DOI: 10.48011/asba.v2i1.1189

Characterization of load curves in a real distribution system based on K-MEANS algorithm with time-series data

Hernán R. Ullón * **Luis F. Ugarte** ** **Eduardo Lacusta Jr.** *** **Madson C. de Almeida** ****

* *School of Electrical and Computer Engineering, University of Campinas, SP, (e-mail: hrullon@dsee.fee.unicamp.br)*
** *School of Electrical and Computer Engineering, University of Campinas, SP, (e-mail: lfugarte@dsee.fee.unicamp.br)*
*** *RGE Sul Distribuidora de Energia S.A., SP (e-mail: lacusta@cpfl.com.br)*
**** *School of Electrical and Computer Engineering, University of Campinas, SP, (e-mail: madson@dsee.fee.unicamp.br)*

Abstract: The modernization of conventional distribution systems in smart grids leads us to face new challenges when dealing with extremely large databases, commonly called Big Data. The accuracy and volume of data have grown significantly with the introduction of Advanced Measurement Infrastructure (AMI). This generates a data tsunami used in different applications of power systems creating great computational efforts, as is the case with the use of a large database of load curves. Due to the patterns that are repeated annually in the demand for active and reactive power in distribution systems, it is necessary to use load clustering methodologies. Based on historical load data, this paper represents a comprehensive approach that uses data mining based on the K-Means clustering method in time-series data for the characterization of real load curves. Besides, a comparative analysis will be presented considering three different distance measurements. This data mining process is presented as a promising method for the recognition of patterns allowing to reduce large databases to some characteristic curves to reduce the computational burden in various applications of power systems. This clustering method is tested using a real database of distribution transformers at UNICAMP.

Keywords: Data mining, Distance measurements, Distribution system, K-Means algorithm, Time-series data

1. INTRODUCTION

The integration of new technologies and the modernization of the measurement and communication systems provide a data tsunami to utilities, allowing the state of the electrical system to be known at all levels and optimizing the operation of its services and products. The big data challenges in distribution systems are related to different sources, such as the information obtained from the different measurement equipment, the patterns of load curves, energy market prices, data management, among others (Ghorbanian et al., 2019).

However, to gain a deep understanding of user energy demand, it is necessary to identify the various patterns that characterize their load curves. In other words, the data must be organized and classified to convert it into easily interpretable information (Restrepo et al., 2018). Due to this, various clustering methods in time series are used to achieve these objectives, allowing better planning of public services and improving their policies.

⋆ The authors would like to thank CAPES, CNPq, FAPESP and CPFL, through ANEEL RD Program on the project PD-00063-3043/2018, for funding this research.

Clustering is a data mining technique in which data with similar features are grouped without advanced knowledge of the definitions of the clusters (Rai and Singh, 2010). Some conventional time-series clustering algorithms are discussed in Aghabozorgi et al. (2015) such as K-Means, K-medoids, or Hierarchical clustering, highlighting the K-Means algorithm as a promising technique for this type of analysis.

In recent years, there have been considerable research efforts to determine typical load curves using clustering algorithms. A clustering technique based on three stages is proposed in Panapakidis et al. (2013). This approach uses the hierarchical algorithm in order to cluster the daily load curves, highlighting that the clustering is done according to the similarity of their shapes, not the energy demand. A diffuse two-stage clustering is proposed in Zakaria and Lo (2009). This technique applied to load curves in different feeders previously used a Principal Component Analysis (PCA) to identify the most predominant features of load curves. The identification of the days for the load curves has been carried out in Benabbas et al. (2008). Visual identification was performed using Kohenen maps. K-Means was used as a complementary method of precision for the

class identification. In Pan et al. (2011), the identification of load curves based on Fuzzy clustering algorithm was used to predict the load of a particular day. As a complement, the Wavelet decomposition was used to decompose the similar-day load into high and low-frequency components to identify the feature of each component. Proclus, a clustering technique based on subspace projection was used for creating load curves (Piao et al., 2014). One of the strengths of this algorithm is that it reduces the influence of noise during the grouping process.

In this context, UNICAMP is currently developing two projects of great impact on the Brazilian community, namely, the Sustainable Campus and Electric Mobility. Both projects have monitoring and control systems through various meters and sensors installed throughout the campus and bus fleet creating great challenges for data analysis and knowledge discovery (Ugarte et al., 2019). This document presents a data analysis based on the examination, cleaning, transformation and modeling of data to draw conclusions about the information inherent in the data and, consequently, to make decisions based on the knowledge discovery.

This paper aims to derive the load profiles of a real distribution system located at UNICAMP by clustering characteristic load curves using the K-Means algorithm to minimize the burden and computational time of the various analyzes performed by the utilities. Besides, a comparative analysis will be carried out considering different types of metrics in order to better characterize electrical demand. Finally, these characteristic clusters will be used in studies of technical losses based on power flows carried out in Open Distribution System Simulator (OpenDSS).

The rest of the paper is organized as follows. Section 2 shows the implementation of the data mining methodology. Section 3 demonstrates the load curve clustering process using K-Means algorithm considering different metrics. Section 4 and 5 shows a result analysis and a application of the clusters to estimate the technical losses in distribution transformers. Finally, the conclusions are drawn in Section 6.

2. METHODOLOGY

The methodology used in this paper was the Knowledge Discovery in Databases (KDD), which is considered a non-trivial procedure to identify patterns that are useful, valid, novel, and understandable in the data (Fayyad et al., 1996). This process has several stages that will be described based on the study carried out.

2.1 Selection

Currently, there are around 330 smart meters deployed in distribution transformers, these perform measurements every 30 seconds, a previous classification by transformer was made to this large database making this study feasible.

The smart meters collected a total of 12 electrical features, however, most of these were classified by phase, which generated datasets of 27 features for each register. A summary is shown in Table 1.

Table 1. Electrical features.

activePowerA	currentA	reactivePowerA
activePowerB	currentB	reactivePowerB
activePowerC	currentC	reactivePowerC
angleA	frequencyA	threephaseActivePower
angleB	frequencyB	threephaseApparentPower
angleC	frequencyC	threephaseReactivePower
apparentPowerA	powerFactorA	voltageA
apparentPowerB	powerFactorB	voltageB
apparentPowerC	powerFactorC	voltageC

Based on the objective of the study of characterizing load curves, our focus variables were the active and reactive powers of each distribution transformer. Due to load distribution in the transformers and for the sake of simplification, this paper considered the following parameters: threephaseActivePower, threephaseReactivePower.

2.2 Preprocessing

Real databases generated by smart meters are normally subject to communication and electricity source problems, which leads to the presence of outliers. The detection of these values was performed using the statistical z-score technique, taking thresholds between 2.5 and 3, according to the analysis of the data nature and a load of each transformer, in addition to being recommended for extensive databases according to Shiffler (1988); Hoaglin (2013), retaining approximately 98% of the data. Although it is true that the visualization of the boxplot in Fig. 1 still shows the presence of outliers, this is because its mustache configuration uses a threshold of 1.5, but that value is not a good scenario when it comes to values obtained by real measurements.

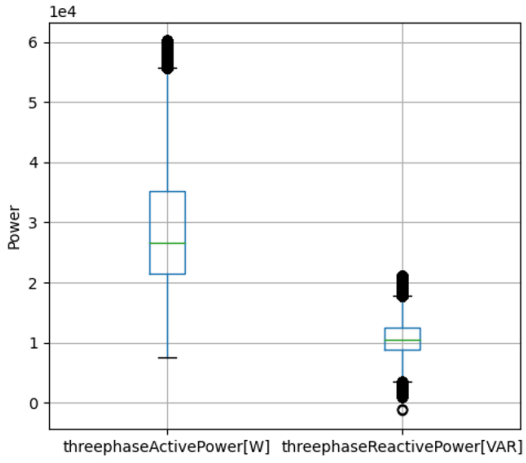


Figure 1. Boxplot for active and reactive power in the meter of the Food Engineering Institute.

Each record was collected as a function of time, this allowed an analysis of the number of records per month and eliminated from the analysis those months that did not meet a uniform number of records compared to the other months, then the scenario is shown in Fig. 2 for the transformer data with Id: 104, corresponding to the Institute of Food Engineering.

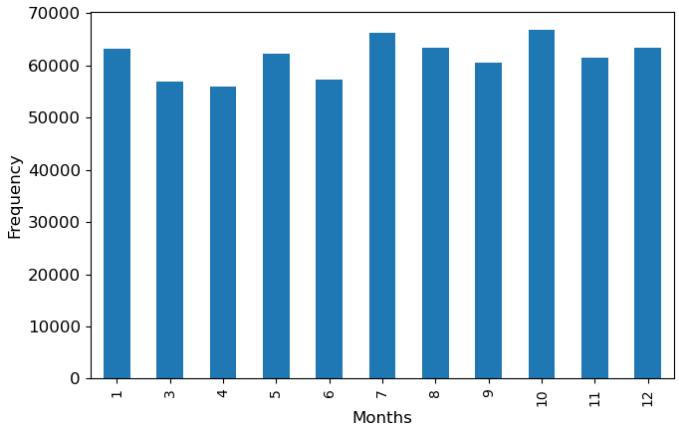


Figure 2. Bar chart of data rate per month for the Food Engineering Institute meter.

2.3 Transformation

So far, the database considered contains records with the values of active and reactive three-phase power, depending on the time in which said measurement was collected for a certain distribution transformer. This makes the data set have a great dimensionality in its number of rows and consequently, it requires a more expensive computational analysis (Aghabozorgi et al., 2015), therefore, a dimensionality reduction technique was applied, which will be described in the following steps:

(1) The raw time series were resampled at a time interval of 30 minutes per measurement, where their grouping was represented by the mean of their respective power values.
(2) When performing this resampling, the presence of missing data was inevitable, these were completed using the interpolation technique.
(3) Each one of the powers created time-based pivots, thus generating a data frame where each row represents a date and in each column, the time measurement was made, in Fig. 3 a graphic representation is presented.

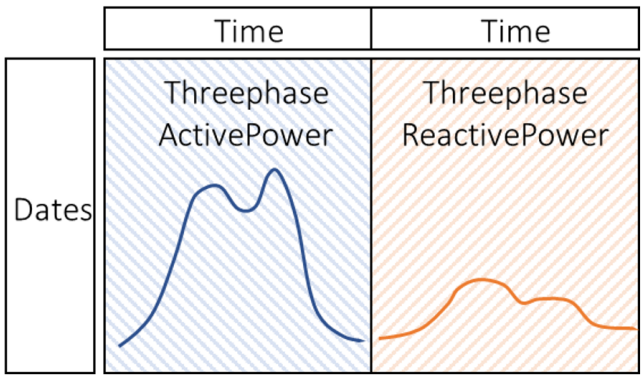


Figure 3. Final transformation of the dataframe

2.4 Data Mining

The chosen methodology for data mining was the Whole time-series clustering, with a features-based approach. According to Aghabozorgi et al. (2015) this methodology has four components:

- Dimensionality reduction
- Distance measurement
- Prototype
- Clustering algorithm
- Evaluation

Although the **dimensionality reduction** component was treated in the previous stage, it is highlighted by its level of impact on the performance of the algorithm, it should control that the reduction maintains the balance between quality and execution time. That is why several samples were tested before making a decision.

The **distance measurements** chosen in this study were: Euclidean, Dynamic Time Warping (DTW), and Soft-DTW. These are considered the most popular and common metrics to measure similarity in the time-series clustering (Aghabozorgi et al., 2015; Cuturi and Blondel, 2017; Salvador and Chan, 2007).

The Euclidean distance is considered the most basic, however, the literature shows that it can be surprisingly competitive (Aghabozorgi et al., 2015). Now we will detail the behavior of this in time-series, Let A and B be two time-series:
$A = [a_1 \ \cdots \ a_r \ \cdots \ a_R]$
$B = [b_1 \ \cdots \ b_r \ \cdots \ b_R]$
It is defined as a cost function between two points of the time series Squared of the Euclidean Distance δ:

$$\delta(a_r, b_c) = (a_r - b_c)^2 \tag{1}$$

From this cost function, a cost matrix is partially constructed between the points of the series, then the final trajectory is obtained from the cost calculated between each pair of points. The DTW distance is a metric that uses the same cost function defined above, however, its objective is to find the optimal alignment between two series that achieves a minimum global cost at the time and also guarantees the continuity of time (Zhang et al., 2017). For a better understanding of these metrics, the behavior is represented graphically in Fig. 4.

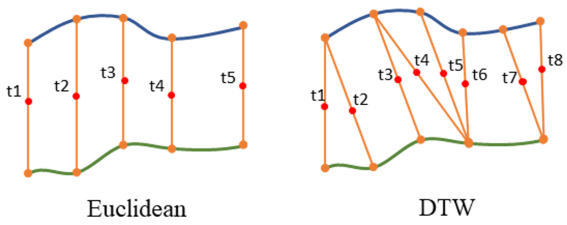


Figure 4. Central trajectory (T) between two series for Euclidean and DTW distance measurements, where $T = [t_1 \ \cdots \ t_r \ \cdots \ t_R]$.

Soft-DTW is a differentiable loss function that makes the method more precise due to its learning fit, and that both its value and its gradient can be calculated with a quadratic complexity of time/space.

According to the literature, this regularization is adequate for average time-series (Cuturi and Blondel, 2017). Before applying the algorithm, a random scheme was chosen as a

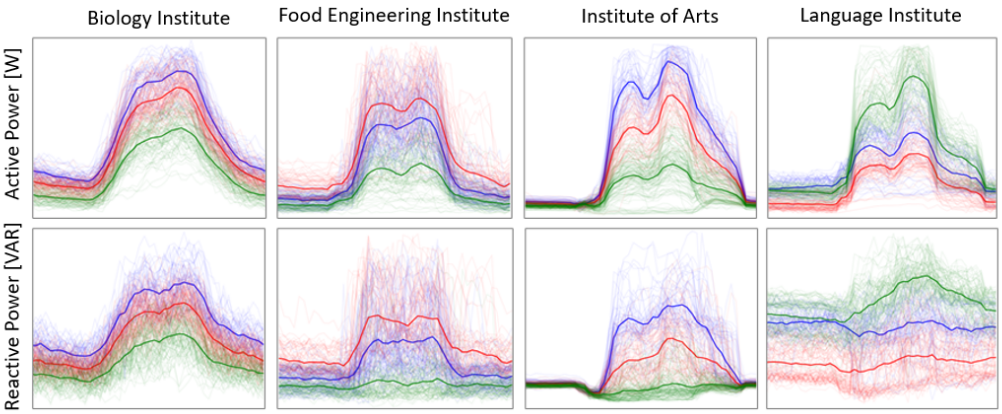


Figure 6. Time-series data for each meter, applying the K-Means algorithm for active and reactive power.

prototype, however, this consideration may vary in future works depending on the volume of data.

The **clustering algorithm** chosen was K-Means. This algorithm is capable of grouping objects into a certain number of groups (K) based on their attributes. Grouping is performed by minimizing the sum of squares of the distances between the group's centroid and its corresponding data (Grigoras et al., 2010).

$$min\,(E) = min\left(\sum_{i=1}^{K}\sum_{x\in C_i} d(x, z_i)\right) \qquad (2)$$

K-Means has demonstrated its solidity with the large number of works that support its performance, and although in the literature it shows to be sensitive to outliers, this problem was taken into consideration by improving the preprocessing and transformation phases(Azad et al., 2014).

The K-Means algorithm is characterized by knowing the number of clusters a priori, for this reason, some techniques allow an optimal search for the K value, among the most commonly used are the silhouette coefficient, the elbow technique. For this particular study, a hybrid technique was used between these two indicators, since the elbow was not as pronounced in the different scenarios, the silhouette coefficient was in charge of evaluating the best K value. An application scenario of the elbow technique for the four analysis meters using the Euclidean distance as a metric is presented in Fig. 5.

Once the aforementioned techniques were applied, the optimal value of K selected for this study was 3, for the analysis transformers since the grouping behavior was uniform. Fig. 6 shows the performance of the K-Means algorithm for a (K = 3) from a graphical point of view of the load curves of some of the transformers with which this methodology was validated.

The KDD process has as its final stage the **Interpretation and Evaluation of Knowledge**, this will be shown in the following chapters.

3. RESULT ANALYSIS

For the sake of simplification, the results will be presented based on the meter with Id: 157, which represents the demand consumed by the Institute of Biology.

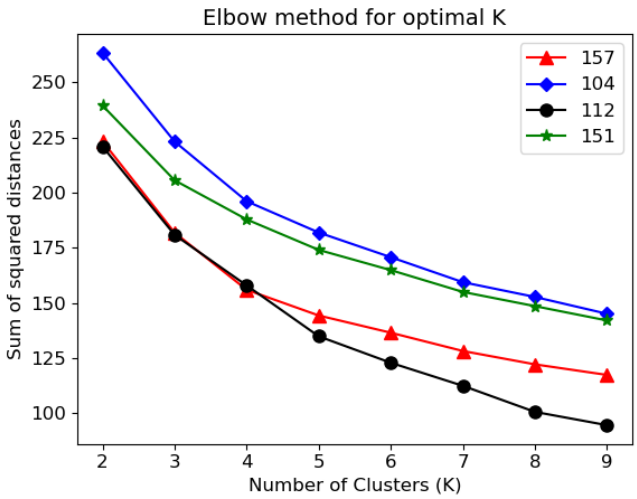


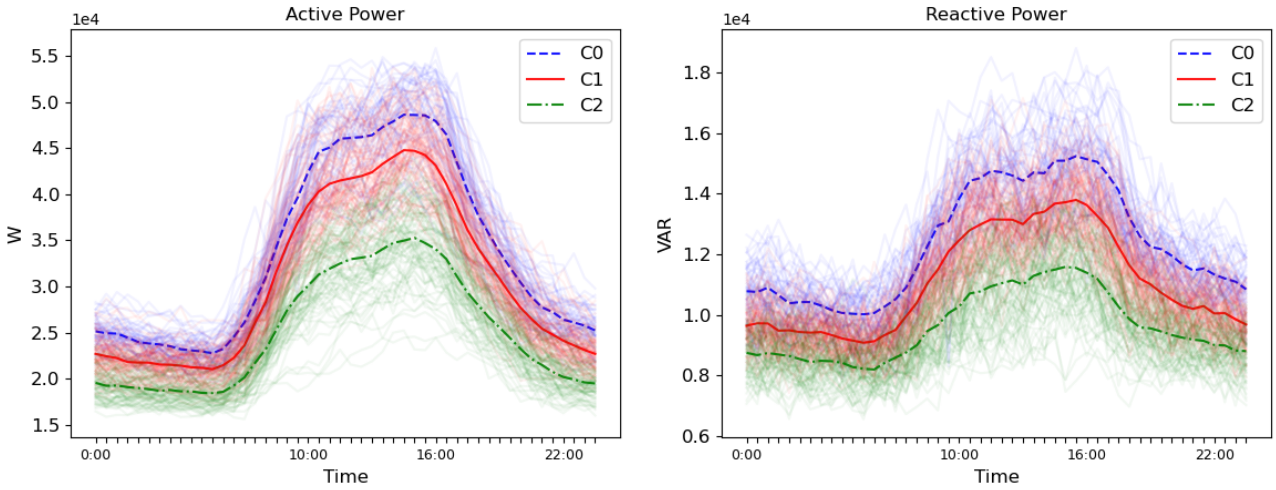
Figure 5. Elbow technique for optimal K search for each meter

The load curve patterns discovered by the K-Means algorithm using the **Euclidean** distance showed an average and smooth behavior in its centroids. The result is shown in Fig. 7, where the centroids C0, C1, and C2 represented a quantity of 53, 95, and 80 days respectively. It should be noted that the greatest number of days were represented by C1, which fits into the average power consumption throughout the day.

The centroids identified using **Soft-DTW** had slight changes concerning those previously obtained, adding small variabilities in the curves of the centroids ending with the smoothness of the curves, as shown in Fig. 8. However, the trend in the number of days per group remained at 55, 94, and 79 for C0, C1, and C2, respectively.

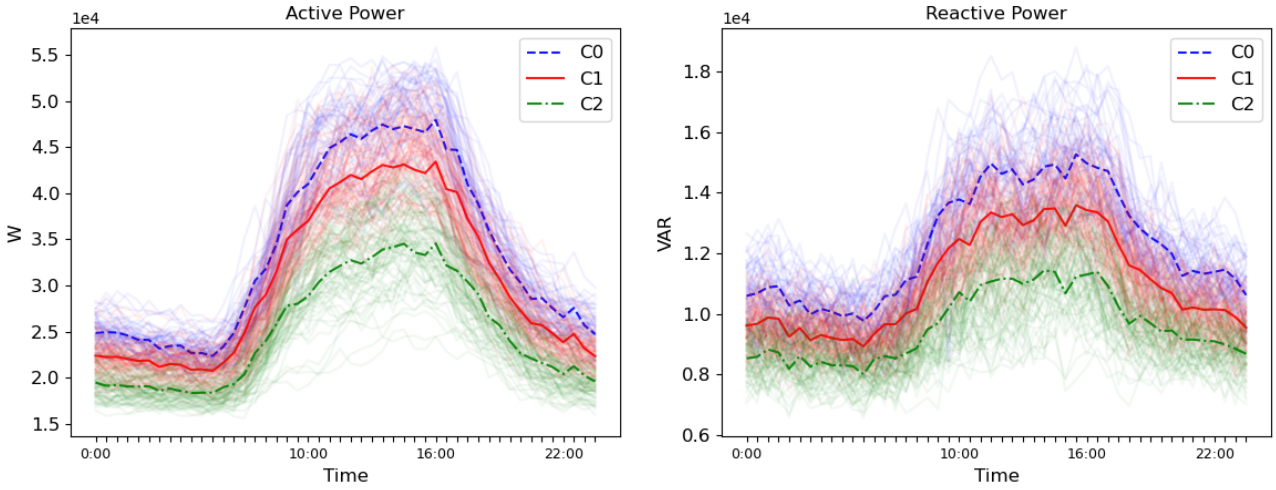
The characterization of the curves obtained using **DTW** generated a very different knowledge from those shown above, large variations were generated between the time intervals in both active and reactive power. This resulted in the superposition of the centroids at the certain timestamp, as shown in Fig. 9.

For the evaluation of the scenarios shown, a highly flexible visualization algorithm called t-SNE was applied. Based



(a) (b)

Figure 7. Clustering of time-series data for Euclidean distance measurement.



(a) (b)

Figure 8. Clustering of time-series data for Soft-DTW distance measurement.

on the dimensionality reduction technique, it manages to represent large datasets most optimally (Wattenberg et al., 2016). In Fig. 10 we can visualize the application of the algorithm that managed to reduce from 98 columns to just 2, each point represents a time series of our dataset.

Finally, the silhouette coefficient was calculated as a validation metric for each of the implemented models. Table 2 shows these values, where it is worth mentioning that the best-performing model was K-Means using Soft-DTW

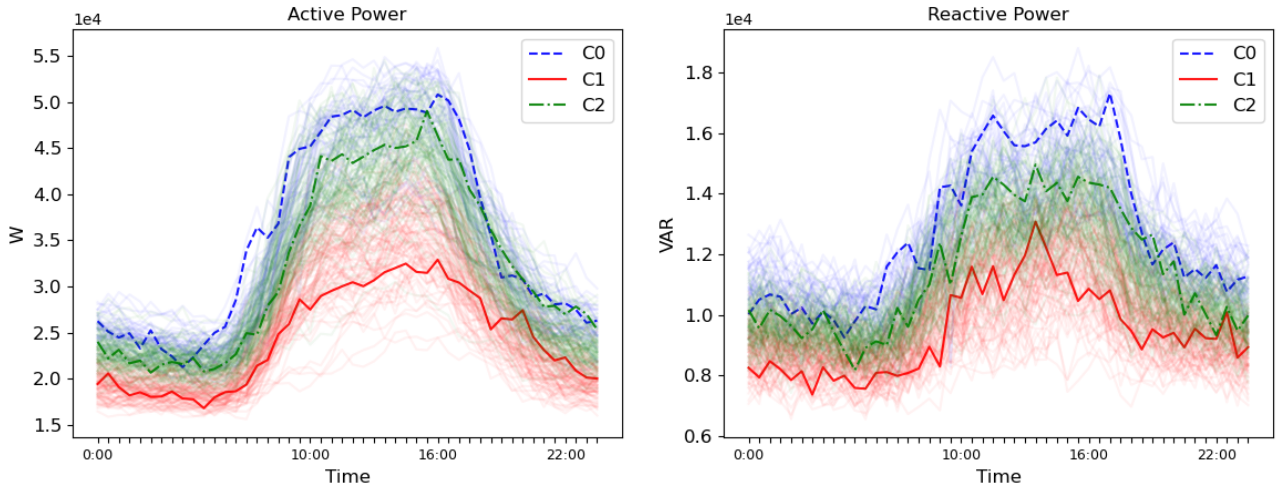
Table 2. Evaluation of Silhouette Coefficient.

Measurements	Silhouette
Euclidean	0.421056
Soft-DTW	0.611914
DTW	0.436307

4. APPLICATIONS IN TECHNICAL LOSS ESTIMATION

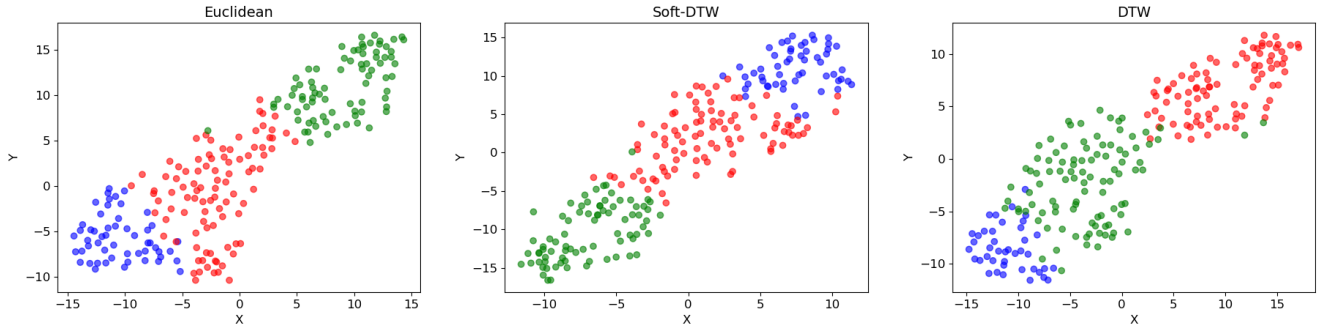
For the utilities, a deep knowledge of the behavior of electricity demand can provide an environment suitable to the proper implementation of the Distribution Management System (DMS). Among the various functions that DMS performs such as contingency analysis, load prediction, power flow, reactive power and voltage control, state estimation, etc., power flow stands out as the principal tool for many applications that the utilities carry out as the loss estimation in the electrical system.

In the context of smart grids, carrying out this type of analysis based on the load flow demands a high burden and computational time. To outline this problem and, in turn, make accurate decisions, this paper presented in the previous sections the obtaining of characteristic load curves that adequately represent the large database obtained from smart meters.



(a) (b)

Figure 9. Clustering of time-series data for DTW distance measurement.



(a) (b) (c)

Figure 10. Visualization of time-series data based on the dimensionality reduction applied by the t-SNE algorithm.

To exemplify the aforementioned in the application of the estimation of technical losses, a distribution transformer of the feeder BGE-06 is connected to a conventional load of the Biology Institute at UNICAMP as shown in Figure 11. The load curves were obtained by using a smart meter connected on the low voltage side considering 228 weekdays. On the other hand, the same process is carried out considering only the characteristic load curves obtained from the clustering process considering the three distances presented in the previous chapters. Considering the 228 real load curves, the energy losses in the distribution transformer of 500 kVA are 7031.70 kWh.

Table 3 shows the estimated technical losses using the three clusters obtained according to the Euclidean distance. When estimating losses considering the real load curves, it was necessary to simulate the 228 days. However, Table 2 shows that it was only necessary to simulate a power flow for each cluster and consequently to estimate the losses. That estimated value is multiplied by the day number that each cluster represents. It can be seen that the estimated losses considering the Euclidean distance are 7380.20 kWh. The percentage difference between the estimated losses through clusters and the real curves is 4.95%.

The same procedure shown in Table 3 considering the Euclidean distance was performed in Tables 4 and 5 for the

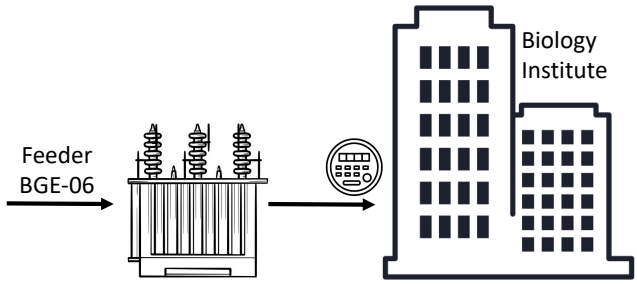


Figure 11. Distribution transformer connected to a conventional load.

Table 3. Technical loss estimation using Euclidean distance

Cluster	Days	Energy Loss per Cluster [kWh]	Total Energy Loss [kWh]
1	53	33.75	1788.75
2	95	32.71	3107.45
3	80	31.05	2484.00
Total	**228**	**97.51**	**7380.20**

DTW and Soft-DTW distances respectively. It can be seen that the estimated losses considering the DTW distance are 7429.62 kWh with a percentage difference between the

losses estimated by means of clusters and the real load curves is 5.65%. On the other hand, considering the Soft-DTW distance, the estimated losses are 7378.91kWh with a percentage difference between the losses estimated by means of clusters and the real load curves is 4.93%.

Table 4. Technical loss estimation using DTW distance

Cluster	Days	Energy Loss per Cluster [kWh]	Total Energy Loss [kWh]
1	41	34.54	1416.14
2	88	30.85	2714.80
3	99	33.32	3298.68
Total	**228**	**98.71**	**7429.62**

Table 5. Technical loss estimation using Soft-DTW distance

Cluster	Days	Energy Loss per Cluster [kWh]	Total Energy Loss [kWh]
1	55	33.77	1857.35
2	94	32.67	3070.98
3	79	31.02	2450.58
Total	**228**	**97.46**	**7378.91**

In comparative terms, the K-Means algorithm using the Soft-DTW distance presented a smaller percentage difference compared to the other two distances. On the other hand, it should be noted that the burden and computational time decreased considerably since it was only necessary to carry out three power flow simulations, one for each characteristic cluster that represents the real load curves.

5. CONCLUSIONS

Due to the high impact projects developed at UNICAMP, Sustainable Campus and Electric Mobility, it is possible to have large amounts of data obtained from various meters and sensors installed throughout the campus. This paper focused mainly on the load curves obtained by the smart meters installed on the secondary side of the distribution transformers at UNICAMP.

Among the various activities that the utilities carry out, the estimation of technical losses in the electrical system stands out. In the context of smart grids, the said process requires great computational efforts due to a large amount of data. To outline this problem, this paper presented a comparative analysis of three metrics, the Euclidean, DTW and Soft-DTW distance, applied to the K-Means algorithm to characterize the load curves that best fit to the real load curves.

The case study was carried out with data obtained from the distribution transformer of the Biology Institute, where the technical losses were compared considering the real load curves and the clustered load curves. Although the results of the metrics have found different patterns from a geometric viewpoint, the technical losses presented a similar behavior.

In addition to reducing the computational burden of various utility applications, the deep knowledge of load curve patterns allows the design of DSM strategies integrating all possible entities in the energy sector.

REFERENCES

Aghabozorgi, S., Shirkhorshidi, A.S., and Wah, T.Y. (2015). Time-series clustering–a decade review. *Information Systems*, 53, 16–38.

Azad, S.A., Ali, A.S., and Wolfs, P. (2014). Identification of typical load profiles using k-means clustering algorithm. In *Asia-Pacific World Congress on Computer Science and Engineering*, 1–6. IEEE.

Benabbas, F., Khadir, M.T., Fay, D., and Boughrira, A. (2008). Kohonen map combined to the k-means algorithm for the identification of day types of algerian electricity load. In *2008 7th Computer Information Systems and Industrial Management Applications*, 78–83. IEEE.

Cuturi, M. and Blondel, M. (2017). Soft-dtw: a differentiable loss function for time-series. *arXiv preprint arXiv:1703.01541*.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., et al. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, 82–88.

Ghorbanian, M., Dolatabadi, S.H., and Siano, P. (2019). Big data issues in smart grids: A survey. *IEEE Systems Journal*, 13(4), 4158–4168.

Grigoras, G., Cartina, G., and Rotaru, F. (2010). Using k-means clustering method in determination of the energy losses levels from electric distribution systems. *World Scientific and Engineering Academy and Society (WSEAS)*, 52–56.

Hoaglin, D.C. (2013). Volume 16: How to detect and handle outliers.

Pan, X., Zhang, P., and Xue, W. (2011). Short-term load forecasting based on fuzzy clustering wavelet decomposition and bp neural network. In *2011 Asia-Pacific Power and Energy Engineering Conference*, 1–4. IEEE.

Panapakidis, I.P., Alexiadis, M.C., and Papagiannis, G.K. (2013). Three-stage clustering procedure for deriving the typical load curves of the electricity consumers. In *2013 IEEE Grenoble Conference*, 1–6. IEEE.

Piao, M., Shon, H.S., Lee, J.Y., and Ryu, K.H. (2014). Subspace projection method based clustering analysis in load profiling. *IEEE Transactions on Power Systems*, 29(6), 2628–2635.

Rai, P. and Singh, S. (2010). A survey of clustering techniques. *International Journal of Computer Applications*, 7(12), 1–5.

Restrepo, J.A., Sierra, S.E., and Rosero, J.A. (2018). Load curve characterization based on real time measurements: Case of study in colombia. In *2018 IEEE PES Transmission & Distribution Conference and Exhibition-Latin America (T&D-LA)*, 1–5. IEEE.

Salvador, S. and Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5), 561–580.

Shiffler, R.E. (1988). Maximum z scores and outliers. *The American Statistician*, 42(1), 79–80.

Ugarte, L.F., Sarmiento, D.N., Mariotto, F.T., Lacusta, E., and de Almeida, M.C. (2019). Living lab for electric mobility in the public transportation system of the university of campinas. In *2019 IEEE PES Innovative Smart Grid Technologies Conference-Latin America (ISGT Latin America)*, 1–6. IEEE.

Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. *Distill*. doi:10.

23915/distill.00002. URL http://distill.pub/2016/misread-tsne.

Zakaria, Z. and Lo, K. (2009). Two-stage fuzzy clustering approach for load profiling. In *2009 44th international universities power engineering conference (UPEC)*, 1–5. IEEE.

Zhang, Z., Tavenard, R., Bailly, A., Tang, X., Tang, P., and Corpetti, T. (2017). Dynamic time warping under limited warping path length. *Information Sciences*, 393, 91–107.