Clustering and Visualization of Electrical Load Profiles using Data Mining Techniques

Ana Aréchiga, E. Barocio, J. J. Ayon
Graduate Program in Electrical Engineering
University of Guadalajara
Guadalajara, Jal. México

H. A. Garcia-Baleon
Data Center Group
INTEL ZPN - Guadalajara Design Center
Zapopan, Jalisco

***Abstract*— Due to the recent deregulated electricity market and the increasing consumption of electrical energy; new possibilities to the electricity suppliers to formulate tariff offers and improve the quality services have been opening to satisfy the daily demand. A key aspect to visualize and interpret a huge volume of energy data is to cluster customers according to their individual electrical load profile similarities. For this purpose, it is proposed to apply dimensionality reduction techniques, such as PCA, Isomap, Sammon Mapping, LLE and SNE. To evaluate the performance of these techniques, a test IEEE 30-bus system is used in this research.**

***Index Terms*—Clustering, Isomap, LLE, load profile, Sammon Mapping, PCA, SNE, visualization.**

I. Introduction

The electrical distribution systems are increasing every day in complexity and scale. As a part of its control and operation, thousands of busses are monitored daily, generating huge volumes of data. From this point, the obtained measurements may be exploited to visualize and identify load profiles that define the consumption behavior of every customer. In addition, the customer profiles, the electrical infrastructure usage and the new tariff offers may be used for planning studies or for operational activities.

Clustering the different types of customers based on groups that display a similar electrical consumption behavior is a key factor to obtain a better understanding of the information that the electrical distribution systems provide. As a result, the knowledge of the electrical consumption allows optimizing the daily demand, making a better balance of the loads, accuracy predictions that improve the quality service, thus achieving to reduce costs levels.

Due to the complexity and the huge volumes of data that are associated to the load profiles characterized by nonlinear and non-oscillatory behavior, it results difficult to visualize and identify patterns, outliers and abnormal behaviors. Different techniques have been used with the objective of organizing, visualizing, and clustering this information. Amongst them, clustering techniques that have been used are: k-means [1], hierarchical algorithms [2], Fuzzy k-means [3] and Follow the leader [4]. Every one of these techniques allows to extract the implicit knowledge of the energy data, as well as the identification of different types of customers. However, clustering techniques are sensitive to outliers, and they present visualization difficulties when their representations is based on dendrograms, especially when the data volume grows or there are groups with different size and densities. Due to these disadvantages, dimensionality reduction techniques are currently being used in order to overcome the difficulties that clustering techniques present. One of the first techniques used for clustering a huge volume of data, related to load profiles, was Principal Component Analysis (PCA) (see, e.g. [5], 6]). However, this technique cannot make a proper clustering when works on nonlinear data. Therefore, in the last decades, a growing number of nonlinear techniques has been developed, namely, Isometric Feature Mapping (Isomap), Sammon mapping, and Locally Linear Embedding (LLE). These techniques allow to obtain better results when it comes to visualization and clustering of load profiles, as reported in [7] and [8]. Considering the above, this paper proposes using a new nonlinear dimensionality reduction technique to cluster load profiles, based on probabilities, named: Stochastic Neighbor Embedding (SNE) [9].

Dimensionality reduction techniques allow clustering a set of data following optimization strategies through minimization or maximization of a cost function. The visualization of the reduced data is based on a dispersion graphic, where every load profile is represented by a point, and these points are grouped by their similarities.

II. Spatio-temporal load profile model

Mathematically, the spatio-temporal load profile model can be defined as follows: Assume a set of data in a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, constituted by n vectors $\boldsymbol{x}_i$ that come from the load profile measurement of the electrical distribution system, in which $\boldsymbol{x}_i$ denotes an observation vector, and i denotes the measurement point:

$$\boldsymbol{x}_i = \{x_1, x_2, \dots, x_m\} \tag{1}$$

This way $\boldsymbol{x}_i$ is a variable of interest at each instant of time m. From this point, a set of observation data is created in a n vectors matrix:

$$\boldsymbol{X} = \{\boldsymbol{x}_1\ \boldsymbol{x}_2\ \boldsymbol{x}_3 \dots \boldsymbol{x}_j \dots \boldsymbol{x}_n\}^T \tag{2}$$

where $\boldsymbol{X}$ is a rectangular matrix that contains space-time information of the electrical system, and n represents the location of the measurements. Also assume that this set of data $\boldsymbol{X}$ can be transformed, through dimensionality reduction techniques, to a new system set of coordinates $\boldsymbol{Y} \in \mathbb{R}^{n\times d}$, formed by n data output points $\boldsymbol{y}_i$ for every vector $\boldsymbol{x}_i$:

$$\boldsymbol{Y} = \{\boldsymbol{y}_1\ \boldsymbol{y}_2\ \boldsymbol{y}_3 \dots \boldsymbol{y}_j \dots \boldsymbol{y}_n\}^T \tag{3}$$

where the $\boldsymbol{Y}$ matrix represents the most relevant information from the input matrix $\boldsymbol{X}$.

This work was supported by the University of Guadalajara under the scientific program namely the PRO-SNI 2016, that provide financial support for develop basic science.

III. DATA MINING TECHNIQUES

To reduce the input dimension data to a sub-space of dimension d, several techniques are tested in this paper. Data dimension may be reduced because it generally contains redundant or correlated information. For this purpose, the original set of data $\boldsymbol{X}$ is transformed to a new set $\boldsymbol{Y}$. Then, the clustering process is formed by a specific number of k clusters. Every cluster contains $n^{(k)}$ load profiles. The used techniques in this paper are: a classic linear technique PCA, and four nonlinear techniques that can be grouped into two types: those that preserve the data distances such Isomap and Sammon mapping, or those that preserve the data topology such as LLE and SNE.

A. Principal Component Analysis (PCA)

Principal component analysis is a linear technique that uses an optimization procedure to carry out the data dimensionality reduction. For this purpose, PCA performs an eigen-decomposition of the data covariance matrix $\boldsymbol{X}$, seeking for a maximum variance and minimum redundancy. This is achieved by transforming to a new set of variables $\boldsymbol{y}_i$ which are a linear combination of the original values $\boldsymbol{x}_i$. Being $\boldsymbol{y}_i$, the called principal components, that allow to explain, through very few variables, most part of the information that the input data contain.

In mathematics terms, PCA perform a linear projection $\boldsymbol{A}$ that maximizes the variance trough a cost function. This may be done by the n principal components of the covariance matrix $cov(\boldsymbol{X})$. This way, PCA solves the problem of eigenvalues as follows:

$$cov(\boldsymbol{X})\boldsymbol{A} = \boldsymbol{\lambda A} \tag{4}$$

where $\boldsymbol{A}$ is an orthogonal matrix which the nth column is the nth eigenvector that corresponds to the highest values $\boldsymbol{\lambda}$ from $cov(\boldsymbol{X})$. The space representation of the output $\boldsymbol{y}_i$ of $\boldsymbol{x}_i$ is calculated with a linear projection $\boldsymbol{Y} = \boldsymbol{XA}$. PCA is identical to the conventional Multidimensional Scaling (MDS) technique, in which the distance matrix $\boldsymbol{D}$ replaces the matrix $\boldsymbol{X}$, and where the inputs $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ represent the Euclidian distances of the inputs space. The conventional scaling technique finds a linear projection $\boldsymbol{A}$ that minimizes the cost function:

$$\min_{j \in k} \boldsymbol{C_{MDS}} = \sum_{i \neq j} \left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 - \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 \right) \tag{5}$$

where the vectors $\boldsymbol{y}_i$, $\boldsymbol{y}_j$ are subjected to be $\boldsymbol{x}_i\boldsymbol{A}$, and $\|\alpha_j\|^2 = 1$ for $\forall j$. And $\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2$ are the Euclidian distances of the output space.

B. Isometric Feature Mapping (Isomap)

Although PCA has been used in many applications successfully, it is not suitable for nonlinear data because its objective is to perform an orthogonal transformation of the data. As an alternative, the Isomap technique solves this problem using geodesics or curvilinear distances.

Isomap is a nonlinear dimensionality reduction technique that uses graphs distances as an approximation of geodesics distances. The current Isomap version is closely related with the MDS technique [10]. The main difference between these two, is the metrics used to calculate the distances between the data. While Euclidian metrics are purely lineal, Isomap has the capacity of nonlinear data measurements by using geodesic distances from Dijkstra's algorithms [11].

In order to calculate the output coordinates matrix $\boldsymbol{Y}$, Isomap follows these steps:

1. Build a graph with k neighbors and radius ε.
2. Weigh the graph through the Euclidian distances of each sample from $\boldsymbol{x}_i$.
3. Compute all pairwise graph distances with Dijkstra's algorithm, and store them in matrix $\boldsymbol{D}$.
4. Convert the distances matrix $\boldsymbol{D}$ into a Gram matrix $\boldsymbol{S}$.
5. Once the Gram matrix is known, compute the eigenvectors decomposition, same way as PCA $\boldsymbol{S} = \boldsymbol{A\lambda A^T}$.
6. Get the representation of $\boldsymbol{Y}$ calculating $\boldsymbol{X} = \boldsymbol{I\lambda A^T}$.

As PCA does, Isomap uses the same cost function in (5) to calculate the projection error, substituting $\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2$ for geodesic distances.

C. Sammon Mapping

Same as PCA and Isomap, Sammon Mapping is a distance preservation technique. However, Sammon Mapping has only one specific purpose, to reduce the dimension of a finite number of points, and it may be considered a PCA variant. Sammon adapts the cost function of the MDS techniques to a normalized cost function that is in function to the inputs space distances. This way, the cost function assigns a weigh that is approximately equal to retain each distance, and seeks the local data output structure to be as similar as the input data. Mathematically, Sammon minimizes the following cost function as follows:

$$\boldsymbol{C_{SMN}} = \frac{1}{\sum_{i<j}^{m} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|} \sum_{i<j}^{m} \frac{\left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| - \|\boldsymbol{y}_i - \boldsymbol{y}_j\|\right)^2}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|} \tag{6}$$

The minimization process of the cost function in (6) has to be done iterative by applying a standard optimization technique which is a variation of Newton's method, named quasi-Newton. This modification involves a Jacobian and Hessian matrices during the solution process. Further the calculation of the output space $\boldsymbol{Y}$ coordinates is written as follows:

$$\boldsymbol{y}_i^{(t+1)} = \boldsymbol{y}_i^{(t)} - \alpha \frac{\left(\frac{\partial \boldsymbol{C_{SMN}}}{\partial \boldsymbol{y}_i}\right)}{\left(\frac{\partial^2 \boldsymbol{C_{SMN}}}{\partial \boldsymbol{y}_i}\right)} \tag{7}$$

where α is a weigh factor between 0.3 and 0.4 values. The advantages of using this method are that it is computationally simple and it obtains results even for nonlinear sets of data, as long as they are not too complex. Sammon Mapping follows these steps:

1. Compute all pairwise distances $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in the input space.
2. Initialize the $\boldsymbol{y_i}$ data coordinates in the output space randomly or using the first principal components, same as PCA.
3. Compute the output data coordinates variation with the equation (7).
4. Update the output space coordinates.
5. Return to step 3 until the cost function $\boldsymbol{C_{SMN}}$ no longer decreases.

D. Locally Linear Embedding (LLE)

On the other side, LLE is a topology preservation technique. It is similar to Isomap in the way it builds the graphic representation of the data. In comparison to Isomap, LLE allows to preserve exclusively the properties of local data. LLE is a nonlinear technique that uses a transformation to preserve local angles, which is related to local distances. The purpose of using this technique is to replace each data point with a linear combination of neighboring data. The data geometry may be represented by linear coefficients that rebuild each data through their neighbor. The reconstruction error is measured by the following quadratic cost function:

$$\boldsymbol{e}(\boldsymbol{W}) = \sum_i \left\| \boldsymbol{x}_i - \sum_{j=1}^{m} w_{ij}\boldsymbol{x}_j \right\|^2 \tag{8}$$

where $\boldsymbol{W}$ contains the eigenvectors w_{ij}coefficients for the $\boldsymbol{x}_i$ reconstruction. In order to make the data projection, LLE assumes that the data has a geometric relation that can be preserved. This is why it is invariant to translations, rotations and up or down scaling when performs the data reduction. To this end, the output projection coordinates should minimize the following cost function:

$$\min_j \boldsymbol{C_{LLE}} = \sum_i \left\| \boldsymbol{y}_i - \sum_{j=1}^{n} w_{ij}\boldsymbol{y}_j \right\|^2 \tag{9}$$

$$\text{s.t. } \left\| y^{(n)} \right\|^2 = 1 \text{ for } \forall\, n$$

where $y^{(n)}$ represents the nth column to the solution for the $\boldsymbol{Y}$ matrix. LLE follows these steps:

1. For every input data x_i compute every k nearest neighbor.
2. Determine the $\boldsymbol{W}$ coefficients that minimize the data reconstruction starting from the equation (8) minimization.
3. Compute the $(\boldsymbol{I}-\boldsymbol{W})^T(\boldsymbol{I}-\boldsymbol{W})$ matrix eigenvalues obtained by solving (9). The output coordinates will be given by the matrix eigenvectors.

E. Stochastic Neighbor Embedding (SNE)

Unlike the previous techniques, Stochastic Neighbor Embedding is a probabilistic technique for dimensionality reduction. SNE doesn't try to preserve the distances between points, but it measures the probability of this points being neighbors. To do so, SNE starts by converting Euclidean distances into conditional probabilities p_{ij}, where p_{ij} marks the probability that $\boldsymbol{x}_i$ chooses $\boldsymbol{x}_j$ as its neighbor through a Gaussian function centered in the $\boldsymbol{x}_i$ point.

In SNE, the probabilities p_{ij} are stored in a $\boldsymbol{P}$ matrix. Subsequently, the output space coordinates, are calculated with a conditional probability q_{ij}, and they are stored in a $\boldsymbol{Q}$ matrix. Assuming a perfect representation, both $\boldsymbol{P}$ and $\boldsymbol{Q}$ matrices must be equal. Thus, SNE minimizes the differences between the probabilities $\boldsymbol{P}$ y $\boldsymbol{Q}$ using the Kullback-Leibler divergences, which are a natural way to measure the distance between two probability distributions. Further, the SNE minimize the next cost function:

$$\boldsymbol{C_{SNE}} = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{10}$$

The cost function minimization in (10), is carried out using a descendant gradient method. This method is initialized in a random way and updates mathematically with every iteration through the following equation:

$$\boldsymbol{y}_i^{(t)} = \boldsymbol{y}_i^{(t-1)} + \eta \frac{\delta \boldsymbol{C_{SNE}}}{\delta \boldsymbol{y_i}} \tag{11}$$

where $\boldsymbol{y}_i^{(t)}$ indicates the solution for the iteration t, and η indicates the learning rate. SNE makes the representation of the output space $\boldsymbol{Y}$, using a number selected by the user, which corresponds to a free parameter σ named $Perplexity$, related to the number of clusters k. *Perplexity* is defined as the Shannon entropy of $\boldsymbol{P}$.

IV. CASE STUDY AND DISCUSSION

In order to evaluate the dimensionality reduction techniques performance, a study that consists of the analysis of the standard IEEE 30-bus system data is developed in this section.

A. IEEE 30-bus system

The electrical system shown in Fig. 1, comprises 19 load busses. These busses form the input data set for matrix $\boldsymbol{X}$. The matrix contains the normalized load profiles that belong to different types of costumers. These are the characteristics that define each customer profile:

- Commercial: high consumption (25-35 kW/h), medium (6-15 kW/h) and low consumption (1-5 kW/h).
- Industrial: high consumption (600-800 kW/h), medium (0 or 1 kW/h), and low consumption (0.01-0.5 kW/h).
- Residential: high consumption (0.1-1.5 kW/h) and low consumption (0.1-0.5 kW/h).

The input data set denoted by $\boldsymbol{X} \in \mathbb{R}^{19x96}$, is displayed at Fig. 2. The consumption curve for each customer is monitored for 24 hours, and it is obtained from the load measurement every 15 minutes, resulting in a $\boldsymbol{x}_i$ vector of 96 samples. The challenge for the dimensionality reduction techniques begins when processing the data shown in Fig. 2, where the different load profiles impose clustering difficulties due to their nonlinear characteristics, similar dynamics, and distance correlations related at the profile voltages. One of the main causes for errors is the distances matrix $\boldsymbol{D}$ estimation.

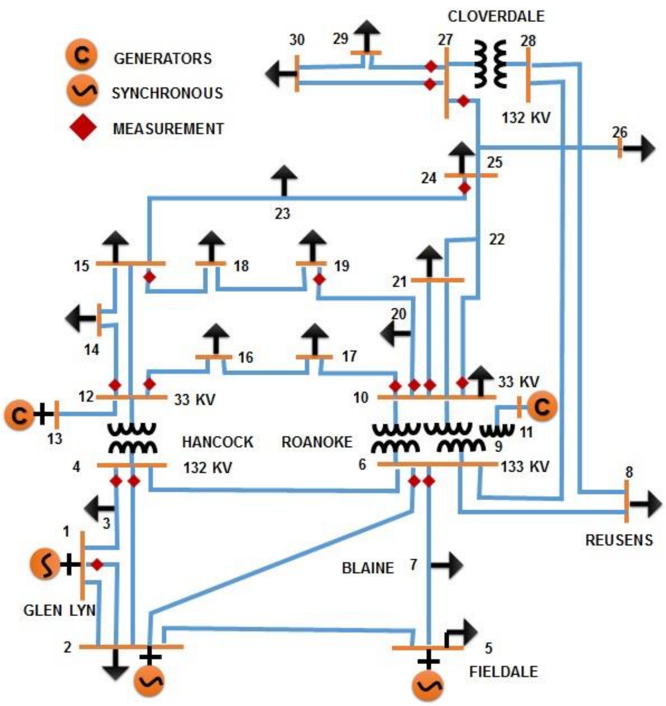

Fig. 1. IEEE de 30-bus system.

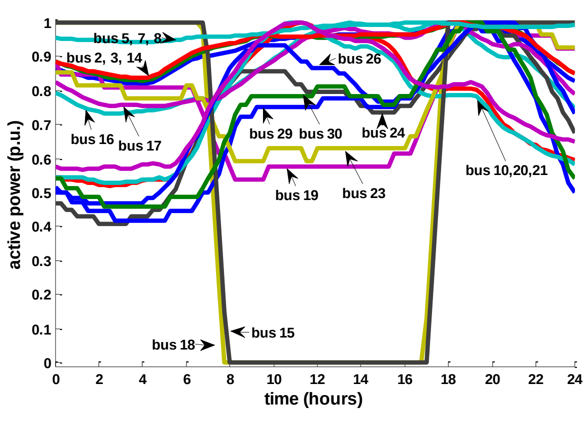

Fig. 2. Load profile extracted from IEEE 30-bus system.

B. Results analysis

Most of the clustering techniques used require to previously define the specific number of clusters. For this study, Isomap and LLE use $k = 8$, SNE requires an entropy measurement parameter named $Perplexity = 2.2$ and Sammon technique, considers that the Newton's method converged after 500 iterations. In order to compare the results obtained by the dimensionality reduction techniques, the techniques performance was evaluated in terms of the following points: The cost function error, computational time, visualization and clustering results.

In Fig. 3, it is shown the total variance of the data set $\boldsymbol{X}$, using the most significant principal components. The first two components are the ones that present the most quantity of variance in the data. The projection of the output space $\boldsymbol{Y}$ of the PCA technique is obtained according to this components. The possibility to explain a high percentage of information (variance), using a low number of components or variables that derive from the data transformation using the dimensionality reduction techniques, provides the justification of why to use dimensionality reduction techniques for clustering and visualization in the load profiles cluster problem.

In Table I, it is shown the comparative analysis of the techniques performance, every technique presents favorable results when it comes to speed computing time. The lower percentage of error in the cost function is reached by the SNE technique. On the other hand LLE technique was unable to find a solution.

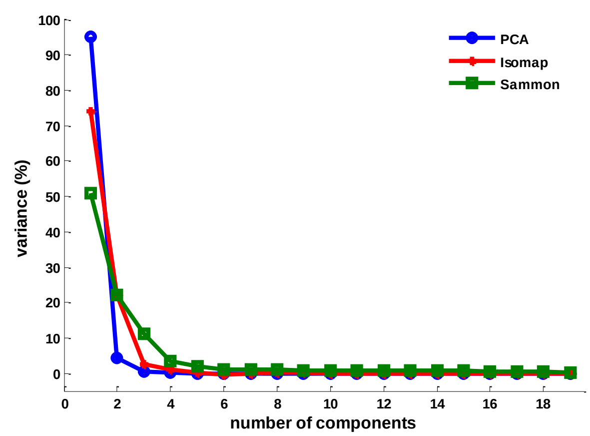

Fig. 3. Variance diagram for PCA, Isomap and Sammon.

TABLE I. Performance comparison techniques.

Technique	Performance Analysis		
	CPU-Time	*Error (%)*	*Complexity*
PCA	0.19 sec.	19.76	$O(D^3)$
ISOMAP	0.27 sec.	13.18	$O(N^3)$
SAMMON	10.16 sec.	6.664	$O(iN^2)$
LLE	1.80 sec.	-	$O(PN^2)$
SNE	3.48 sec.	0.15	$O(N^2)$

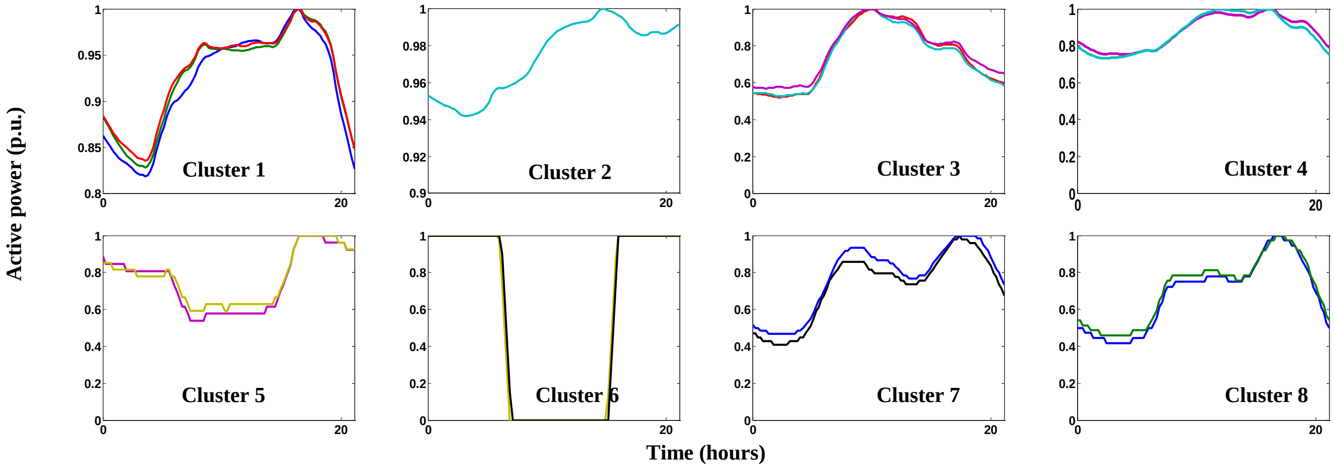

Fig. 9. Clustering results for the SNE technique using a perplexity $Perp = 2.2$.

Figs. 4-8, show the results for the clustering of load profiles measurements from the IEEE 30-bus system. The data reduction results are contained in the output matrix $\boldsymbol{Y} \in \mathbb{R}^{19x2}$. According to these results, the comparison and evaluation of visualization of each technique will be made.

In PCA and Isomap, it is observed that both techniques only find four clusters, because each load profile presents similar consumption behavior. Both techniques, despite of using a different distance metrics for data clustering, show difficulties when it comes to separating these four clusters, due to the close distances between the profiles, the distances between the clusters are constant, while the variance grows, making the similarities grow with them.

The results shown in Fig. 6 present a best separation in comparison with the results from PCA and Isomap. Because the normalized cost function preserves the distances between all the data, Sammon Mapping accomplishes to define every cluster, weighting short distances, and the similarities between these clusters will approach to zero, accomplishing to separate every cluster. Meanwhile in Fig. 7, it can be observed that LLE technique does not find a solution because of the nature of the data is too complex to LLE that include nonlinear characteristics and similar dynamics.

In Fig. 8 the results obtained with SNE technique are shown, the results reveal that the data structure is captured, that means that the load profiles orientation was preserved successfully and achieves the cluster's distribution according to the different customers. The SNE technique cost function, splits the data and assigns a larger cost for the $\boldsymbol{y}$ point's usage widely separated nearby $\boldsymbol{x}$ points, (that means, by using a small q_{ji} value, that represents a very large p_{ji} value), however, there will be a small cost to the close $\boldsymbol{y}$ points usage, that represent distant $\boldsymbol{x}$ data points.

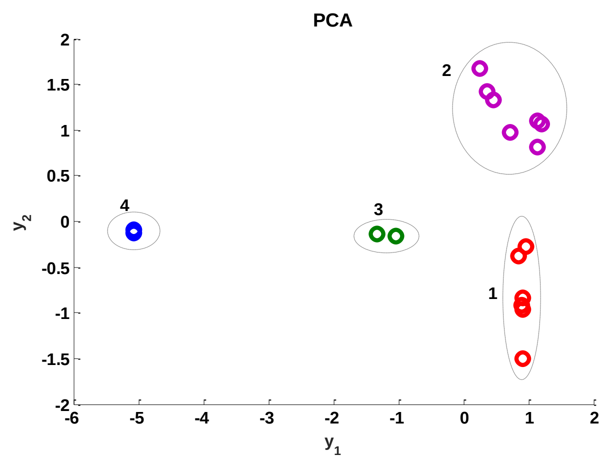

Fig. 4. PCA clustering results for load profiles.

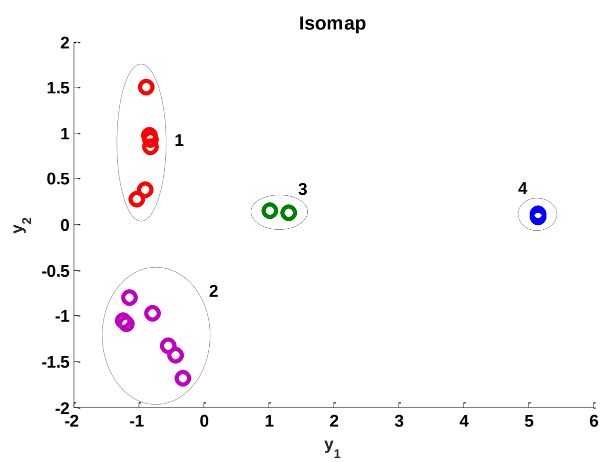

Fig. 5. Isomap clustering results for load profiles.

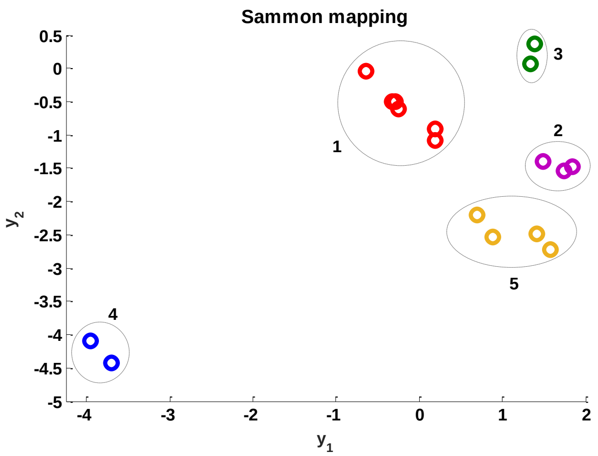

Fig. 6. Sammon mapping clustering results for load profiles.

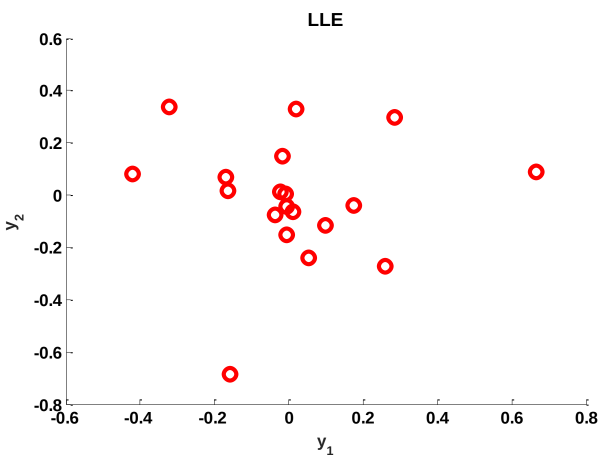

Fig. 7. LLE clustering results for load profiles.

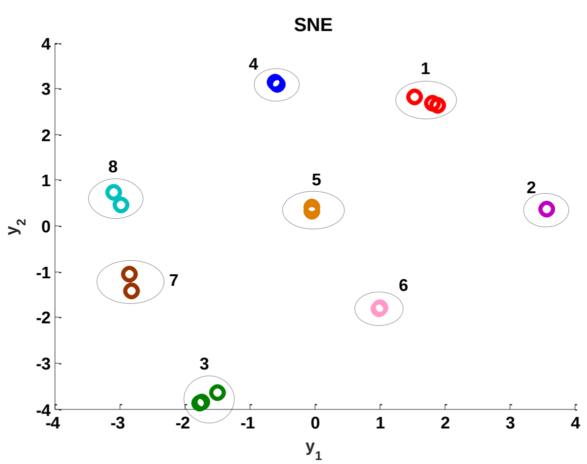

Fig. 8. SNE clustering results for load profiles.

Due to the fact that SNE showed superiority for clustering and visualization, Fig. 9 shows the clustering results obtained by using this technique. Each cluster was split according to the types of customers (commercial, industrial and residential) successfully. Lastly, detailed results for each cluster are shown in Table II.

TABLE II. Detailed clustering results obteined by SNE technique.

# Cluster	***# Bus***	***Type of Customer***
1	2, 3, 14	Commercial high consumption
2	5, 7, 8	Industrial high consumption
3	10, 20, 21	Commercial low consumption
4	16, 17	Commercial medium
5	19, 23	Industrial medium
6	15, 18	Industrial low consumption
7	24, 26	Residential high consumption
8	29, 30	Residential low consumption

V. Conclusions

Clustering techniques are extremely useful to the electricity suppliers to clustering different types of customers based in the consumption curve similarities. The study that was carried out in this paper shows the most relevant aspects of the dimension reduction techniques, from the clustering and visualization context. Results obtained by Sammon and SNE techniques shown an optimal performance for clustering load profile data with nonlinear, non-oscillatory and smooth dynamic characteristics. At the stage of this research, both techniques are suitable to process a small amount of data associated to load profiles. In a follow-up paper, two aspects will be evaluated: a larger volume and higher dimensional datasets will be used to test these techniques and the incorporation of clustering indexes may help to improve the clustering and visualization results.

VI. References

[1] Wu Junjie. "Advances in K-means Clustering", Springer theses. 2012. China.

[2] J. H. Ward. "Hierarchical grouping to optimize an objective function", J. Amer. Stat. Assoc., vol. 58, p.p. 236-244, 1963.

[3] J. C. Bezdek. "Pattern Recognition with Fuzzy Objective Function Algorithm", New York: Plenum, 1981.

[4] Y. H. Pao y D. J. Sobajic. "Combines use of unsupervised and supervised learning for dynamic security assessment", IEEE Transaction Power System., vol. 7, No. 2, p.p. 878-884, Mayo 1992.

[5] I. T. Jolliffe. "Principal Component Analysis", Springer series in statistic, 2da ed,. p.p. 32-500. 1986. New York.

[6] G. Chicco, R. Napoli y F. Piglione. "Comparisons Among Clustering Techniques for Electricity Customer Classification", IEEE Trannsactions on Power Systems, vol. 21, No. 2, Mayo 2006.

[7] L. van der Maaten, E. Postman ans J can den Herik "Dimensionality reduction: A comparative Review" Tilburg centre for Creative Computing, October 26, 2009.

[8] N.R Sakthivel, B. B. Nair, M. Elangovan, V. Sugumaran, y S. Saravanmurugan, "Comparison of Dimensionality reduction techniques for the fault diagnosis of mono block centrifugal pump using vibrations signals", El Sevier Engineering Science and Technology an International Journal, 2014.

[9] G. Hinton, y S. Roweis. "Stochastic Neighbor Embedding", Department of Computer Science, University of Toronto,. 8 p.p.

[10] J. Leeuw, y W. Heiser. "Theory of Multidimensional Scaling", Handbook of Statistics, vol. 2, p.p. 285-316, 1982.

[11] E. W. Dijkstra. "A note on two problems in conexion with graphs", Numerische Mathematik, vol. 1, p.p. 269-271, 1959.