NoSQL Data Storage and Clustering Large Volume of Data from Smart Metering Systems with Impact on Electricity Consumption Peak and Tariff Settings

Simona-Vasilica Oprea
Adela Bâra
Bucharest University of Economic Studies, Romania
simona.oprea@csie.ase.ro, bara.adela@ie.ase.ro
Dan Preoțescu
Romanian Energy Center

Abstract

Recently, large volumes of electricity consumption data are pouring constantly from smart meters and other sensors that count for millions or even milliards of records. Our purpose in this paper is to handle such data and extract valuable information until it becomes stale. Sometimes, additional data such as meteorological, motion-sensitive, door position data, results from surveys, tariffs, etc. come together with the electricity consumption and increase the number of records. In this case, NoSQL solutions are utilized to process and analyze the entire volume of data. In this paper, we propose a data processing framework for electricity data set that comes from a trial smart metering implementation period that took place from 1st January to 31st December 2010 in Ireland. The main purpose is to cluster the consumers based on similarities regarding theirs 30-minute consumption, show their impact on the electricity consumption peak that could be used as an input in establishing real-time tariffs based on peak coefficient.

Key words: clustering, big data, NoSQL, electricity consumption, real-time tariff
J.E.L. classification: L94, C55, C38, C92, E21

1. Introduction

The progress of the smart metering systems and sensors have led to big data solutions for processing and analyzing large volumes of data that requires attention as soon as the data is fresh and can assist the decision makers to enhance business (Diamantoulakis, Kapinas and Karagiannidis, 2015), (Oprea, Bâra and Diaconita, 2019), (Zhou, Yang and Shen, 2017). Data sources are also multiple not coming only from sensors but social and professional networks, complex questionnaires that are usually gathering and revealing useful insights related to the consumers' customs, preferences and expected behaviors.

After handling large amount of data through means of NoSQL data bases and programming environments, clustering the electricity data and analyzing the contribution of each consumer or group to the consumption peak are proposed in the current paper. The analysis of the clusters is significant in terms of tariff settlement for the electricity suppliers. Using boxplots at the level of each cluster, the consumption at peak is analyzed for its variability to the median values. They also uncover the outliers and the significance of the whiskers.

This paper is structured in 5 sections. In the second section, a couple of related paper to the data clustering aspects were enumerated. The third section is dedicated to the research methodology that is described using a comprehensive flowchart. Results are depicted in the fourth section, whereas in the fifth section the conclusion is drawn.

2. Literature review

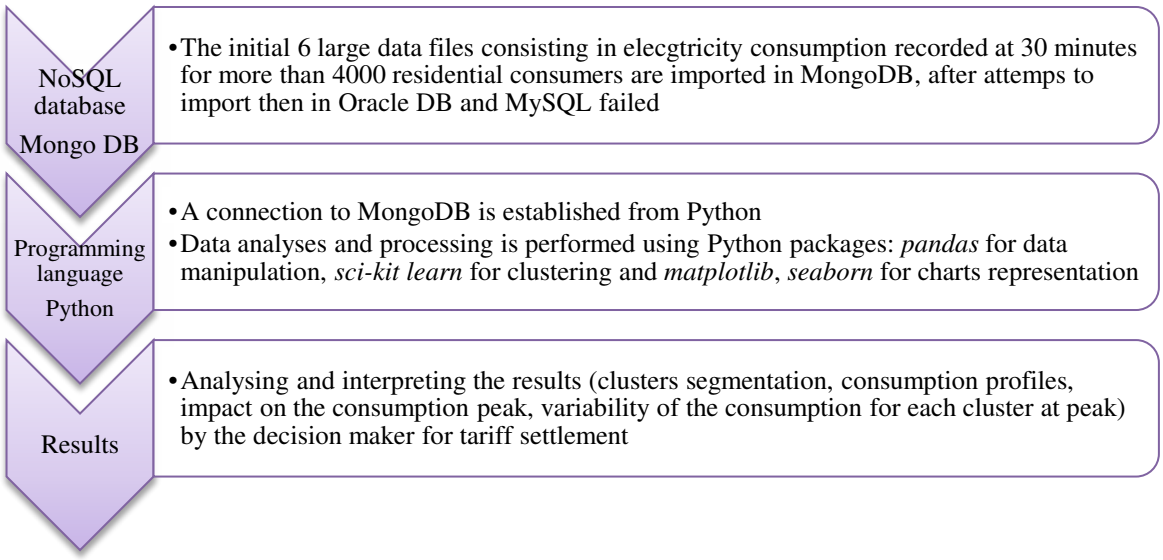
Clustering the electricity consumption data has had several objectives. Many papers distinctly have approached this evolving topic that continues to grab attention of researchers as a consequence of the changing context of the demand side management strategies. (Chicco, 2012) provided an overview on the performance evaluation of some clustering methods in case of electricity consumption without considering the large volume of data and NoSQL alternative for the relational databases. A fuzzy-oriented clustering model was developed in (Zhou, Yang and Shao, 2017) considering the monthly consumption of residential consumers and the concept of pattern mining. However, grouping the electricity consumers based on monthly consumption is a particular approach. (Wang et al., 2016) assessed the electricity consumption behavior and tendency of the consumers considering big data solutions. (Hernández et al., 2012) approached several classification and clustering methods to identify patterns in electricity loads for industrial business. (Kim, 2015) used clustering algorithms to answer to the question "Electricity consumption and economic development: Are countries converging to a common trend?" (Motlagh et al., 2015) assessed the impact of the local electricity production on the electricity consumption behavior considering the renewable sources availability and the feed-in tariff. Clustering methods were also applied for detecting theft using smart metering systems and consumption patterns (Jokar, Arianpoo and Leung, 2016). Also, clustering the electricity consumers assist the decision makers in creating electricity profiles or patterns that are useful for many applications: settlement, understanding behaviors and future tendencies, market strategies, tariff setting, etc. (Gouveia and Seixas, 2016), (Yildiz et al., 2017), (Hayn, Bertsch and Fichtner, 2014), (Rodrigues et al., 2003).

3. Research methodology

The electricity consumption data is stored in six text data files, with three attributes: consumer id for over 4000 residential consumers, date and time information and consumption data recorded at 30 minutes. The consumption data files size is 2,539 MB containing over 157 million of rows. Additional files containing results of pre and post-surveys, proposed time-of-use tariffs, and reports are available (UMassTraceRepository, Smart* Data Set for Sustainability, 2007), (Comission for Energy Regulation - CER, 2011).

The relational databases (e.g. Oracle Database and MySQL) failed to import or process the consumption files. Taking into account the large volume of the data sets, NoSQL solution MongoDB is used to store the data and Python to process and analysis it. The flowchart of the proposed methodology is shown in Figure 1.

Figure no. 1. Flowchart



Source: Authors' contribution

After the processing of the data stored in NoSQL database (MongoDB) system by using Python code, some of the data were sent to a second processing phase to uncover significant information and knowledge for decision makers. Combinations of various techniques, such as grouping, aggregation and group functions, descriptive statistics, statistical analysis, plotting and data mining using *.sql* queries and *.xls* sheets, etc. are applied to enhance the analysis capabilities.

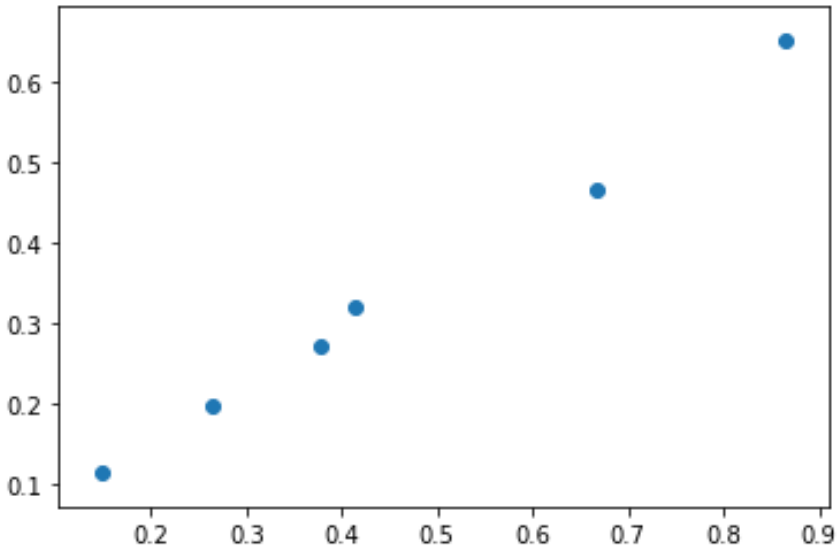
As for the clustering method, in order to group the electricity consumers based on the similarities among their individual consumption and to reveal the characteristic daily profile, we applied k-means method available in sci-kit learn provided by Python. For identifying the number of optimal number of clusters, the Elbow method is used.

4. Results

The electricity consumption data is processed in Python using the facilities of *dataframes* and Spyder as IDE for a better visibility and control of variables. Initially, the data is imported from Mongo DB to *pandas dataframe* with a structure of 168 (7x24) columns meaning the hourly consumption for each hour of the week and 4225 rows meaning the ID meter.

Before processing, the consumers are clustered into 6 clusters using k-means algorithm implemented by *sci-kit learn* package in Python, considering the Elbow method in selecting the number of clusters. Actually, the number of clusters depends on the data sample so it can be considered variable. Hence, the optimal number of clusters is 6. In Figure 2, the centroid of each cluster is given.

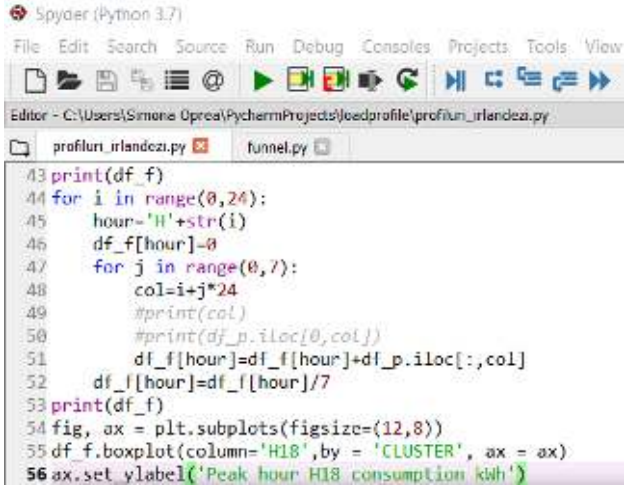
Figure no. 2. Centroids of the clusters



Source: Authors' contribution

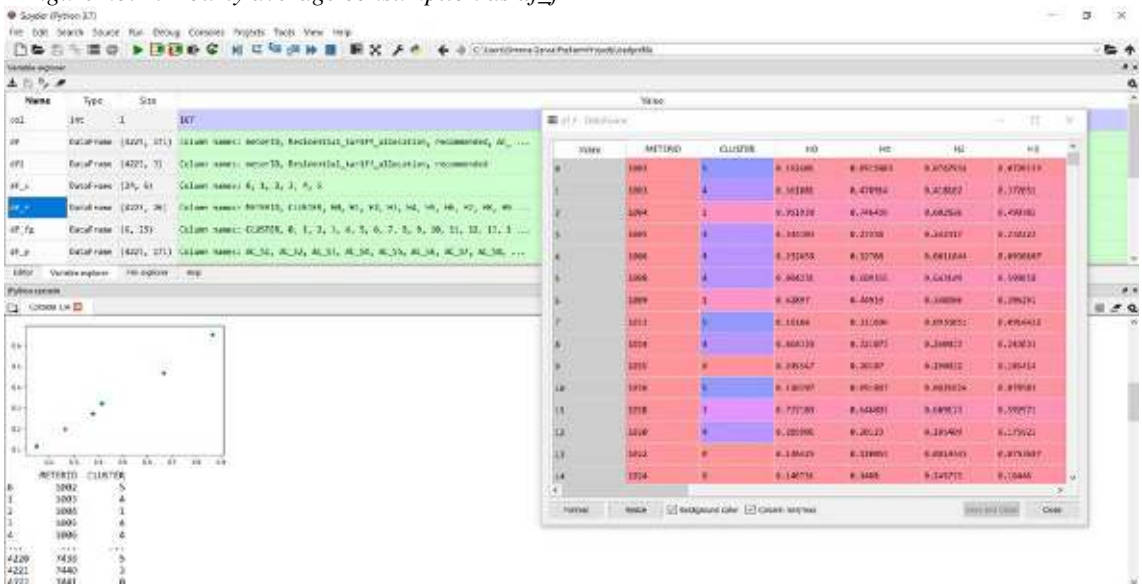
To represent the hourly consumption for each meter id as in *df_f*, the consumption from the same hour of the day was summed up and then divided by 7, obtaining the average hourly consumption for 24 hours regardless the days of the week (Figure 3). So instead of seven daily profiles as in Figure 4, that would have been difficult to represent, we obtained one single daily profile for each cluster as in Figure 5.

Figure no. 3. Data processing in Python from the initial dataframe df_p (4225x171) to df_f (4225x26)



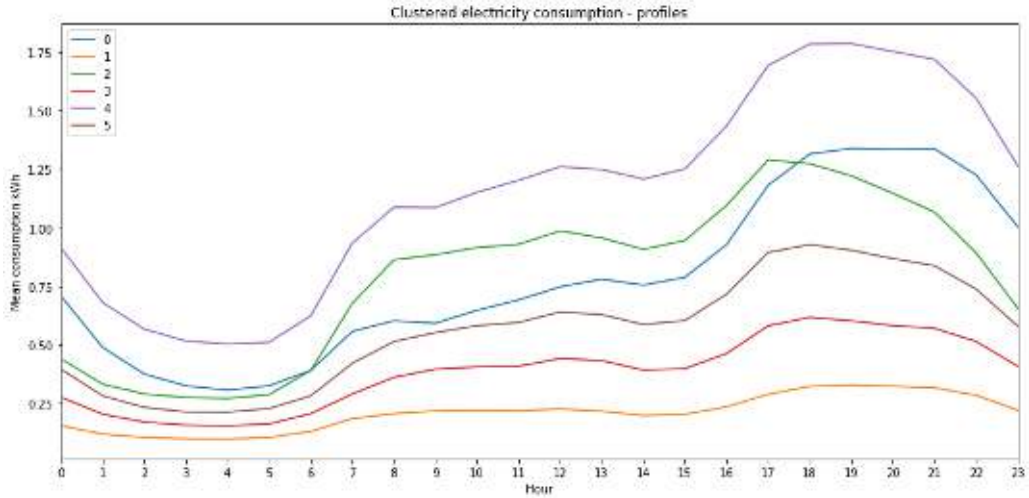
Source: Authors' computation

Figure no. 4. Hourly average consumption as df_f



Source: Authors' computation

Figure no. 5. Load profile for each cluster

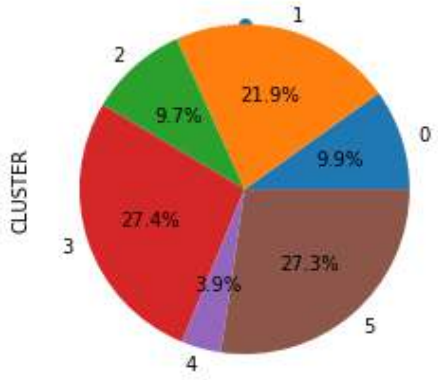


Source: Authors' computation

Based on similarities of the electricity consumption, the residential consumers are grouped with k-means algorithm into six clusters (from 0 to 5). The optimal number of clusters is given by Elbow method. Starting from the clusters, the hourly average consumption is shown in Figure 5. These daily load profiles differ mainly in amplitude. However, one exception is provided by cluster 0 - blue and cluster 2 - green that intersect twice. Cluster 2 - green has a shorter consumption peak period (around 17) compared with cluster 1 (from 18 to 21), although during the day (from 6 to 17), cluster 2 consumption is higher. Cluster 1 - orange has a special allure characterized by lowest, almost flat daily consumption. Cluster 3 - red and 5 - brown are similar, differing only in amplitude, while cluster 4 - purple is characterized by highest clear morning, noon and evening peaks and night off-peak. Its off-peak consumption level exceeding other clusters' peak consumption level (cluster 3 - red and cluster 1 - orange).

The distribution of consumers allocation in the above-mentioned clusters is given in Figure 6. Most of the consumers (more than 76% of the total) belong to clusters with the lowest electricity consumption, namely cluster 1 – orange, cluster 3 – red and cluster 5 – brown. Cluster 4 – purple with the highest consumption consists in only 4% of the total consumers, while the other two clusters green and orange sum up around 20% of the total consumers.

Figure no. 6. Segmentation of members for each cluster

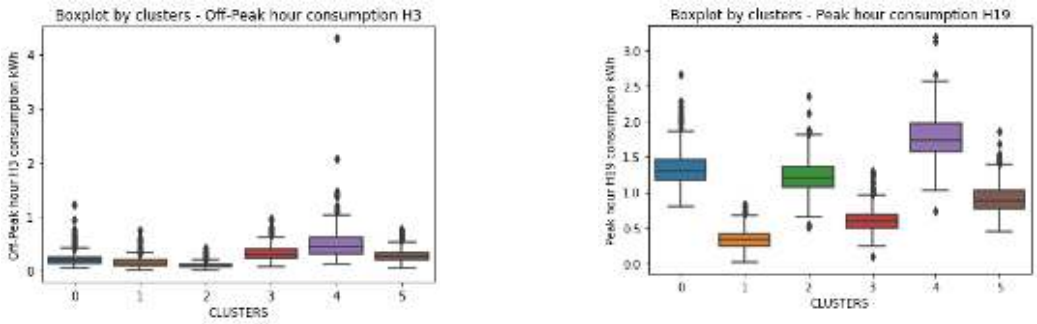


Source: Authors' contribution

The variability of the consumption profiles at peak hour (and at off-peak too) can be analysed by box plotting the consumption data for each cluster as in Figure 7. The highest variability of data belongs to cluster 2 while the smallest belong to cluster 1. Half of the data in case of cluster 2 is spread from 1.4 to 1.8, one whisker that goes up to 2.3 and the other one to 0.8, and biggest outlier to 3.4.

Another interesting feature of the input data is the spread of the hourly consumption values at the level of each cluster. Let us pick two specific hours: evening peak hour (H19) and night off-peak hour (H3).

Figure no. 7. Box plotting the consumption at peak and off-peak hour for each cluster



Source: Authors' computation

As it can be seen in box plots provided in Figure 7, the spread of the consumption values at the level of clusters at night is smaller and dispersion of the values around the mean for each cluster is also smaller. Also, the highest variety of data is recorded for cluster 4 – purple with the highest consumption level.

The contribution to the consumption peak/off-peak of each cluster is calculated as percentage of the cluster consumption at peak/off-peak hours from the total consumption. The clusters contributions are summurized as following: cluster 0 – peak 21%/off-peak – 18%, cluster 1 – peak 8%/off-peak – 11%, cluster 2 – peak 17%/off-peak 23%, cluster 3 – peak 16%/off-peak 14%, cluster 4 – peak 28%/off-peak – 18%, cluster 5 – peak 10%/off-peak 16%. Base on these contributions, the tariff system can be designed so that to reflect the apport of each cluster to the consumption peak/off-peak. However, it is necessary to repeat calculation at a few months as the clusters can significantly change in terms of number of members or contributions to the peak/off-peak.

5. Conclusion

The large volume of raw electricity consumption data consisting in more than 157 million of records from smart meters was stored and procced in Mongo DB that is a NoSQL database whereas the relational databases failed. Then, for further processing the data was transferred to Python and other tools for extracting useful insights for decision makers.

As it is well-known that the pattern of the consumption is sensitive to the level of the electricity tariff rates, in this study the consumption data of 4225 residential consumers is analyzed. In this context, we aimed at identifying the contribution group of consumers or clusters to the consumption peak or off-peak and accordingly encourage or discourage the consumption.

The impact on the consumption peak is measured as a coefficient that is recommended to be timely computed to uncover the behavior of individual consumers or certain consumers grouped in clusters. This way, only the consumers that highly impacts the consumption peak would be charged with higher hourly rates, whereas the consumers that significantly contributes to the off-peak consumption should be charged with lowest rates.

As further work, formalizing the impact on the electricity tariff is the next step. Also, more data sets will be analyzed and the sensitivity to the tariff rates will be considered. We should further analyze whether the rate should be established based on the individual consumer or groups of consumers' contributions to the consumption peak or off-peak.

6. Acknowledgment

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CCCDI – UEFISCDI, project title "Multi-layer aggregator solutions to facilitate optimum demand response and grid flexibility", contract number 71/2018, code: COFUND-ERANET-SMARTGRIDPLUS-SMART-MLA-1, within PNCDI III

7. References

- Chicco, G., 2012. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy.*
- Gouveia, J.P. and Seixas, J., 2016. Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy and Buildings*.
- Hayn, M., Bertsch, V. and Fichtner, W., 2014. Electricity load profiles in Europe: The importance of household segmentation. *Energy Research and Social Science*.
- Hernández, L., Baladrón, C., Aguiar, J.M., Carro, B. and Sánchez-Esguevillas, A., 2012. Classification and clustering of electricity demand patterns in industrial parks. *Energies.*
- Jokar, P., Arianpoo, N. and Leung, V.C.M., 2016. Electricity theft detection in AMI using customers' consumption patterns. *IEEE Transactions on Smart Grid.*
- Kim, Y.S., 2015. Electricity consumption and economic development: Are countries converging to a common trend? *Energy Economics.*

- Motlagh, O., Paevere, P., Hong, T.S. and Grozev, G., 2015. Analysis of household electricity consumption behaviours: Impact of domestic electricity generation. *Applied Mathematics and Computation.*
- Rodrigues, F., Duarte, J., Figueiredo, V., Vale, Z. and Cordeiro, M., 2003. A comparative analysis of clustering algorithms applied to load profiling. In: *Lecture Notes in Artificial Intelligence* (Subseries of Lecture Notes in Computer Science).
- Wang, Y., Chen, Q., Kang, C. and Xia, Q., 2016. Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications. *IEEE Transactions on Smart Grid.*
- Yildiz, B., Bilbao, J.I., Dore, J. and Sproul, A.B., 2017. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Applied Energy.*
- Zhou, K., Yang, S. and Shao, Z., 2017. Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study. *Journal of Cleaner Production.*