Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Electrical Consumption Patterns through Machine Learning

To cite this article: Amelec Viloria *et al* 2020 *J. Phys.: Conf. Ser.* **1432** 012093

Electrical Consumption Patterns through Machine Learning

Amelec Viloria[1], Alexa Senior Naveda[2], Hugo Hernández Palma[3], William Niebles Núñez[4], Leonardo Niebles Núñez[5]

[1,2] Universidad de la Costa, Barranquilla, Atlántico, Colombia
[3, 5] Universidad del Atlántico, Puerto Colombia, Atlántico, Colombia.
[4]Universidad del Sucre, Sincelejo, Sucre, Colombia.

[1]**Email**: aviloria7@cuc.edu.co

Abstract. Electricity distribution companies have been incorporating new technologies that allow them to obtain complete information in real time about their customers´ consumption. Thus, a new concept called "Smart Metering" has been adopted, giving way to new types of meters that interact in an interconnected system. This will allow to make data analysis, accurate forecasts and detecting consumption patterns that will be relevant for the decision-making process. This research focuses on discovering common patterns among customers from data collected by smart meters.

1. Introduction

Electric power represents a very important resource during the last years. The advance of technology and its massive arrival in cities and homes has generated a high demand that is increasing every day. From street lighting to lighting and household appliances, such as refrigerators and televisions, need to be constantly supplied with electricity for keeping running [1] [2].

This strong demand has led to concentrate research efforts on improving the efficiency of the generation, distribution and consumption of electricity in homes and industries. With the same objective in mind, electricity distribution companies have been incorporating new technologies that allow them to obtain complete information in real time on the consumption of their customers [3].

For electricity distribution companies, it is very important to make projections of total consumption in future periods. These projections allow them to have a better basis for decision making, and thus reduce the risk involved in uncertainty. Depending on the quality of the data, it is possible to make forecasts with very high precision, for which there are several models and also researches that are dedicated to applying and evaluating these models [4].

In this instance, the discipline that has had greater boom is Data Mining [5] [6] [7], since it provides a series of techniques that allow grouping and effectively characterizing patterns of electricity consumption; secondly, machine learning allows applying accurate forecast models. These analyses are not only aimed at favoring and improving distribution operations on the part of companies, but also have the aim of offering consumers an analysis of their own consumption, and thus promoting a conscious use to safeguard energy efficiency.

This research studies the recognition of patterns and the application of predictive models for electricity consumption, based on data provided by intelligent meters. The following main topics will be addressed [8] [9]:

- Advances and applications in Intelligent Metering
- Research Guidelines on Pattern Recognition and Forecasting of Electricity Consumption

- Data mining techniques applicable to pattern recognition and forecasting of electricity consumption.
- Pre-processing, representation and clustering of consumption patterns.
- Time series and forecast models.
- Application and evaluation of forecast models for electricity consumption.

2. Databases of electricity consumption

The data used in the experimentation were obtained from intelligent meters of electrical consumption. These meters are owned by an electricity distribution company in the Caribbean region of Colombia. The electricity consumption database consists of two files (.csv), containing measurements recorded between the years 2017 and 2018.

The measurements are expressed in KWh, and are recorded in intervals of 15 minutes, starting at 00:00:00 hrs. In this way, each meter is expected to record 96 consumption data daily, thus completing 24 hours of consumption at 15-minute intervals. Table 1 shows details of the database files.

It considers:
Expected number of records per meter = 98 * Number of days
Expected number of records = Number of meters * Expected number of records per meter

Table 1. Details BDD Consumption 2017

File BDD 2017	
Size (GB)	1,99
Range of readings (start/end)	02-05-2017 / 30-11-2018
N° of months	12
N° of days	289
N° of meters	1.345
Expected number of total records	34.785.410
Actual no. of total records	16.061.878
Expected number of records per meter	28.410
Average actual number of records per meter	13.950
No. of meters with above-average records	698
No. of meters with low average records	701

Together, the two files record the consumption of 1,687 customers, for a total of 54,078,592 measurements. Figure 1 best express the distribution of the number of records in 2018.

The database of electricity consumption has different attributes, which are: Reading Date, CIM Code, Reading Type Description, Value, Reading Quality Code, Received Date, Mac Address and Meter Serial. Each attribute is briefly described below [10] [11] [12].

- Reading Date: Date and time of reading.
- Reading Type Description: Description of the type of data, where "Delivered 15 min Energy KWh" is the active energy supplied to the customer.
- Value: Energy consumed (without considering the constant), in an interval of 15 min.
- Reading Quality Code: Refers to the rating or quality of the reading sent by the meter. It can take the following values: Configuration Changed, Overflow, DST Flag, Partial Interval, LP Recording Stopped, Clock Set Backward, Test Data, Clock Set Forward, Long Interval, Data Valid, Power Fail, Skipped Interval. Most registers have the attribute "Data Valid" and "DST Flag", thus indicating a correct reading of consumption.
- Received Date: Date the information was received.
- Mac Address: Meter network identifier.

- Serial Meter: Serial number of the meter.

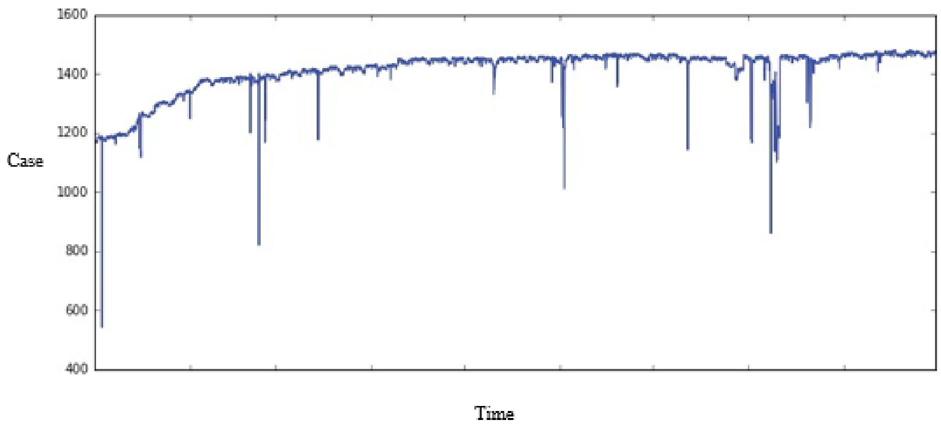

Figure 1. Number of measurements by date, year 2018

3. Processing

The project was developed using the Python programming language, version 3.5. The vast number of specialized libraries for data analysis make this language the most appropriate for conducting experiments [13].

In order to cluster and group customers with similar consumption patterns, it is necessary to represent in some way the individual consumption of each customer. Four types of representation were created in this research. The first one consists of a daily consumption pattern; the second one is a weekly consumption pattern (Monday through Friday); the third one describes a weekend consumption pattern (Friday and Sunday); the fourth one is a pattern based on the Fourier Fast Transform.

For each performance, a complete analysis was made and another based on peak hours. The complete analysis considers all consumption measurements at any time and date. Meanwhile, peak hour analysis is performed on all measurements recorded between 18:00 and 23:00 hours during the months of April, May, June, July, August and September [14]. It is important to include this analysis, since in this period charges are incorporated as additional winter energy for customers with BT1 tariff. In addition, the charge for power consumed is approximately 10 times higher than the charge for power during off-peak hours for customers with BT4 or AT4 tariffs.

For both types of analysis, the data were grouped into two modalities: general summary and summary by season of the year. Each summary contains the consumption pattern of all clients in the study. The general summary considers the whole range of possible dates according to the type of analysis; while the seasonal summary restricts dates according to each station for later analysis.

3.1 Consumption pattern

A customer's daily consumption pattern consists of a vector of N continuous measurements at 15-minute intervals; each vector element corresponds to the average consumption at each time interval. The N value depends on the type of schedule analysis used. In the case of the full schedule analysis, N will take a value of 96, which is equivalent to a lapse of 24 continuous hours, recorded from 00:00 hrs. to 23:45 hrs. In the case of peak hour analysis, N will take a value of 20, which is equivalent to a lapse of 4 continuous hours, recorded between 18:00 and 23:00 hrs.

To apply this representation, only the days on which the customer has a complete record of 98 intervals were considered. In this way, the average consumption in each hour (hi) will always be calculated on the same n-value.

Figure 2 shows a general summary graph, with the daily consumption pattern (PD) of all customers, using a full analysis (FULL).

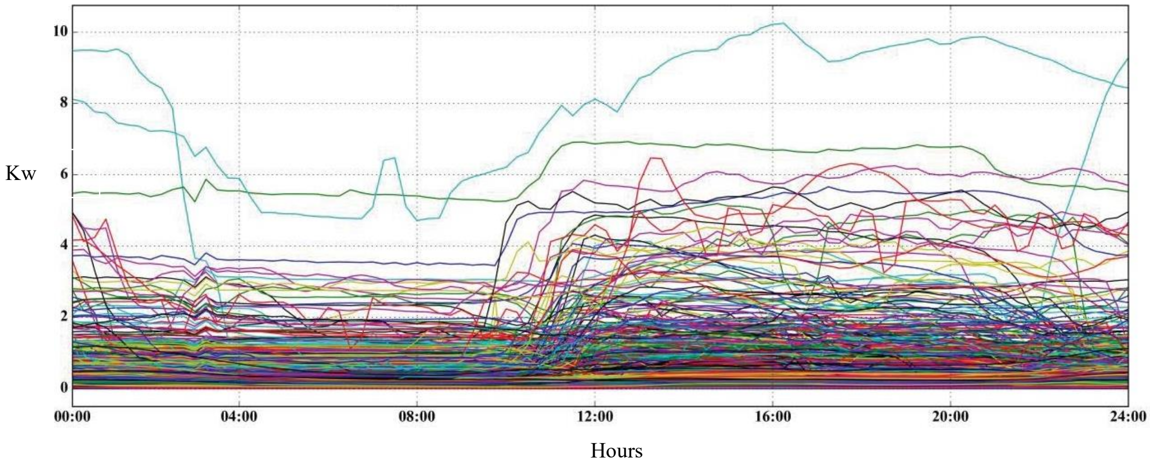

Figure 2. Overview - PD | FULL

Figure 2 reveals that there are customers with consumption much higher than the rest, reaching records even higher than 10 KWh. There is also an increase in consumption levels from around 12:00 hrs. onwards. The image shows a general graph of the consumption pattern based on the Fourier Transform (PF) for all customers, with full schedule analysis.

4. Results

Fast Fourier Transform is a reversible function, i.e. it can be applied inversely to obtain the original series in the time domain. In addition, for processing purposes, the number of input samples must be two power; thus, the calculation time is considerably reduced.

In this research, a number of 2578 samples were used to construct the consumption pattern with full schedule analysis. For peak time analysis, the number of samples to be used were 685. Both sample quantities shall consider a continuous time series of 22 and 24 days respectively.

To determine the time series to be used, a graphical analysis was made of the number of readings over the year 2017. From this, it was possible to identify a period between April 10 and May 10, which has a stable number of meter records greater than 1454. Figure 3 shows a general graph of the consumption pattern based on the Fourier Transform (PF) for all customers, with full schedule analysis.

After generating the different consumption patterns for each client, the research proceeds to the application of the clustering algorithm to group those clients that present a similar behavior. The algorithm chosen is the K-means.

In order to obtain a more accurate result, the optimal number of clusters to be used was evaluated using the Silhouette metric. To do this, the K-means algorithm was applied several times, using different number of clusters. The number of clusters to be evaluated was defined between 4 and 10, based on the number of tariffs with the highest number of clients in the study. Then, each result was evaluated with Silhouette, and according to it, the optimal number of clusters was established at 10, for each consumption representation.

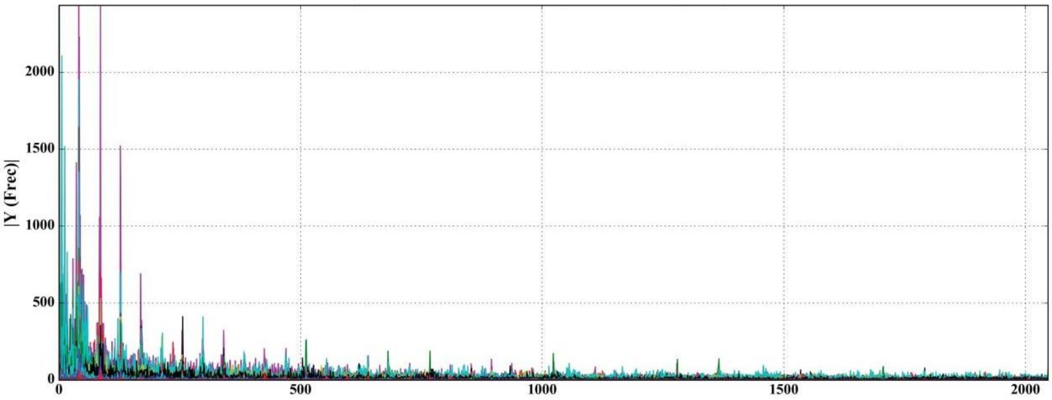

Figure 3. Overview - PF | FULL

Finally, the K-means algorithm is applied for the last time with the number of clusters determined in the previous step. The parameters used in the implementation through the sklearn library are the following [15]:

init='k-means++', n_init=10, max_iter=300, tol=0.0001, precompute_distances='auto', verbose=0, random_state=None, copy_x=True, n_jobs=1, algorithm='auto'

Table 2 shows that, despite being a small group of customers, each cluster represents more than 14% of total consumption. In the case of cluster 1, customers represent only 1.052 % of the total number of meters, but as a percentage of consumption, they represent 12.265% of the total. To ratify what has been seen in Figure 3, cluster 1 has an average consumption much higher than the rest, with a value of 3.478 KWh.

Table 2. indicators, Representative clusters - PD | FULL | General

Cluster	Sum consume (KWh)	Maximum value (KWh)	Minimum value (KWh)	Average (KWh)	Percentage meters (%)	Percentage consume (%)
1	4543,457	6,124	0,478	3,028	1,147	12,369
5	4368,254	1,754	0,000	0,654	7,111	12,785
7	5410,974	2,345	0,002	0,741	4,578	15,468
10	4520,544	0,876	0,001	0,214	15,583	12,647

Table 3 shows that, in terms of averages, clusters 8 and 9 are similar, although the latter registers a slightly higher level of consumption. On the other hand, clusters 1 and 5 have a higher average than the previous ones, and they also differ from each other. Cluster 1, which has 23 customers, has an average consumption of 2,986 KWh; while cluster 5 has an average consumption of 6,352 KWh, with customers reaching values of up to 8,457 KWh around 21:00 hrs.

Table 3. Descriptive Indicators, Representative Clusters - PD | HP | General

Cluster	Sum consume (KWh)	Maximum value (KWh)	Minimum value (KWh)	Average (KWh)	Percentage meters (%)	Percentage consume (%)
1	1876,366	4,478	1,074	2,987	1,247	14,587
5	1252,454	9,147	2,358	5,365	0,369	12,365
8	1365,741	3,159	0,047	2,578	4,478	18,174
9	1746,025	2,747	0,247	1,785	3,254	14,647

5. Conclusions

The recognition of patterns and the forecast of electricity consumption is a measure that cannot wait. The need to increase energy savings and efficiency forces distribution companies to invest in the implementation of technology such as smart meters. That is why today the smart metering system is making a very fast entrance to the electricity market or sector.

In the present study, the discovery of patterns of electricity consumption was dealt with, in order to know the behavior of customers. This would allow distribution companies to create recommendations or flexibilize tariffs, seeking to improve consumption efficiency and cost optimization. To this end, the K-means clustering algorithm was applied, the results of which allowed to identify up to 10 groups of customers with similar behavior, using different consumption representations. In this way, it was possible to identify those customer segments that present the greatest impact on demand, whether at peak times or without time restrictions.

References

[1] Pretnar, A. The Mystery of Test & Score. Ljubljana: University of Ljubljana. Retrieved from: https://orange.biolab.si/blog/2019/1/28/the-mystery-of-test-and-score/ (2019).

[2] Joana M. Abreu, Francisco Camara Pereira, Paulo Ferrao, using pattern recognition to identify habitual behavior in residential electricity consumption, Energy and Buildings, Vol. 49, June 2012, pp. 479-487, ELSEVIER

[3] Yasser, A. M., Clawson, K., & Bowerman, C.: Saving cultural heritage with digital make-believe: machine learning and digital techniques to the rescue. In Proceedings of the 31st British Computer Society Human Computer Interaction Conference (p. 97). BCS Learning & Development Ltd. (2017).

[4] Khelifi, F. J., J. (2011). K-NN Regression to Improve Statistical Feature Extraction for Texture Retrieval. *IEEE Transactions on Image Processing*, 20, 293-298.

[5] Abdul Masud, M., Zhexue Huang, J., Wei, C., Wang, J., Khan, I., Zhong, M.: Inice: A New Approach for Identifying the Number of Clusters and Initial Cluster Centres. Inf. Sci. (2018). https://doi.org/10.1016/j.ins.2018.07.034

[6] Martins, L.; Carvalho, R.; Victorino, C.; Holanda, M.: Early Prediction of College Attrition Using Data Mining. 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1075-1078 (2017)

[7] Leo Breiman, Random Forests, Machine Learning, Vol. 45, Issue 1, October 2001, pp. 5- 32, Springer.

[8] A.S.Ahmad, et al., A review on applications of ANN and SVM for building electrical energy consumption forecasting, Renewable and Sustainable Energy Reviews, Vol. 33, May 2014, pp. 102–109

[9] Witten, I.; Frank, E.; Hall, M.; Pal, C.: Data Mining Practical Machine Learning Tools and Techniques. Elsevier 4th Ed, pp. 167-169 (2016).

[10] Clustering – EcuRed. Disponible vía web en http://www.ecured.cu/Clustering. Revisado por última vez el 29 de marzo de 2017.

[11] Sanchez L., Vásquez C., Viloria A., Cmeza-estrada (2018) Conglomerates of Latin American Countries and Public Policies for the Sustainable Development of the Electric Power Generation Sector. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.

[12] Perez, R., Inga, E., Aguila, A., Vásquez, C., Lima, L., Viloria, A., & Henry, M. A. (2018, June). Fault diagnosis on electrical distribution systems based on fuzzy logic. In *International Conference on Sensing and Imaging* (pp. 174-185). Springer, Cham.

[13] Perez, Ramón, Carmen Vásquez, and Amelec Viloria. "An intelligent strategy for faults location in distribution networks with distributed generation." *Journal of Intelligent & Fuzzy Systems* Preprint (2019): 1-11.

[14] Chakraborty, S., Das, S.: Simultaneous variable weighting and determining the number of clusters—A weighted Gaussian algorithm means. Stat. Probab. Lett. 137, 148–156 (2018). https://doi.org/10.1016/j.spl.2018.01.015

[15] Bucci, N., Luna, M., Viloria, A., García, J. H., Parody, A., Varela, N., & López, L. A. B. (2018, June). Factor analysis of the psychosocial risk assessment instrument. In International Conference on Data Mining and Big Data (pp. 149-158). Springer, Cham.