Characterization of electricity demand based on energy consumption data from Colombia

Santiago Toledo-Cortés[1,2], Juan Sebastián Lara[2], Álvaro Zambrano[3], Fabio Augusto González Osorio[2], Javier Rosero García[3]

[1]Department of IT and Process Optimization, Faculty of Engineering, Universidad de La Sabana, Campus Puente del Común, Chía, Colombia
[2]MindLab Research Group, Department of Systems and Industrial Engineering, Faculty of Engineering, Universidad Nacional de Colombia, Bogotá, Colombia
[3]EM&D Research Group, Department of Electrical and Electronic Engineering, Faculty of Engineering, Universidad Nacional de Colombia, Bogotá, Colombia

Article Info

Article history:

Received Oct 29, 2022
Revised Jan 7, 2023
Accepted Feb 3, 2023

Keywords:

Clustering
Customer segmentation
Electric consumption profile
Electric demand
K-means
Load classification
Smart meter

ABSTRACT

The development of dynamic energy distribution grids to optimize energy resources has become very important at the international level in recent years. A very important step in this development is to be able to characterize the population based on their consumption behaviour. However, traditional consumption meters that report information at a monthly rate provide little information for in-depth analysis. In Colombia, this has changed in recent years due to the implementation and integration of advanced metering infrastructure (AMI). This infrastructure allows to record consumption values in small time intervals, and the available data then allows for the execution of many analysis mechanisms. In this paper we present an analysis of the electricity demand profile from a new dataset of energy consumption in Colombia. A characterization of the users demand profiles is presented using a k-means clustering procedure. Whit this customer segmentation technique we show that is possible identify customer consumption patterns and to identify anomalies in the system. In addition, this type of analysis also allows to assess changes in the consumption pattern of users due to social measures such as those resulting from the coronavirus disease (COVID-19) pandemic.

Corresponding Author:

Javier Rosero García
EM&D Research Group, Department of Electrical and Electronic Engineering, Faculty of Engineering, National University of Colombia
Carrera 45 # 26-85, Bogotá, 111321, Colombia
Email: jaroserog@unal.edu.co

1. INTRODUCTION

The increasingly massive integration of advanced metering infrastructure (AMI) has led to energy consumption data collection in a way we have never seen before. One possible application consists of being able to characterize users based on the way they demand electrical energy, analysing how the demand curve varies from hour to hour. This would allow us to know the different types of users that make up a distribution system.

In Colombia, the implementation of AMI is at an early stage. Although some progress has been made from a regulatory perspective, the deployment is still based on small pilots that have been implemented in different areas of the country. However, these pilots continuously generate data that can be an input to identify the types of electricity consumers that exist in the country. Additionally, by identifying additional variables such as those related to the geographical location of the metering points such as average temperature or height above sea level, the impact of these types of variables on consumption can easily be evidenced.

In this paper we present a new database processed from the collection of power consumption measurements in Colombia between 2017 and 2021. The data were provided at the discretion of five regional grid operators, and therefore we had no influence or decision on the dates of coverage or the number of samples. Each sample in the dataset contains information related to energy consumption levels, and in addition, for each user there is information on the type of user and the geographic location. Each user is registered among three possible classes: residential, commercial, or industrial. The geographic location allows to relate climate and altitude information. Although Colombia is in the tropics and has no climatic seasons (the average temperature varies little throughout the year), there are marked differences in different parts of the country. The climate depends directly on the altitude above the sea level, which in inhabited areas ranges from 0 m.a.s.l. (as in the north of the country) to above 2,600 m.a.s.l. (as in the capital Bogotá). Average annual temperatures can then range from 12 °C to beyond 30 °C, and this is expected to have a direct impact on electricity consumption behaviour. For these reasons we will take this information into account in the segmentation analysis.

We perform data processing and subsequent user segmentation using the k-means clustering method. K-means is an unsupervised machine learning method that automatically clusters data depending on the similarity between their features and has been previously used successfully in similar tasks [1]. With our method we show that: i) the average consumption in each hour of the day for a user is a good basis for representing the type of consumption of that user; ii) K-means successfully captures the characteristics of the data in the feature space created for them, obtaining easily interpretable clusters that can be related to the climatic variables of the country; and iii) it is possible to track the effects on the consumption behaviour produced by major social restrictions, which may be helpful in future planification processes.

The rest of the paper is organized as follows: in section 2 we will give a brief overview of the related work about user characterization in power grids. In section 3 we will look in detail at the characteristics of the dataset and the processing of the data. In section 4 we present the results of the clustering process and in section 5 we present the conclusions of our work.

2. RELATED WORK

The current need to build reliable grid services has led to many research efforts in energy consumption profile analysis. Much of the previous work has focused separately on analysing data coming exclusively from industrial consumption, or exclusively from buildings or residential areas. It is notable however that, regardless of the source, data collection plays a major role in these processes. This is due to the variability of the data structure, and the fact that this information can usually be easily complemented by geographical, meteorological and socio-economic information.

After collection, the final clean-up and adequacy of the data depends very much on the objective of the analysis. These can range from demand response analysis, through energy system management, to control analysis or energy flexibility performance indicators [2]. In any case, for the customer segmentation task, a customer feature space needs to be constructed. For this, average daily load profiles have been used [3], [4], where the available daily load profiles of each user are averaged, in order to identify large-scale patterns. Analyses can also be made on the basis of peak demand periods and the percentage overlap of these periods in the day. This is in order to study how likely it is that a user can respond to consumption stimuli in different time periods [5], [6]. Or it can be analysed directly using daily load profiles, without further processing, so that day-to-day consumption variations of users can be studied [7], [8].

Regarding the analysis methods, again, the use of various models depends on the final objective. However, clustering methods are practically mandatory for the task of user characterization and segmentation [2]. Classified within the family of unsupervised learning methods, these methods allow finding clustering patterns according to the similarity of the data representation and therefore allow finding average trends of energy use. This provides a basis for finding latent relationships between descriptors and external factors in the data and is a good first input for the task of characterizing the population and general demand behaviour.

There are several options for clustering. K-means, one of the most famous, works by optimizing the average distance between each of the data to an assigned centroid. The number of centroids is chosen in advance and is a parameter to be explored from the direct interpretation of the results. Derived from this method is the fuzzy c-means, which makes cluster boundaries more flexible and allows overlapping of clusters, or algorithms for hierarchical clustering, which allows grouping by levels. Although there is no consensus on the preference of a particular method to perform clustering tasks [2], and while often the results reached with each algorithm may be the same, for the demand characterization task, k-means was found to be more consistent [9].

Clustering methods have been implemented for a variety of purposes related to building energy consumption, such as pattern recognition [10], identification of abnormal energy behaviour [11], general characterization of building energy demand [12], demand management in the industrial [13] and residential sectors [14], forecasting of building energy consumption [15], and peak demand [16]. These techniques are also used for a variety of applications, such as identifying priority targets for energy efficiency programs [17], optimizing equipment sizing, energy storage, grid operation, renewable energy integration [18] and commercial offers [19]. Studies have mainly focused on households and later mixed industrial and commercial buildings, as reported in [16].

Finally, the comparison of clustering methods has been extensively studied [20]. For example, the performances of k-means and hierarchical clustering algorithms have been investigated by Quintana *et al.* [11], Satre-Meloy *et al.* [16], Chicco *et al.* [19], who also compared them with fuzzy k-means and follow- the-leader algorithms, and Xu and Massachuse [5], who also tested adaptive k-means and symbolic aggregate approximation methods. Beyond all this exploration, k-means is still the most used algorithm for performance analysis of non-residential buildings [20], and has proven to be the best clustering method for the analysis of residential buildings [21].

3. METHOD

3.1. Dataset

The collection of data of each grid operator begins with a formal request from them. Once the communication channel has been initiated, the signing of a confidentiality agreement is proposed, which will guarantee that the information will only be used for academic purposes. Once the agreements have been signed, the information is delivered by the grid operators. In this sense, different means have been used for this purpose depending on the amount of data to be delivered. A first way is through access to cloud repositories where access to information is allowed. Another way to deliver this information is through the delivery of physical hard drives which contain the different variables that are collected from the AMI system. It is important to mention that AMI measurements from 6 grid operators were integrated. However, it was evident that each operator handled its own form of data storage, making its integration difficult.

The data was collected from 166,630 smart meters belonging to five regional grid operators. This generated 3,104,870,233 records with consumption information. This database is the first to integrate metering information from different network operators in Colombia. The specific details on the number of meters and records by grid operators are given in Table 1. In addition, the type of user and its geographical information is also available which is further is used in the characterization analysis.

Table 1. Detail of the number of meters and the number of records by grid operator in the database

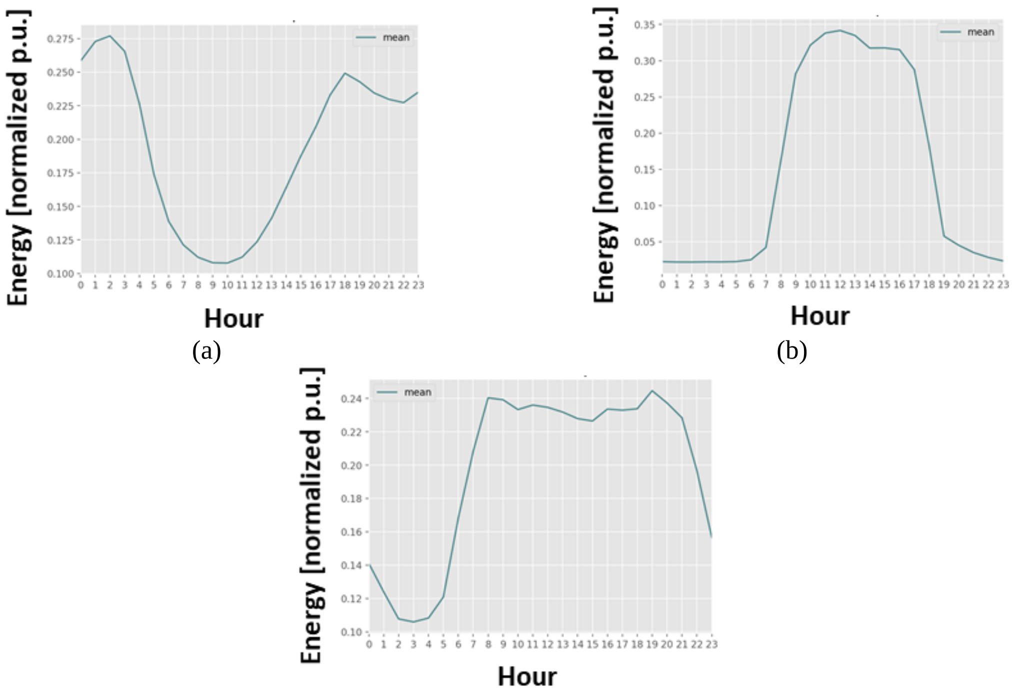
Grid operator	N° of meters	N° of records
Operator A	34,495	21,088,752
Operator B	50,723	598,883,091
Operator C	7,004	7,423,829
Operator D	73,793	2,470,092,041
Operator E	615	7,382,520

3.1.1. Pre-processing

For optimised storage purposes, a first filter of the raw data is made leaving only measurement values, date and time, and sensor identification. This is stored in Parquet files. This allows the subsequent manipulation and cleaning to be done in Python, using Dask [22] and Pandas [23], [24]. Further cleaning of the data consists of filtering out only the active energy data, standardising the date and time information. This allows to select those samples where measurements are available for the 24 hours of the day.

3.1.2. Feature selection

Once the pre-processing is made, a database is created where each sample consists of the measurement value for each of the 24 hours of the day. Then a grouping by sensor is made, and in each group the values corresponding to each hour of the day are averaged. If there are, for instance, five records for the same sensor, each one of them with measurements in the 24 hours of the day, from five different days, at the end of the process we may obtain a single record in which for each hour of the day we will have an average value. Then, for that sensor we will have the mean consumption for each one of the 24 hours of the day. In this way, we have a set of 24 unique features for each sensor. Figure 1 shows the demand curves generated in this way for different users: Figure 1(a) corresponds to a residential profile, Figure 1(b) depicts a commercial profile, and Figure 1(c) an industrial profile. The quantitative details of the final dataset thus created are described in Table 2.



(a) (b)

(c)

Figure 1. Load profiles samples after data prepossessing and normalization, (a) a load profile of a residential user, (b) a load profile of a commercial user, and (c) a load profile of an industrial user

Table 2. Detail of the final number of mean records by grid operator in the database. Each mean record consists of 24 values corresponding to the mean of the measured consumption for each hour of the day

Grid Operator	N° of Meters
Operator A	8400
Operator B	47798
Operator C	6291
Operator D	62595
Operator E	77

3.2. Clustering

Clustering is a processing approach that seeks to find groupings in the data based on its possible intrinsic patterns. It does not use any kind of labelling to learn how to segment the data. Therefore, it is widely used precisely in the tasks of customer segmentation, exploratory data analysis, pattern recognition and information compression. Among the many clustering algorithms, k-means appears within the family of unsupervised machine learning methods, posing an optimization problem based on the initial choice of centroids that will later define the clusters. K-means has been tested repeatedly and has been shown to be the best for energy user segmentation purposes, both at industrial and residential level [20], [21], so we used it for our data segmentation procedure.

3.2.1. K-means

K-means is an unsupervised machine learning method that works by finding a set of optimal centroids by minimizing the inertia, given by (1):

$$\min_{S} \sum_{i=0}^{n} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2 \tag{1}$$

where C is the set of clusters, $\{\mu_i\}_i$ is the set of centroids and $\{x_i\}_i$ is the training data. Inertia is a measure of coherence or cohesion of the resulting clusters. The number of clusters is a parameter that must be chosen before starting the optimisation process, and therefore must be explored and chosen according to the same measure of inertia, or to other performance metrics such as the silhouette score [25], as well as a final inspection and interpretation of the results. The initialization of the centroids can be made randomly, although this may affect the final convergence of the algorithm. While convergence is guaranteed, k-means can reach a local minimum. For this reason, the procedure is initialized several times, using for this different initialization techniques that seek to find well-spaced centroids to begin the optimization process [25].

4. CUSTERING RESULTS AND DISCUSSION

The experimental implementation was made in Python using Scikit Learn [26] and Dask [22] with a Scikit Learn backend. Initialization of the centroids is made using the kmeans++ algorithm, exploring 10 different initializations, and a maximum of 300 iterations. Experiments were performed by testing a number of clusters in a range from 2 to 60, and the inertia graph is inspected. From this, a fixed number of clusters is chosen for inspection of the centroids. These centroids represent the average demand curve of all users in the same cluster. The inspection of these centroids and their interpretation will verify whether the chosen number of clusters is adequate or not.

4.1. Cluster analysis

40 final clusters were chosen. The centroid of each cluster accounts for the average consumption profile. 23 clusters have a residential consumption profile and 17 have a non-residential consumption profile. The decision as to whether a cluster corresponded to residential or non-residential profile was made on the basis its shape of consuming energy. In addition, based on the geographical information of the sensors, and using information from the Colombia Institute of Hydrology, Meteorology and Environmental Studies (IDEAM) [27], we could categorize each cluster according to the predominant altitude and climate associated to the samples of the cluster. The complete description of altitude and climate categorization can be found in Tables 3 and 4 respectively. It is worth noting that during this process, measurements for some users evidenced that although they are classified as residential users by the grid operator, they present a behaviour of a commercial user.

Table 3. Description of altitude categorization criteria

Altitude category	Description
High	Higher than 2000 meters above sea level
Medium	Between 1000 and 2000 meters above sea level
Low	Less than 1000 meters above sea level

Table 4. Description of the climate categorization criteria

Climate category	Description
Cold	Average temperature below 18 °C (~64 °F)
Temperate	Average temperature between 18 °C (~64 °F) and 24 °C (~75 °F)
Warm dry	Average temperature above 24 °C (~75 °F)
Warm Humid	Average temperature above 24 °C (~75 °F)

4.1.1. High altitude/cold climate

23 clusters correspond to users located in a high-altitude region. In Colombia, this automatically implies a cold climate. 13 out of those 23 clusters have a residential consumption profile and can be classified in four types of users:

a) Customers with progressive increase in consumption Figure 2: this type of customer is characterized by low energy consumption in the early hours of the morning, gradually increasing consumption throughout the day until peak consumption is reached at around 20:00 hours.
b) Customers residential Figure 3: this group of residential users has two consumption peaks throughout the day. The largest peak is in the morning hours and the second peak is in the evening hours.
c) Customers with double peak consumption Figure 4: this type of customer is characterized by two consumption peaks throughout the day. The main peak occurs in the evening at around 20:00 hours. However, in the morning there is an increase in demand that may be associated with the use of electrical appliances before going out to do their daily chores.
d) Customers with low consumption during the day and peak demand in the evening hours Figure 5: this type of customer has a stable low consumption during the hours of the day. From 18:00 hours onwards, there is a significant increase in energy consumption until it reaches its peak at around 21:00 hours.

The 10 remaining clusters have a non-residential profile, and can be categorized in three groups:

a) Bell-like commercial curve Figure 6: this type of customer has a stable low consumption during the hours of the day.

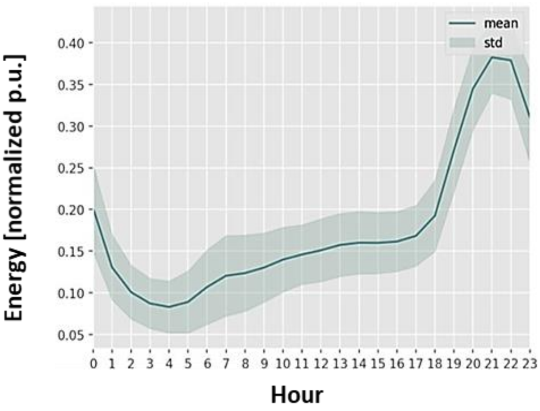


Figure 2. Residential cluster mean of customers with progressive increase in consumption

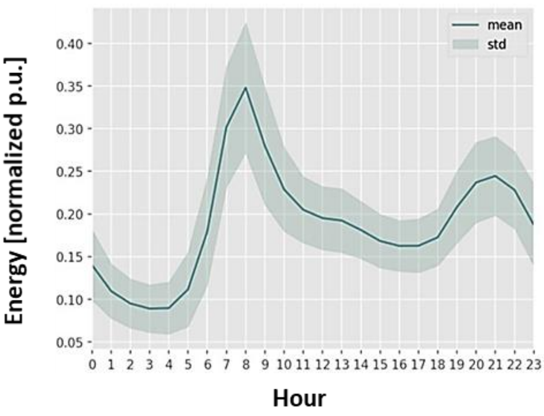


Figure 3. Residential cluster mean of customers with two consumption peaks throughout the day. The higher peak occurs in the morning

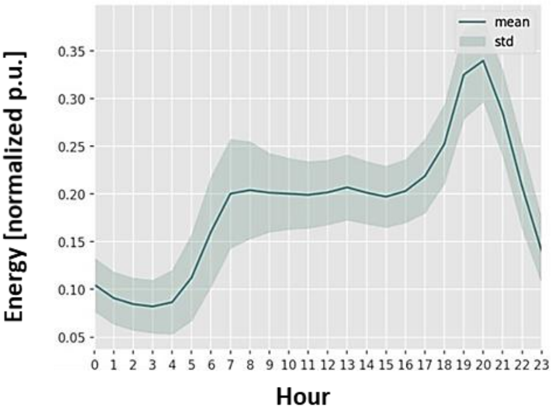


Figure 4. Residential cluster mean of customers with two consumption peaks throughout the day. The higher peak occurs in the evening

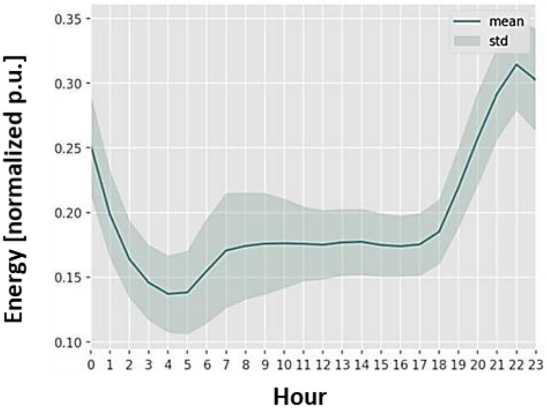


Figure 5. Residential cluster mean of customers with low consumption during the day and peak demand in the evening hours

b) Double-peak commercial curve Figure 7: this type of customer has a bell-shaped curve with two peaks throughout the day. The first peak occurs during the day. Subsequently, there is a decrease in energy consumption, followed by a further increase in energy consumption in the evening hours.

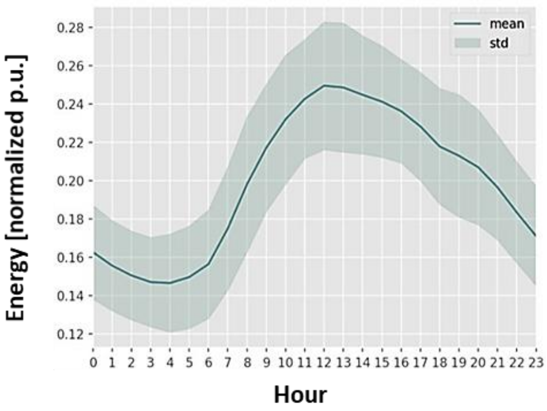


Figure 6. Non-residential cluster mean of customers with a bell-like commercial curve

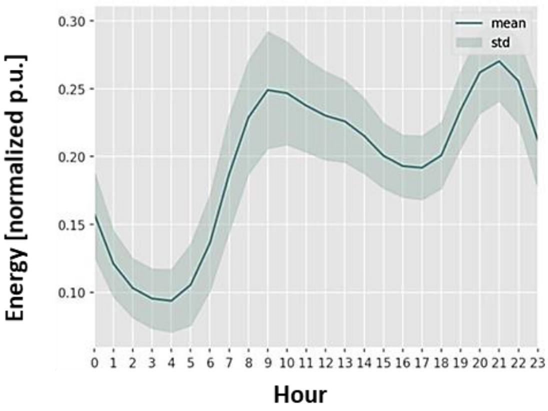


Figure 7. Non-residential cluster mean of customers with a double-peak curve

c) Night-time commercial customers Figure 8: this type of customer is characterised by having their peak energy consumption hours at night and in the early hours of the morning. Their energy consumption starts at 18:00 hours, with peak demand at around 20:00 hours, and they maintain significant consumption during the night and early morning hours, with consumption decreasing from 07:00 hours onwards.

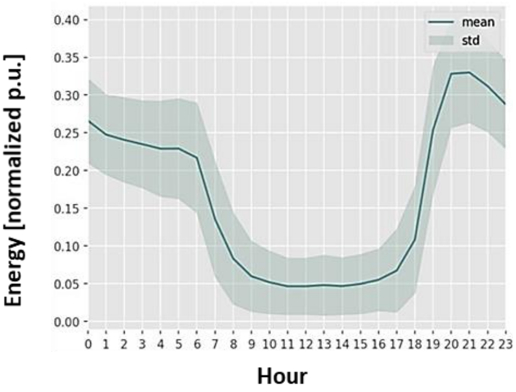


Figure 2. Night-time commercial mean customer

4.1.2. Medium altitude/temperature climate

Fifteen clusters correspond to users located in a medium altitude region. This means users living in a temperate climate. Nine clusters have a residential consumption profile, and can be classified in two types of users:

a) Customers with higher consumption in the early hours of the morning Figure 9: for this type of user, consumption is low during the day and gradually increases, with peak demand at around 18:00 hours.

b) Residential customers Figure 10: this group of residential users has a peak demand in the early hours of the morning that can be associated with the use of electrical appliances. It maintains a stable behaviour throughout the day until 17:00 hours when it increases its demand until it reaches the peak.

The 6 remaining clusters have a non-residential profile, and can be categorized in two groups:

a) Commercial customers Figure 11: the energy consumption is bell-shaped. It is worth mentioning that the start of energy demand for this type of user is around midday.

b) Night customers Figure 12: this type of customer has a higher consumption at night and in the early hours of the morning than in the evening.

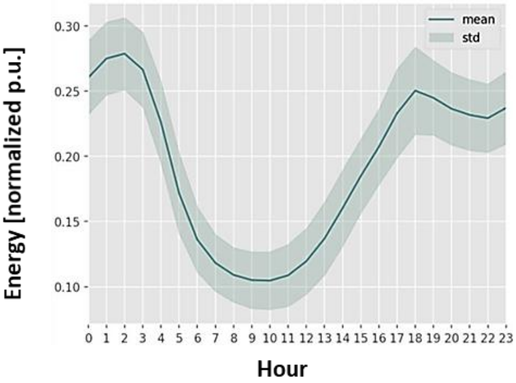


Figure 9. Residential cluster mean of customers living in temperate climates with higher consumption in the early hours of the morning

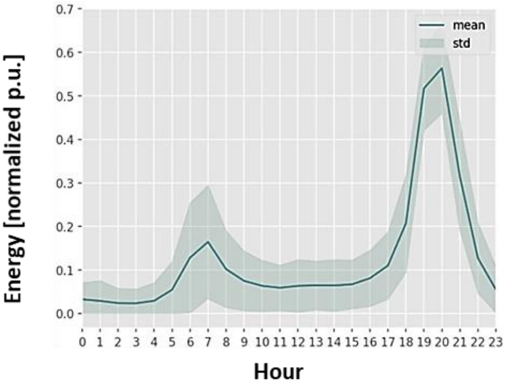


Figure 10. Residential cluster mean of customers living in temperate climates with higher consumption in the evening

4.1.3. All altitudes/all climates

Two final clusters do not present a predominant altitude/climate relation. One of them Figure 13 correspond to a residential profile. This type of residential users is characterized by having peak demand at night and a progressive increase in demand during the day. And the other one Figure 14 presents an industrial-like behaviour, that has a constant consumption throughout the 24 hours with variations in consumption at specific times and of short duration.

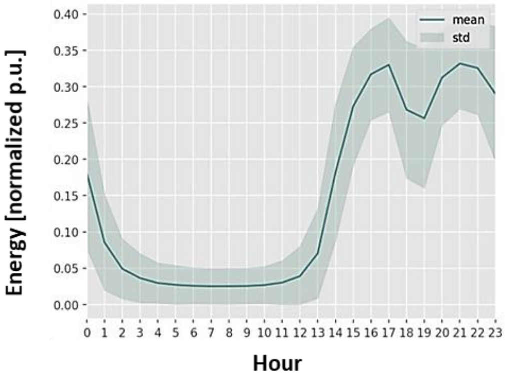


Figure 11. Cluster mean of commercial customers living in temperate climates with higher consumption after midday

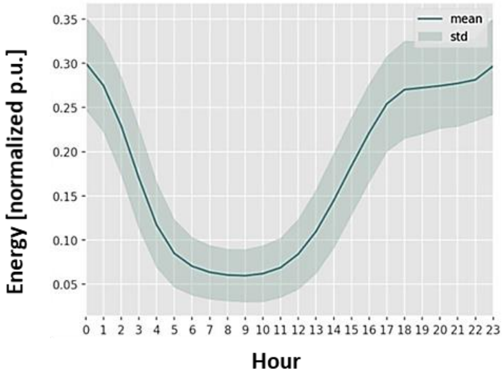


Figure 12. Cluster mean of commercial customers living in temperate climates with higher consumption at night

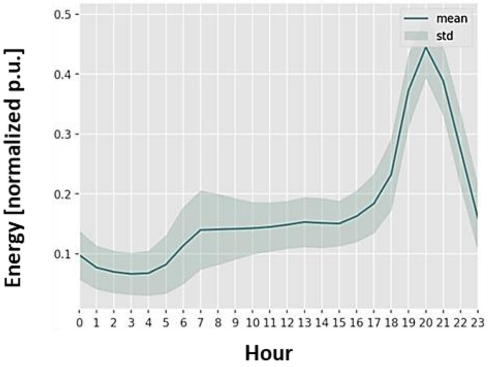


Figure 13. Generic residential profile. This cluster does not have a predominant presence of costumers in any particular altitude/climate

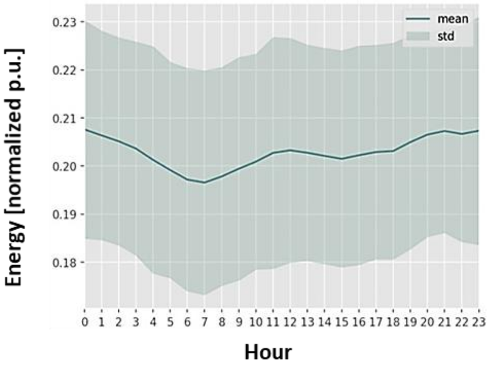


Figure 14. Generic industrial profile. This cluster does not have a predominant presence of costumers in any particular altitude/climate

4.2. User type coherence by cluster

In each cluster we can see how the different attributes mentioned above are distributed, especially the type of user (residential, commercial, industrial) and the socio-economic classification. In this sense, it is remarkable to see that in 11 of the 23 clusters with a commercial profile, most of the samples correspond to users originally categorized as residential users. For instance, in Figure 15 we can see the average load profile of a cluster composed by 542 users, where only 22 (0.5%) of them are registered by the grid operator as a commercial user. In fact, 490 of the users in that same cluster (more than 90%) are registered as residential users. On the other hand, for all of the residential-like clusters, most of the users belonging to those clusters are indeed registered as residential users by the grid operator.

4.3. Pandemic-related effects

Data collected before and after the 24 March 2020 is available for one of the grid operators. On this date, the nationwide quarantine begun in Colombia. Therefore, a simple analysis could be made about the effects of the quarantine in the load profile of the users. 211 meters were selected with the complete data collection for this task including measurements of the grid operator operating in island territory. The pre-pandemic dates cover from January 1, 2020, to March 23, 2020. Pandemic dates (or quarantine dates) cover from March 24 to December 31, 2020.

A separate cluster analysis was made over the data from the two periods of time. K-means inertia on pre-pandemic data suggests 7 clusters while for pandemic data it suggests 5 clusters. To have a better understanding on the changes of the different load profiles, for each cluster learned from the pre-pandemic data, we calculated the mean load profile of the elements of the same cluster but using the pandemic data. Therefore, we would have a direct comparison of the consumption behaviour for the same users in the two periods of time. The results are shown in Figures 16 and 17.

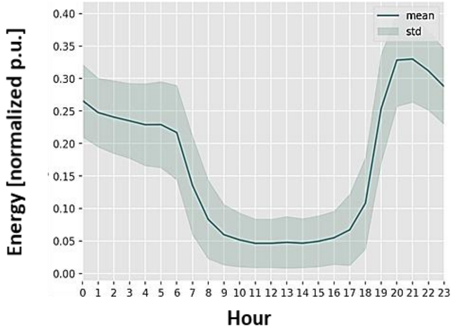


Figure 15. Sample of a commercial cluster centroid. This load profile clearly corresponds to a commercial-night club establishment. However, only 0.5% of the cluster samples are originally categorized as commercial users

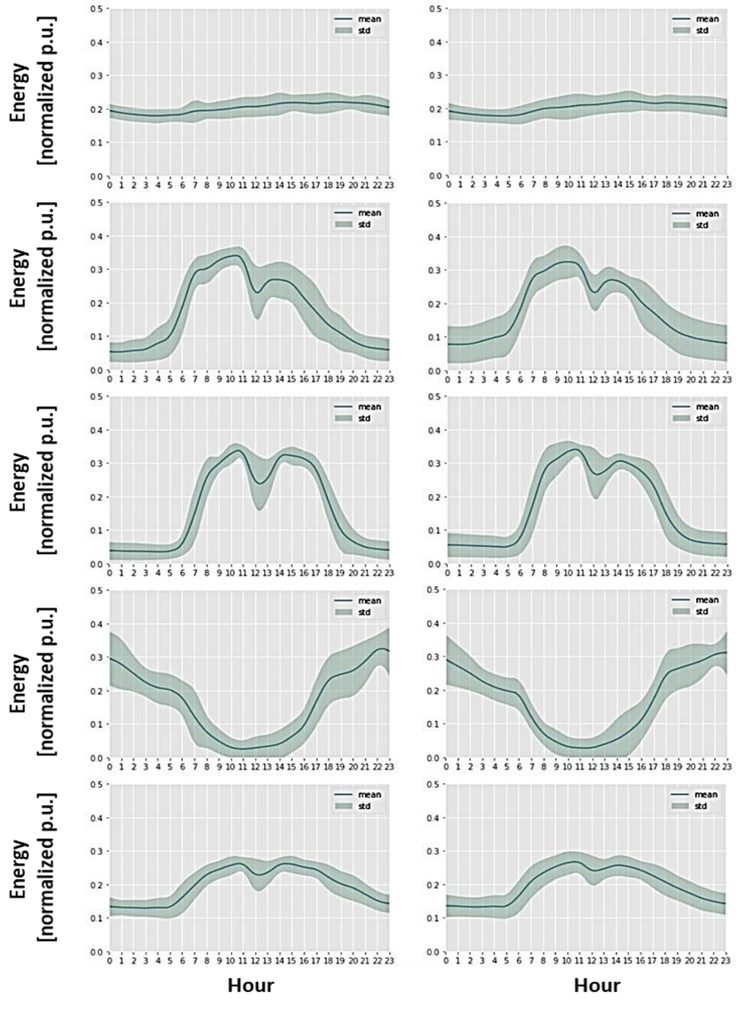


Figure 16. Clusters of users with few pandemic-related changes in energy consumption patterns. Left column: five of the seven clusters found for the pre-pandemic period. Right column: the mean load profile corresponding to the same users in each cluster, but in the pandemic period

Figure 16 confirms that for the first five clusters there was not a significant change in the power demand behaviour caused by the pandemic lockdown and restrictions. Three of these unchanged clusters are mostly composed by commercial customers, one is residential, and the other one is industrial. The clusters for which there was some changes are all composed mostly by commercial users. Figure 17 shows two slight trends: some flattening of the consumption curve, and some changes in at least one of the peaks of demand.

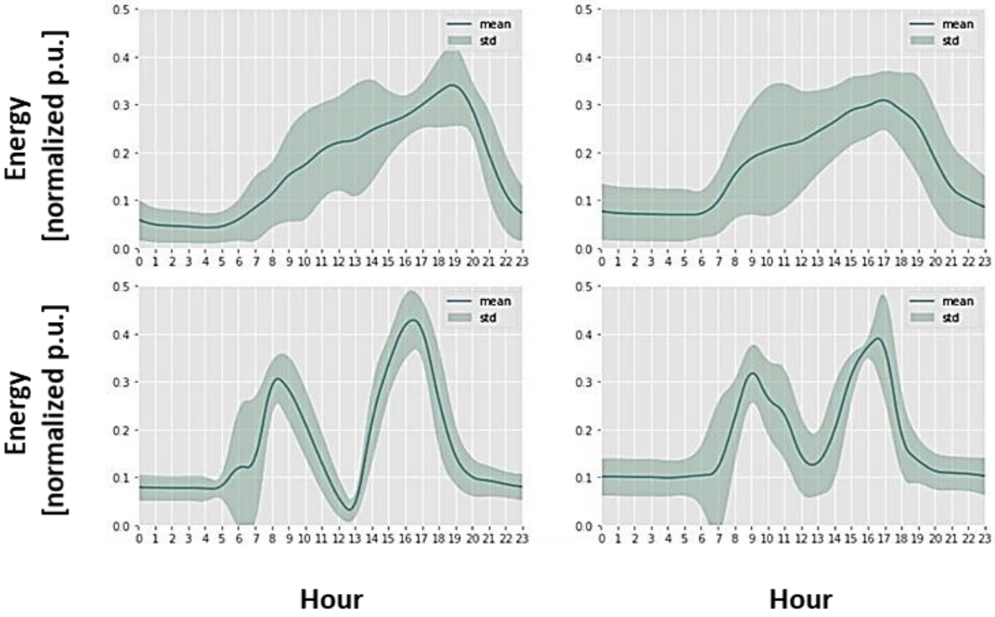


Figure 17. Clusters of users with visible pandemic-related changes in energy consumption patterns. Left column: two of the seven clusters found for the pre-pandemic period. Right column: the mean load profile corresponding to the same users in each cluster, but in the pandemic period

5. CONCLUSION

In this work we presented an energy demand analysis of a new dataset of load profiles from Colombia. Data collection was made by five grid operators and was used to perform a customer segmentation by means of a k-means clustering. The number of clusters was explored by means of the inertia and the direct inspection of the centroids. In this way, 40 clusters were found. The following step was a visual analysis. The objective was to find out what are the predominant types of consumption in a population, understanding what are the main factors that influence the form of consumption.

Climate is the first key explanatory factor for consumption. And if this is so, it means that it is dominated using air conditioning elements. The climate consistency makes it possible to make such conclusions. And it is much more common to use air conditioners in hot climates. In cold climates, on the other hand, it is very unusual to use heaters. The next explanatory factor is the type of user, and in this sense the distinction is simple: there is residential and non-residential behaviour. Non-residential users are identified by having bell-type consumption curves that concentrate their consumption around noon. Through this exercise it was possible to identify users classified as residential by the grid operator but who have a commercial type of behaviour. Likewise, it was possible to identify various types of residential consumption curves, associating some of these curves with typical behaviours of the areas where the meters were located.

This kind of analysis is useful to make an initial categorisation of new customers. Just by knowing the climate and the type of user, a grid operator can narrow down the possibilities associated with the demand curve. This information is also relevant to detect erroneous information registered by the operators, with respect to the type of user. For example, as each network operator has a type of profile registered for each user, we can compare the coincidence of the type of profile registered with the type of profile of the cluster to which it belongs. Thus, we find that in more than 47% of the commercial clusters, most of the users are registered as residential consumers. This anomalous behaviour is not observed in any of the residential profile clusters. In other words, in these clusters, most users are indeed residential.

Another important aspect is to observe how the information from the AMI meters and the characterization analysis allow us to identify the impact that large-scale events can have on energy consumption habits. We saw this in the comparative analysis between pre-pandemic and pandemic seasons. Analysis of the users for whom we had records for the two time periods showed that, surprisingly, most consumption patterns changed very little but, if they did change, there was a tendency for the consumption curve to flatten and, therefore, a change in the magnitude of peak demand.

Overall, we show the effectivity of conducting a characterization process. We also show how this allows for a better understanding of the users and their consumption habits. This can be the starting point for improving the services offered to them and even initiating new business models.

ACKNOWLEDGEMENTS

This research was supported by Electrical Machines Drives Research Group (EMD) from Universidad Nacional de Colombia and project Think Green on the island of San Andres, BPIN 2016000100002 EEDAS ESP.

REFERENCES

[1] H. Li, Z. Wang, T. Hong, and M. A. Piette, “Energy flexibility of residential buildings: A systematic review of characterization and quantification methods and applications,” *Advances in Applied Energy*, vol. 3, Aug. 2021, doi: 10.1016/j.adapen.2021.100054.

[2] S. Yilmaz, J. Chambers, and M. K. Patel, “Comparison of clustering approaches for domestic electricity load profile characterisation-implications for demand side management,” *Energy*, vol. 180, pp. 665–677, Aug. 2019, doi: 10.1016/j.energy.2019.05.124.

[3] X. Wang, K. Smith, and R. Hyndman, “Characteristic-based clustering for time series data,” *Data Mining and Knowledge Discovery*, vol. 13, no. 3, pp. 335–364, Sep. 2006, doi: 10.1007/s10618-005-0039-x.

[4] F. McLoughlin, A. Duffy, and M. Conlon, “A clustering approach to domestic electricity load profile characterisation using smart metering data,” *Applied Energy*, vol. 141, pp. 190–199, Mar. 2015, doi: 10.1016/j.apenergy.2014.12.039.

[5] S. Xu and Massachuse, “Household segmentation by load shape and daily consumption,” 2017.

[6] I. Dent, T. Craig, U. Aickelin, and T. Rodden, “Variability of behaviour in electricity load profile clustering; who does things at the same time each day?,” in *Advances in Data Mining. Applications and Theoretical Aspects*, 2014, pp. 70–84.

[7] I. Benítez, A. Quijano, J.-L. Díez, and I. Delgado, “Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers,” *International Journal of Electrical Power & Energy Systems*, vol. 55, pp. 437–448, Feb. 2014, doi: 10.1016/j.ijepes.2013.09.022.

[8] J. Kwac, J. Flora, and R. Rajagopal, “Household energy consumption segmentation using hourly gata,” *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 420–430, Jan. 2014, doi: 10.1109/TSG.2013.2278477.

[9] T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang, “A new index and classification approach for load pattern analysis of large electricity customers,” *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 153–160, Feb. 2012, doi: 10.1109/TPWRS.2011.2167524.

[10] M. S. Piscitelli, S. Brandi, and A. Capozzoli, “Recognition and classification of typical load profiles in buildings with non-intrusive learning approach,” *Applied Energy*, vol. 255, Dec. 2019, doi: 10.1016/j.apenergy.2019.113727.

[11] M. Quintana, P. Arjunan, and C. Miller, “Islands of misfit buildings: Detecting uncharacteristic electricity use behavior using load shape clustering,” *Building Simulation*, vol. 14, no. 1, pp. 119–130, Feb. 2021, doi: 10.1007/s12273-020-0626-1.

[12] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, “Clustering analysis of residential electricity demand profiles,” *Applied Energy*, vol. 135, pp. 461–471, Dec. 2014, doi: 10.1016/j.apenergy.2014.08.111.

[13] M.-A. Richard, H. Fortin, A. Poulin, M.-A. Leduc, and M. Fournier, “Daily load profiles clustering: a powerful tool for demand side management in medium-sized industries,” in *Conference: ACEEE summer study on Energy Efficiency in Industry*, 2017, pp. 160–171.

[14] A. Rajabi *et al.*, “A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications,” *Energy and Buildings*, vol. 203, Nov. 2019, doi: 10.1016/j.enbuild.2019.109455.

[15] G. Chicco, R. Napoli, and F. Piglione, “Load pattern clustering for short-term load forecasting of anomalous days,” in *2001 IEEE Porto Power Tech Proceedings (Cat. No.01EX502)*, 2001, vol. 2, pp. 1–6, doi: 10.1109/PTC.2001.964745.

[16] A. Satre-Meloy, M. Diakonova, and P. Grünewald, “Cluster analysis and prediction of residential peak demand profiles using occupant activity data,” *Applied Energy*, vol. 260, Feb. 2020, doi: 10.1016/j.apenergy.2019.114246.

[17] A. Lavin and D. Klabjan, “Clustering time-series energy data from smart meters,” *Energy Efficiency*, vol. 8, no. 4, pp. 681–689, Jul. 2015, doi: 10.1007/s12053-014-9316-0.

[18] E. C. Bobric, G. Cartina, and G. Grigoras, “Clustering techniques in load profile analysis for distribution stations,” *Advances in Electrical and Computer Engineering*, vol. 9, no. 1, pp. 63–66, 2009, doi: 10.4316/aece.2009.01011.

[19] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, “Customer characterization options for improving the tariff offer,” *IEEE Transactions on Power Systems*, vol. 18, no. 1, pp. 381–387, Feb. 2003, doi: 10.1109/TPWRS.2002.807085.

[20] M. Bourdeau *et al.*, “Classification of daily electric load profiles of non-residential buildings,” *Energy and Buildings*, vol. 233, Feb. 2021, doi: 10.1016/j.enbuild.2020.110670.

[21] L. Czétány *et al.*, “Development of electricity consumption profiles of residential buildings based on smart meter data clustering,” *Energy and Buildings*, vol. 252, Dec. 2021, doi: 10.1016/j.enbuild.2021.111376.

[22] D. D. Team, “Dask: Library for dynamic task scheduling,” 2016. https://dask.org (accessed Feb. 28, 2023).

[23] J. Reback *et al.*, “Pandas-dev/pandas: Pandas 1.0.1.” Zenodo, 2020, doi: 10.5281/ZENODO.3644238.

[24] W. McKinney, “Data structures for statistical computing in Python,” in *Python in Science Conference*, 2010, pp. 56–61, doi: 10.25080/Majora-92bf1922-00a.

[25] K. R. Shahapure and C. Nicholas, “Cluster quality analysis using silhouette score,” in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2020, pp. 747–748, doi: 10.1109/DSAA49011.2020.00096.

[26] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] “IDEAM - Institute of Hydrology, Meteorology and Environmental Studies,” http://www.ideam.gov.co/ (accessed Oct. 18, 2022).

BIOGRAPHIES OF AUTHORS

Santiago Toledo-Cortés received his bachelor's degree in mathematics from the National University of Colombia. He received a M.Sc. degree in applied mathematics, and currently is a candidate for a Ph.D. degree in systems and computing science, also at the National University of Colombia. He is currently an Assistant Professor in the Department of IT and Process Optimization at Universidad de La Sabana. His research areas include computer vision, natural language processing, multimodal machine learning, kernel methods, and general topics related to machine learning theory and fundamentals. He can be contacted at email: santiagotoco@unisabana.edu.co or https://sites.google.com/unal.edu.co/santiagotoledo-cortes.

Juan Sebastián Lara is a biomedical engineer from Universidad del Rosario and Escuela Colombiana de Ingeniería - Julio Garavito. He has a master's degree in systems and computer engineering from the National University of Colombia. His research interests are related to machine learning and deep learning theory, natural language processing, signal processing, computer vision, and the application of machine learning in different fields. He can be contacted at email: juselara@unal.edu.co and at his webpage: http://juselara.com.

Álvaro Zambrano is an electrical engineer from Universidad Distrital Francisco Jose de Caldas. He received a master's degree in electrical engineering from the National University of Colombia. His fields of study include quality power, modelling of electrical power systems, reliability in power systems and smart grids. He is interested in demand side management for energy efficiency, phasor measurement units, renewable energy integration and software development. He can be contacted at email: aazambranop@unal.edu.co.

Fabio Augusto González Osorio received his bachelor's degree in computing systems engineering, the M.Sc. degree in mathematics from the National University of Colombia, Bogotá, and the MSc and PhD degrees in computer science from the University of Memphis. He is currently a full professor in the Department of Computing Systems and Industrial Engineering at the National University of Colombia, where he leads the Machine Learning, Perception and Discovery Laboratory. He can be contacted at email: fagonzalezo@unal.edu.co and at his webpage: https://dis.unal.edu.co/~fgonza.

Javier Rosero García received his BSC in Electrical Engineering from the Universidad of Valle, Cali, Colombia, in 2002. He worked for construction and maintenance of power systems and substations in Bogotá, Colombia, between 2002 and 2004. He received a PhD degree from the Technical University of Catalonia (UPC) in 2007 and master's in administration from National University of Colombia, in 2020. Dr Rosero received IEEE Senior Member and the IEEE AESS Harry Rowe Mimno award for excellence in technical communications for 2007 from the Aerospace and Electronic Systems Society (AESS) IEEE 2007. He is currently a Full Professor in the Universidad Nacional de Colombia's Electrical and Electronic Engineering Department in Bogota. Dr. Rosero has published more than 100 referred journal and conference papers. His research interests are focused on the areas of modelling, diagnosis and control of electrical machines and drives, electric mobility, and smart grids. He can be contacted at email: jaroserog@unal.edu.co and at his webpage: https://www.researchgate.net/profile/Javier-Rosero-Garcia.