A Probabilistic Load Modelling Approach using Clustering Algorithms

M.S. ElNozahy, *Student Member, IEEE,* M.M.A.Salama, *Fellow, IEEE,* and R. Seethapathy, *Senior Member, IEEE*

Abstract**-- In this paper, a novel probabilistic load modeling approach is presented. The proposed approach starts by grouping the 24 data points representing the hourly loading of each day in one data segment. The resulting 365 data segments representing the whole year loading profile are evaluated for similarities using principle component analysis; then segments with similar principal components are grouped together into one cluster using clustering algorithms. For each cluster a representative segment is selected and its probability of occurrence is computed.**

The results of the proposed algorithm can be used in different studies to model the long term behavior of electrical loads taking into account their temporal variations. This feature is possible as the selected representative segments cover the whole year. The designated representative segments are assigned probabilistic indices that correspond to their frequency of occurrence, thus preserving the stochastic nature of electrical loads.

Index Terms**—Clustering algorithms, principal component analysis, probabilistic load modeling, validity indices.**

I. INTRODUCTION

IN power systems studies, electrical loads are usually represented using either deterministic or probabilistic models. In the deterministic modeling approach, system loads are modeled as constant power sinks; this constant power can be: i) the average daily, monthly or seasonal load demand curves obtained from the utility historical data for the area under study [1]. ii) the worst case scenario for the study under consideration such as when the load demand is minimal and a distributed energy resource is generating its maximum output [2]; iii) constant values that correspond to different loading conditions (peak, average and minimum loading) [3, 4]. Deterministic models study only predetermined situations; thus, they are not suitable for assessing the long term behavior of loads or modeling their random behavior. Such random behavior greatly impacts the performance of power networks.

In the probabilistic modeling approach, electrical loads are modeled as random variables that follow pre-determined probability density functions (pdfs). In [5], the load is modeled using a beta pdf, and the resulting cumulative distribution function (cdf) is discretized into equal steps. Finally, the probability of each step is computed and used in the random generation of different load levels. In [6], the authors used the k-means clustering algorithm to reduce the number of steps of the cdf. These models have the disadvantage that generation of the hourly load levels takes place in a memoryless random fashion that does not consider the chronological nature of load variations; the value of the load level generated at hour i depends only on the value of the random variable generated at hour i and the pre-determined pdf representing the load. This situation results in irregular daily load curves with load spikes; whereas in reality, the load level at hour i has some correlation with the preceding load levels (i-1, i-2, etc.). This correlation is justified by the fact that the aggregated loading at each transformer or substation does not experience any sudden spikes; nevertheless, this loading changes in a smooth manner. Another disadvantage of the previous modeling approach is that the consequential load levels are not synchronized with the rest of the system parameters; it may happen that peak load levels are randomly generated during an off-peak period. The previous drawback was resolved in [7, 8] by using a different pdf to represent the load level for each hour of the day during each month of the year; however, the huge computational burden associated with computing the parameters for 288 different pdfs to generate random variables renders this solution practically infeasible.

From the previous discussion, it becomes evident that a robust load modeling technique should satisfy the following requirements: i) can be used to assess the long term behavior of electrical loads; ii) should have a minimal computational burden to be practically feasible; iii) should consider the stochastic nature of system loads.

In this paper, a probabilistic load modeling approach that fulfills all the previously mentioned requirements is proposed. The rest of the paper is organized as follows: section II presents the different stages of the proposed load modeling approach, section III includes the results and finally, section IV concludes the paper.

II. DESCRIPTION OF THE PROPOSED APPROACH

The proposed load modeling approach starts by grouping the 24 data points representing the loading conditions during a certain day in a data segment. The resulting 365 data segments are evaluated for similarities, using principle component

M.S. ElNozahy and M.M.A. Salama are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, N2L 3G1, Canada (e-mail: mnozahy@uwaterloo.ca). Dr. Salama is also a Visiting Professor at King Saud University, Saudi Arabia.
R. Seethapathy is with Hydro One Network Inc. Toronto, On, Canada.

978-1-4799-1303-9/13/$31.00 ©2013 IEEE

analysis, and similar segments are grouped together into the same cluster using clustering techniques. For each cluster a representative segment is selected and its probability of occurrence is computed. Different stages of the proposed algorithm are explained in the following sections.

A. Data collection stage

In this research, the load profile being studied is adopted from the IEEE-RTS system presented in [9]. This system provides the hourly peak load as a percentage of the annual peak load. The provided data is used to form the load percentage matrix P (365 days x 24 data point/day); this matrix is used in the rest of the analysis.

B. Data pre-processing stage

In this stage, the annual data set is processed using principal component analysis (PCA) to extract the most important features of each day. PCA is a feature extraction tool that is used to extract relevant information from a confusing data set. They can also reduce a complex data set to a lower dimension one while retaining, as much as possible, the variation present in the data to reveal the hidden structures that underlie it and filter out the noise [10]. This reduction is simply achieved by means of an orthogonal linear transformation that is used to re-express the correlated data set in terms of new (and hopefully) more meaningful bases where the components of the transformed data are uncorrelated. The first coordinate in the new system coincides with the direction of the greatest variance of the original data and is defined as the first principal component. The second coordinate (which is orthogonal to the first) lies in the direction of the second greatest variation of the data, and so on. The number of selected principal components was chosen such as to maintain 90% of the variance within the data. The results reveal that the required variance can be maintained keeping only the first principal component, as shown in Table 1.

TABLE I
PERCENTAGE VARIANCE MAINTAINED AFTER APPLYING PCA

Number of principal components	1	2	3	4	5
% of variance kept within the data	93.1	96.2	98.5	99.8	99.9

PCA is used in this research to perform three functions:

i) Extracting the important features (with the greatest variance) of each day,

ii) Reducing the noise in the data set,

iii) Reducing the dimensionality of the data set, by neglecting the higher principle components.

Reducing the dimensionality of the data set is of great importance in clustering applications; for example, in hierarchical clustering algorithms the time complexity is at minimum of the order $O(n^2)$ [11], where n is the number of total objects. Therefore, by applying PCA, the total number of objects is reduced from 8760 (365 days x 24 data point/day) to 365 (365 days x 1 data point/day). This reduction corresponds to 576 times reduction in the computation time (and burden).

C. Data clustering stage

In this stage, days with similar principle components are grouped into one cluster and treated as one unit. Later on in the representative selection stage, only one representative segment will be chosen to represent the whole cluster.

Clustering algorithms are broadly classified into exclusive and overlapping algorithms as shown in Figure 1. Exclusive clustering algorithms are those in which each data segment belongs to only one cluster, whereas in overlapping clustering (also known as fuzzy c-means clustering) each data segment may belong to more than one cluster with different degrees of membership. Exclusive clustering can be further classified into hierarchical and partitional clustering. Partitional clustering directly divides data segments into a pre-determined number of clusters without building a hierarchical structure, whereas hierarchical clustering seeks to build a hierarchy of clusters with a sequence of nested partitions, either from singleton clusters to a cluster including all data segments or vice versa. The former is known as agglomerative hierarchical clustering, and the latter is called divisive hierarchical clustering. Divisive clustering algorithms need to compute (2^n – 1) possible divisions for a cluster with n data points, which is very computationally intensive [11]. Therefore, agglomerative methods are usually preferred, and only they are included in this research. Partitional clustering can be classified into the famous k-means clustering algorithm and the model-based clustering (also known as probabilistic clustering or a mixture of Gaussians clustering). In model-based clustering, each cluster can be mathematically represented by a parametric distribution, like Gaussian (continuous) or Poisson (discrete). The entire data segments are therefore modelled by a mixture of these distributions. The probabilistic clustering algorithm seeks to optimize the parameters of the mixture model so as to "cover" the data segments as much as possible. This is a very computationally intensive process and, thus model-based clustering is not considered in this research. In this research, the following clustering techniques will be considered:

i) K-means,
ii) Fuzzy C-means (FCM),
iii) Hierarchical.

The following classes of hierarchical clustering are considered

- Single linkage,
- Complete linkage,
- Average linkage,
- Ward's linkage.

In the results evaluation stage, the performances of different clustering techniques are compared to select the most accurate technique based on a proposed performance index.

D. Representative selection stage

The goal of this stage is to select a representative data segment for each cluster to represent all the data segments within the cluster. The representative data segments can be selected in two different ways:

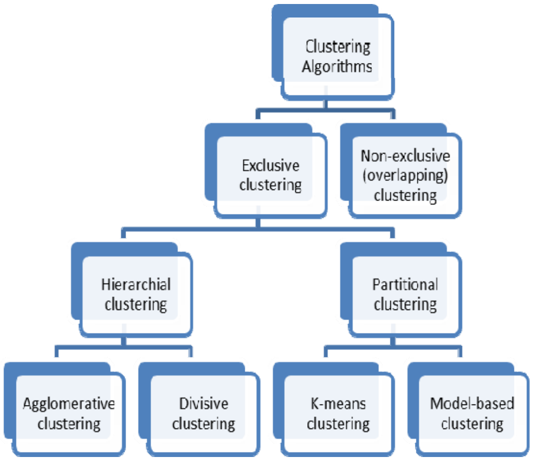

Fig. 1. Classification of clustering techniques

i) The mean representative: In this method, the representative data segment is the one formed by the calculated mean (barycenter) of all segments in the cluster.

ii) The median representative: the representative data segment is the one that is originally included in the cluster and is the closest to the calculated centroid (the median).

In the results evaluation stage, the performances of the two alternatives are compared to select the most accurate method.

E. *Probability evaluation stage*

A question arises here: are all the data representatives equally important? In other words, how can we reflect the probability of occurrence of different data representatives? To answer this question, we consider this example with two data clusters, cluster A having x data segments, and cluster B having 2x data segments. It is clear that the data representative representing cluster B is twice as important as the one representing cluster A as the former represents double the number of segments and so should be repeated twice as often. Thus, the frequency of occurrence of a certain data representative (and hence its relative importance) depends on the number of segments within the cluster it is representing with respect to the total number of data segments. For this reason, a Representative Probability Index (RPI) is proposed in this research and is given as:

$$RPI_i = \frac{n_i}{\sum_{j=1}^{k} n_j} \tag{1}$$

Where:
RPI_i: Representative probability index for the data segment representing cluster i.
n_i: Number of data segments within cluster i.
n_j: Number of data segments within cluster j.
k: total number of clusters.

The probabilistic RPI probability index represents the probability of different data representatives.

F. *Results evaluation stage*

In the previous stages, it is noticeable that different alternatives are possible for the clustering process. These alternatives should be compared for the best alternative. In this stage, we try to find an answer for the following questions:

- What clustering algorithm yields the best results? And what is the best alternative when selecting the cluster representative?
- What is the optimal number of clusters to be used with each clustering algorithm?

To answer these questions, we should have some kind of criteria or indices that can compare different clustering alternatives. Such indices are called cluster validity indices and can be broadly classified into the following types [12]:

i) External (supervised) validity indices: These indices evaluate the clustering based on some user-specific intuition, and thus they require the knowledge of external information about the data.

ii) Internal (Unsupervised) validity indices: These are the most widely used validity indices. They evaluate the clustering based on some metrics within the resulting clustering schema itself. This purpose is usually achieved by measuring the cohesion within each cluster and the separation between different clusters. Examples of these indices are the Dunn index [13], the Davies Bouldin (DB) Index [14], the Average Partition Density (APD) index [15] and the Xie - Beni (XB) index [16].

iii) Relative validity indices: Such indices use either internal or external indices to compare different clustering structures obtained from the application of different clustering algorithms or by applying the same algorithm but with different parameters. The aim of the relative criteria is to choose the best clustering schema for a certain application.

It is obvious that since the best clustering algorithm is to be chosen from different alternatives, relative validity indices are the most suitable indices to achieve that goal; however, another question arises: should we employ internal or external validity indices in the relative comparison? Internal validity indices measure the quality of clustering in terms of compactness of each cluster and the separation between different clusters. The compactness of each cluster is mainly evaluated by calculating the distance between the data segments or the distance between the segments and the centroid of the cluster. Consequently, these indices tend to prefer the clustering algorithm that produces more clusters containing single segments because, for these clusters (singletons), the compactness is zero. This preference does not serve the main purpose of the proposed algorithm, which is to group days with similar PV output profiles together. Another disadvantage of internal validity indices is that there is no single index that can be used to compare all clustering algorithms; for example, the famous Xie - Beni (XB) index can only be used with fuzzy clustering algorithms. For these reasons, external validity indices will be used in the relative comparison between different clustering algorithms.

As the purpose of this platform is to group segments with similar profiles together in the same cluster, the index used should be able to express the power/time mismatch between the original segments included in each cluster and the representative segment for the whole cluster. An average power/time mismatch (APTM) index is proposed here; the steps for calculating the proposed index are as follows:

i) For the n power segments included in each cluster, calculate the cluster representative.

ii) Form an annual fictitious power vector R (365 days × 24 data point/day) by replacing each data segment with the representative segment of the cluster to which this data segment belongs.

iii) Calculate the APTM index as follows:

$$APTM = \frac{\sum_{i=1}^{24}\sum_{j=1}^{365}\left|\frac{P(i,j)-R(i,j)}{P(i,j)}\right|}{24\times 365} \qquad (2)$$

Where:
APTM: Average power time mismatch index.
P (i, j): AC Power output of the i^{th} hour at the j^{th} day in the original AC power vector.
R (i, j): AC Power output of the i^{th} hour at the j^{th} day in the fictitious AC power vector.

The proposed index has the advantage that it considers not only the power mismatch between the original power segments within a certain cluster and their corresponding representative segment, but also considers this power mismatch in a chronological manner, i.e., Power/Time mismatch. For example, if the APTM index for a certain clustering alternative is 10%, it is possible to say that the resulting representative segments, if considered with their corresponding Representative Probability Indices (RPIs), are able to represent the complete data set with an average error of 10% per data point.

Thus, for each clustering alternative, the APTM index will be calculated for different numbers of clusters, and the best clustering alternative is the one that requires the least number of clusters to satisfy a minimum APTM threshold.

III. RESULTS OF THE PROPOSED ALGORITHM

A. Selecting the best clustering alternative

In this section, the performance of different clustering algorithms is evaluated. For each clustering algorithm, the performance index (APTM index) should be computed for two cases:

- Using the mean representative,
- Using the median representative.

i) K-means clustering: To overcome the initialization problem of the k-means algorithm, each case is repeated 10 different times and the best run is taken. Results are shown in Figure 2.

ii) Fuzzy C-means clustering: In these techniques, the data segment belongs to more than one cluster with different degrees of membership; however, in order to calculate the APTM index, each data segment should be included in only one cluster. Thus, each data segment is assigned only to the cluster to which it belongs with the highest degree of membership (Maximum membership rule [17]). The results of applying FCM clustering are shown in Figure 3.

iii) Hierarchical clustering: The performance of different kinds of hierarchical clustering is evaluated. It was found that ward’s linkage hierarchical clustering has the least APTM index for any number of clusters. Results of applying ward’s linkage clustering to the data set are shown in Figure 4.

iv) Conclusion: From these figures, it is clear that k-means clustering using median representatives yields the best results as it has the smallest APTM index among all clustering techniques for any number of clusters.

B. Selecting the optimal number of clusters

Selecting the optimal number of clusters is a famous dilemma in clustering applications: choosing a large number of clusters ensures that only a small number of segments (which are really similar to one another) are grouped together within the same cluster; however, the data dimensionality is not greatly reduced. On the other hand, selecting a very small number of clusters ensures that the data dimensionality is reduced; however, it may happen that some data segments which are not really similar to one another are grouped together. Thus the reduced data set does not represent the original data set accurately. For such reasons, the optimal number of clusters is never known apriori and is determined based on some kind of compromise between the computational burden reduction and accuracy.

In this research, the optimal number of clusters is selected based on the proposed APTM index. Previous results show that the APTM index for different clustering algorithms decreases rapidly as the number of clusters increases, and then it starts to saturate at around 4%. It can thus be deduced that for 4% average error allowance in the data set under study, k-means clustering with median representatives requires the least number of clusters (6 clusters) making it possible to represent the complete loading profile (365 days × 24 data points/day) using only six representative data segments (6 days × 24 data point/day), with an average error of 4%. The selected representative segments are shown in Figure 5, and their corresponding RPI indices are given in Table 2.

TABLE 2
PERCENTAGE RPI INDICES FOR THE SIX REPRESENTATIVE DATA SEGMENTS

Data Segment	1	2	3	4	5	6
RPI index	11.8	13.2	12.6	23.6	16.4	22.4

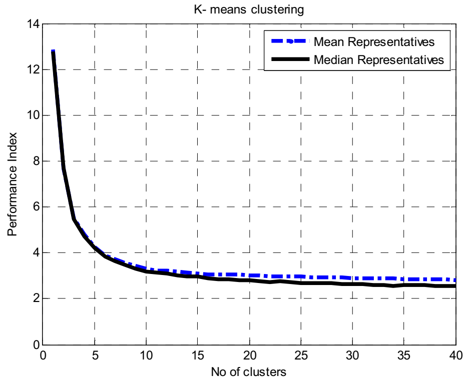

Fig. 2. K-means clustering results

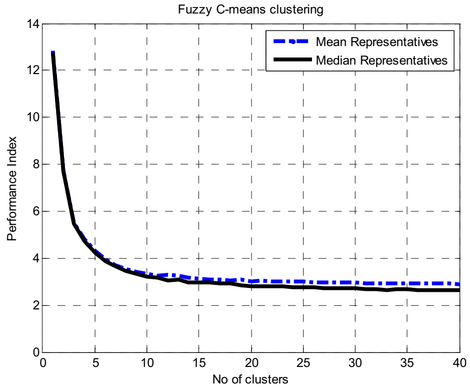

Fig. 3. FCM clustering results

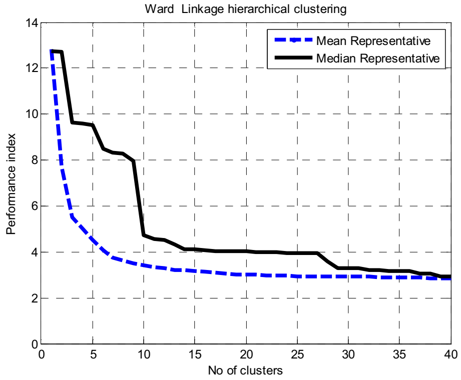

Fig. 4. Ward's Linkage hierarchical clustering results

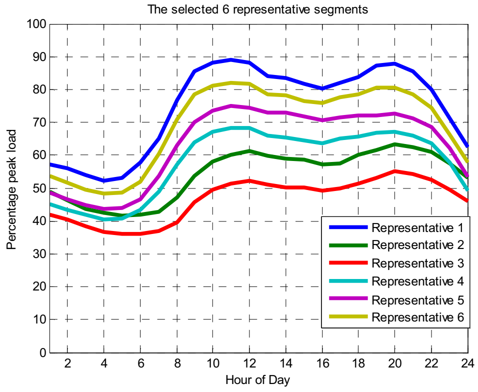

Fig. 5. The representative data segments

IV. Conclusions

This paper presents a probabilistic modeling approach for electrical loads extracted from the IEEE-RTS system. The proposed model avoids many of the drawbacks of the models described in the literature; unlike the deterministic modeling approaches, the proposed model can mimic the long-term behaviour of electrical loads because the representative data segments cover the whole year. Probabilistic modeling approaches described in the literature generate hourly load levels in a memoryless fashion, resulting in irregular load curves with spikes. Using daily segments in the proposed model generates smooth daily load curves without any spikes, thereby, maintaining the chronological nature of the loading profile.

References

[1] W. A. Omran, M. Kazerani, and M. M. A. Salama, "A Clustering-Based Method for Quantifying the Effects of Large On-Grid PV Systems," Power Delivery, IEEE Transactions on, vol. 25, pp. 2617-2625, 2010.

[2] H. Liu, L. Jin, D. Le, and A. Chowdhury, "Impact of high penetration of solar photovoltaic generation on power system small signal stability," 2010, pp. 1-7.

[3] A. F. Povlsen, "Impacts of power penetration from photovoltaic power systems in distribution networks," International Energy Agency, February 2002.

[4] K. Myers, S. Klein, and D. Reindl, "Assessment of High Penetration of Photovoltaics on Peak Demand and Annual Energy Use," Public Service Commission of Wisconsin January 2010.

[5] S. W. Heunis and R. Herman, "A probabilistic model for residential consumer loads," Power Systems, IEEE Transactions on, vol. 17, pp. 621-625, 2002.

[6] C. Singh and Q. Chen, "Generation system reliability evaluation using a cluster based load model," Power Systems, IEEE Transactions on, vol. 4, pp. 102-107, 1989.

[7] J. A. Jardini, C. M. V. Tahan, M. R. Gouvea, S. U. Ahn, and F. M. Figueiredo, "Daily load profiles for residential, commercial and industrial low voltage consumers," Power Delivery, IEEE Transactions on, vol. 15, pp. 375-380, 2000.

[8] M. Espinoza, C. Joye, R. Belmans, and B. DeMoor, "Short-Term Load Forecasting, Profile Identification, and Customer Segmentation: A Methodology Based on Periodic Time Series," Power Systems, IEEE Transactions on, vol. 20, pp. 1622-1630, 2005.

[9] C. Grigg, P. Wong, P. Albrecht, R. Allan, M. Bhavaraju, R. Billinton, Q. Chen, C. Fong, S. Haddad, and S. Kuruganty, "The IEEE reliability test system-1996. A report prepared by the reliability test system task force of the application of probability methods subcommittee," Power Systems, IEEE Transactions on, vol. 14, pp. 1010-1020, 1999.

[10] Shlens, Jonathon. "A tutorial on principal component analysis." Systems Neurobiology Laboratory, University of California at San Diego (2005).

[11] R. Xu and D. C. Wunsch, Clustering: Wiley-IEEE Press, 2009.

[12] K. Ferenc, L. Csaba, and B. Attila, "Cluster validity measurement techniques," 2005.

[13] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," 1973.

[14] D. L. Davies and D. W. Bouldin, "A cluster separation measure," Pattern Analysis and Machine Intelligence, IEEE Transactions on, pp. 224-227, 1979.

[15] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 11, pp. 773-780, 1989.

[16] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 13, pp. 841-847, 1991.

[17] S. Miyamoto, H. Ichihashi, and K. Honda, Algorithms for fuzzy clustering: methods in c-means clustering with applications vol. 229: Springer Verlag, 2008.