculated this index for both a k-means and a FCM clustering analysis of the customer data for numbers of clusters ranging from 3 to 10. The results of this analysis are shown in Table 4 and Figure 9.

Table 4. Index values for k-means clustering and FCM given numbers of clusters from 3 to 10. Lower index values represent more similarity within clusters (Wang, Li and Yang 2015).

Algorithms	The index value of the clustering results under different given number							
	3	4	5	6	7	8	9	10
k-means	0.51	0.45	0.38	0.32	0.40	0.38	0.43	0.45
FCM	0.48	0.47	0.42	0.38	0.37	0.35	0.45	0.48

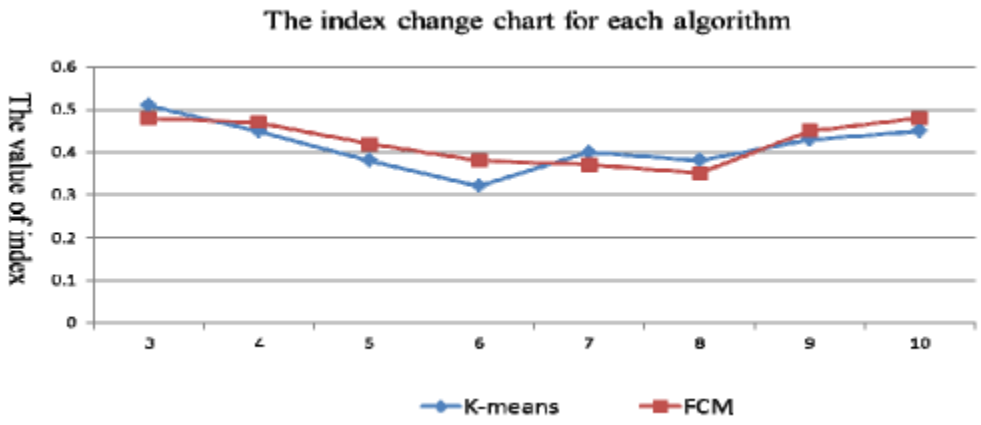

Figure 9. Plotted index values for k-means clustering and FCM given numbers of clusters from 3 to 10. Lower index values represent more similarity within clusters (Wang, Li and Yang 2015).

Lower index values are preferred because they indicate that there is more similarity within clusters, and that the data points are accurately grouped. This analysis shows that k-means where k = 6 is the best clustering operation because it has the lowest index value. We can see the results of this in Figures 10 and 11.

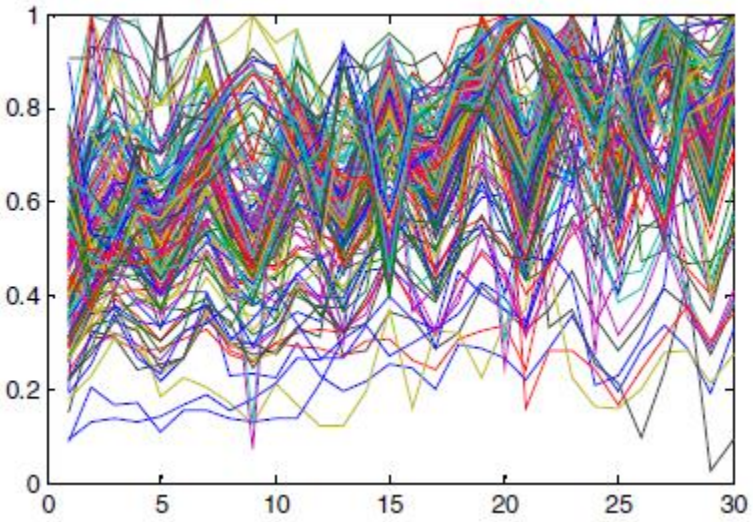

Figure 10. Electricity consumption curves for a set of customers in Southern China. This is before the clustering operation. Exact units were not provided but the description in the study states that these curves show electricity consumption (y-axis), over time (x-axis) (Wang, Li and Yang 2015).

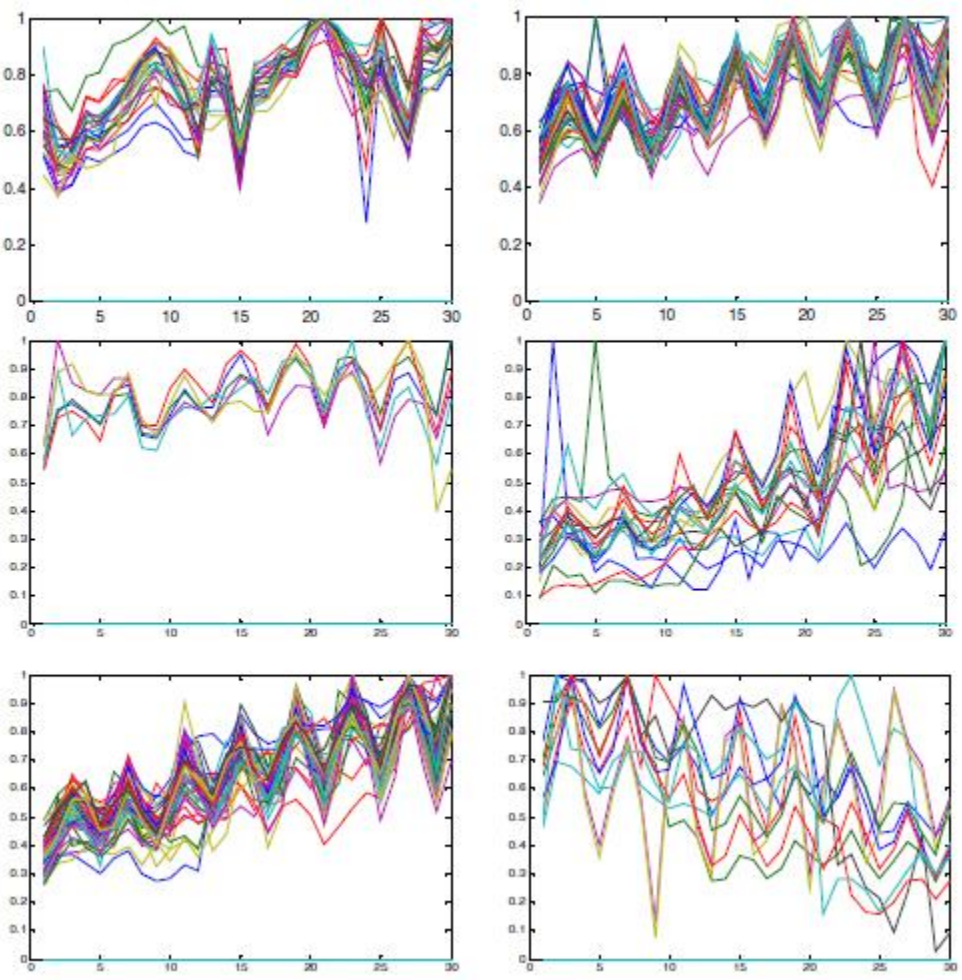

Figure 11. The results of k-means clustering (with k=6) on electricity consumption curves for a set of customers in Southern China. Exact units were not provided but the description in the study states that these curves show electricity consumption (y-axis), over time (x-axis) (Wang, Li and Yang 2015).

The plots in Figure 11 illustrate the purpose and intended end result of applying clustering methods to electricity consumption data from customers. Utilities or smart grid operators can use this information to design programs that better manage the energy demands of different customer types, and understanding the composition of electricity customers in a region can help the utility build out a more stable and efficient grid.

1.4 Conclusions and Future Work

The data mining methods covered in this paper demonstrate the diverse approaches that analysts can take for breaking down customer classification problems. To select the right method for the classification problem, data scientists must consider the size of the dataset, processing speed, how often the analysis is required, accuracy in customer classification with respect to the needs of the utility, and a many other factors.

K-means clustering is a commonly used algorithm with relatively fast processing speeds, but it requires the user to estimate the proper number of clusters to define. Its approach of condensing several data points to a single cluster centroid also results in a loss of information about the full dataset. As an alternative, fuzzy c-means clustering captures more information about the observations and in some cases results in a more accurate analysis, but can be slow when processing large datasets, and also requires the user to provide the final number of clusters to define. Hierarchical clustering has an advantage over k-means and fuzzy c-means clustering in that it can determine the number of clusters in a dataset independent of user input. It also produces cluster tree plots which can be useful for interpreting the customer classifications defined by the algorithm. However, researchers have demonstrated that it can be slow for processing large datasets.

Following the discussion on these algorithms, this paper looked at an alternative approach to clustering. The self-organizing map is a neural-network-based approach to dimension reduction that can condense high-dimensional datasets down to a form that can be processed by relatively slow clustering methods. This presents a solution to one of the major obstacles to effective electricity customer classification: that each customer has a unique daily load profile that can contain many data points to represent its daily energy use, and that a typical analysis may need to process hundreds or thousands of customers to be effective. Self-organizing maps give analysts a way to simplify all of this data, which opens up options to use other clustering or classification methods.

Finally, this paper presented an example of clustering validity assessment. With so many methods available for classification, analysts need ways to compare the effectiveness of different techniques. Developing an assessment index like the one described in section 1.3.5 presents one option for finding the best method for classifying a set of customers. The examples provided in

this paper show how well data mining methods suit the problems faced by utilities, but also show that there is no best way to classify customers in every case. Data scientists and analysts must develop a detailed understanding of the pros and cons of all of the tools available to them, and they must consider the limitations of their available computing power and data collected on customers.

This paper presents an introduction to the topic of customer classification for electric grids, and focuses on some of the more commonly-used classification techniques that use demand profile or energy consumption data. With smart grid technologies, the available data that utilities have on their users may increase dramatically, which may open up a number of additional methods for customer classification. The "internet of things" trend has provided ways for consumers to track energy use at a more granular level, and the expansion of commercially available distributed generation (primarily residential photovoltaic systems) means that consumers can have a more detailed look at the energy flowing into and out of their residence. To add to this, local energy storage in the form of electric vehicles is being promoted as a feature in many of the latest models (Amjadi 2010). All of these energy sources and their corresponding data, if made available to utilities, can open up opportunities for managing the flow of energy. For example, utilities might develop programs that tap into residential distributed energy generation or storage to reduce the midday peak demand problem, allowing utilities to reduce the power generation capacity that they must keep online. To take this one step further, forecasting techniques could provide added value to the customer classification methods outlined in this report. Once we understand the different types of energy users in a network, and once we can estimate their needs in the future, we can predict what tomorrow's energy demand will be and develop incentive structures to manage it efficiently. As topics of future work, it would be worthwhile to research data mining methods that can incorporate energy data from customer-owned distributed generation or storage systems, as well as forecasting techniques that can be used to estimate energy demand at some point in the future. These challenges, as well as customer classification, involve millions of data sources, and as such, data mining is a field uniquely suited to solving these problems.

2. References

American Public Power Association. 2016. "U.S. Electric Utility Industry Statistics." *publicpower.org.* Accessed June 2016. http://www.publicpower.org/files/PDFs/USElectricUtilityIndustryStatistics.pdf.

Amjadi, Zahra. 2010. "Power-Electronics-Based Solutions for Plug-in Hybrid Electric Vehicle Energy Storage and Management Systems." *IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS* 57 (2): 608-616. http://ieeexplore.ieee.org.proxygw.wrlc.org/xpls/abs_all.jsp?arnumber=5256209&tag=1.

Arif, Mohammad Taufiqul, Amanullah M. T. Oo, and A. B. M. Shawkat Ali. 2013. "Estimation of Energy Storage and Its Feasibility Analysis." *Intech* Chapter 2, Page 47. http://www.intechopen.com/books/energy-storage-technologies-and-applications/estimation-of-energy-storage-and-its-feasibility-analysis.

Figueiredo, Vera, Fatima Rodrigues, Zita Vale, and Joaquim Borges Gouveia. 2005. "An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques." *IEEE Transactions on Power Systems* 20 (2): 596-602. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1425550&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F59%2F30784%2F01425550.

Ghosh, Soumi, and Sanjay Kumar Dubey. 2013. "Comparative Analysis of K-Means and Fuzzy C-Means Means Algorithms." *International Journal of Advanced Computer Science and Applications (IJACSA)* 4 (4): 35-39. https://thesai.org/Downloads/Volume4No4/Paper_6-Comparative_Analysis_of_K-Means_and_Fuzzy_C_Means_Algorithms.pdf.

Kim, Y., A. Aravkin, and A. Zondervan. 2016. "Analytics for understanding customer behavior in the energy and utility industry." *IBM Journal of Research and Development* 11:1-11:13. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7384570&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D7384570.

Kim, Young-Il, Jong-Min Ko, and Seung-Hwan Choi. 2011. "Methods for generating TLPs (typical load profiles) for smart grid-based energy programs." *2011 IEEE Symposium on*

Computational Intelligence Applications In Smart Grid (CIASG). http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?tp=&arnumber=5953331&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F5940251%2F5953191%2F05953331.pdf%3Farnumber%3D5953331.

Pentland, William. 2013. "Stress Testing Solar PV Growth Forecasts." *Forbes Energy*, May 5. http://www.forbes.com/sites/williampentland/2013/05/05/stress-testing-solar-pv-growth-forecasts/#2be698972327.

Poullikkas, Andreas. 2013. "A comparative assessment of net metering and feed in tariff schemes for residential PV systems." *Sustainable Energy Technologies and Assessments* 1-8. https://www.researchgate.net/publication/257745865_A_comparative_assessment_of_net_metering_and_feed_in_tariff_schemes_for_residential_PV_systems.

Singh, Nidhi, and Divakar Singh. 2012. "Performance Evaluation of K-Means and Hierarchical Clustering in Terms of Accuracy and Running Time." *International Journal of Computer Science and Information Technologies* 4119-4121. http://www.ijcsit.com/docs/Volume%203/vol3Issue3/ijcsit2012030357.pdf.

Steinbach, Michael, Levent Ertoz, and Vipin Kumar. 2004. "The Challenges of Clustering High Dimensional Data." *New Directions in Statistical Physics.* http://www-users.cs.umn.edu/~kumar/papers/high_dim_clustering_19.pdf.

Verdu, S. V., M. O. Garcia, C. S. Blanes, F. J.G. Franco, and A. G. Marin. 2008. "Application of Self-Organizing Maps for Classification and Filtering of Electrical Customer Load Patterns." *International Journal of Power and Energy Systems* 28 (1): 83-90. https://www.researchgate.net/publication/228715103_Application_of_self-organizing_maps_for_classification_and_filtering_of_electrical_customer_load_patterns.

Wang, Yanlong, Li Li, and Qinmin Yang. 2015. "Application of Clustering Technique to Electricity Customer Classification for Load Forecasting." *Proceeding of the 2015 IEEE International Conference on Information and Automation* 1425-1430. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7279510&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D7279510.

Zakaria, Zuhaina, Mohd Najib Othman, and Mohamad Hadi Sohod. 2006. "Consumer Load Profiling Using Fuzzy Clustering and Statistical Approach." *4th Student Conference on Research and Development (SCOReD 2006)* 270-274. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4339352&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4339352.