Solution to Handle Missing Data in Create Typical Load Graph

Le Hong Lam\(^1\), Nguyen Huu Hieu\(^1\), Ngo Minh Khoa\(^2\), Nguyen Thanh Minh\(^3\) and Tran Tan Vinh\(^1\)

\(^1\) Falcuty of Electrical Engineering-University of Science and Technology-The University of Danang
\(^2\) Quy Nhon University
\(^3\) Thua Thien Hue Power Company
E-mail: lhlam@dut.udn.vn, nhhieu@dut.udn.vn, ngominhkhoa@qnu.edu.vn, minhnt3@cpc.vn and ttvinh@dut.udn.vn

Abstract. Today, the development of the demand is very fast, so the management of demand and the planning of purchasing power is a challenge for Electricity companies. Moreover, the continuous data collection is very difficult especially for Vietnam since the types of measuring equipment are too old or not enough budget to mass equip. Therefore, the paper analyses data processing methods to develop a tool to build typical load graphs for distribution grids corresponding to different load components in order to optimize the distribution grid. The calculation tool is integrated with the proposed algorithm and built-in Matlab and applied to the actual power grid in Thua Thien Hue province for various load components.

1. Introduction
The transmission network includes medium voltage grid (6, 10, 22 kV) and low voltage grid (220/380 kV) which is the last step in power system in order to transmit energy to load. Following [1], The Vietnam Electricity set up a target to reduce energy losses more than 8% in 2015 to 6.5% in 2020. According to the calculation of EVN, from now to 2020, it is necessary to invest 7.9 billion USD/year for Vietnam Power System, in which 25% the investment is used for reducing energy losses purpose. Until 2020, distribution network has been planned to build more than 120.000 km medium voltage line, about 85.000 MVA of distributed substations and about 93.000 km low voltage line. With a rapid expansion, transmission network will face to a serious issue of power losses. Moreover, according to the plan, in July of 2019, "the competitive retail market" will be debuted to operate in early 2020 [2]. Therefore, a research of creating typical load is always a critical issue of power companies as well as in the development of electricity market and some related problems. Therefore, this is an urgent scheme which requires the high accuracy and efficiency from methodologies and applications.

However, to effectively reduce power losses, it is necessary to obtain an accuracy typical load graph. In fact, the selection of typical load graph is processed by two methods: (i) selecting the load graph of a random day in a year or (ii) using the maximum load graph. Selecting a random load graph to become the typical load graph leading to the error in compensating reactive power such as: wrong capacity, wrong location, continuously closing or opening equipment will reduce the life cycle of equipment. Moreover, continuously gathering data is very difficult due to updating every 30 minutes, thus the data can not perfectly 100% and the missing data occurs frequently. Especially, fixing this issue in a short time is very difficult in Vietnam. According to Circular 19 of Ministry of Industry and Trade [3]:

---

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Published under licence by IOP Publishing Ltd
Art. 15: determine measurement error of load pattern and correction methods, for example, a value of 0, notification status, comparison test, check of load factor. Example: value is zero, notification status, comparison check, check load factor;

Art. 16: correction and estimation of measurement data for electric load samples, for example: correction and estimation of measurement data for electric load samples, linear interpolation, analogy dates, automatic estimation;

Art. 18: method of creating load graph, for example: bottom-up method and top-down method;

Art. 21: develop typical load graph by working days (month, year) or holidays.

Currently, the methods of creating typical load graphs being used in the world are based on the Clustering data algorithm [4], also known as Machine Learning method, divided into two groups: (i) unattended linear clustering algorithm and (ii) unattended linear clustering algorithm. Typically, unattended linear clustering algorithms are often chosen because of their simplicity and ease of implementation, which can handle quite large datasets including K-Means [5] and Fuzzy K-Means [6] are two typical methods. Currently, the K-Means method has been proposed for application in reducing load losses [7].

In this paper, the authors summarize the data missing errors that are being encountered in the electricity and propose a method based on the circular of the Ministry of Industry and Trade [3]. The two algorithms K-Means and Fuzzy K-Means commonly used in the creation of load graphs are also compared each other in this paper. Finally, a tool was developed to build a typical load graph by day and season using both methods for Thua Thien Hue Power Company. The tool was developed to take into account the lack of data that electricity companies are facing and thus highly applicable in practice.

2. Creating typical load graph

2.1. Use clustering methods to build a typical load graph

2.1.1. Introduction to data clustering method. The data clustering problem is a major branch of the unattended field of study (see Fig. 1), where the data described in the problem are unlabeled. In this case, the algorithm will try to cluster data into groups with similar characteristics, but at the same time the characteristics between those groups must be as different as possible. The number of data clusters can be determined by experience or can be automatically determined by the algorithm.

![Figure 1. Schema method of clustering data](image)

Application of data clustering: Applied in many fields such as: Business, Biological, Library, Insurance, Medical, Agriculture;
2.1.2. Cluster classification. Data clustering is divided into two categories in Fig. 2: (i) hard data clustering and (ii) soft data clustering:

- **Hard data clustering** (or clear clustering) is a method of assigning each object to one and only one cluster and clearly defining boundaries between clusters. Some algorithms: K-Means algorithm, K-Medoids algorithm;

- **Soft data clustering** (or fuzzy clustering) is a method that allows each object can belong to one or more data clusters and there is ambiguity or fuzzy boundaries between clusters: Fuzzy C-Means algorithm.

![Figure 2. Schema method of clustering data](image)

Currently, the algorithm that is widely used to study clustering problems is the K-Means algorithm and the Fuzzy K-Means algorithm, so the author chose this algorithm to solve the above problem.

2.2. Compare K-means and Fuzzy K-means

K-means [5] is always faster than Fuzzy K-means [6] in all data sets that contain regular scatter clusters or irregular models. Fuzzy K-means is an algorithm based on repeated fuzzy calculations, so that it finds clustering faster than expected.

In addition, Fuzzy K-Means can provide the best results for overlapping data sets and is relatively better than the K-Means algorithm. Unlike K-Means, data points must belong to a cluster center where data points are assigned membership of each cluster center as a result of data points that can belong to more than one center cluster.

However, K-Means and Fuzzy K-Means are still simple algorithms, the number of k clusters is decided by humans. This leads to inaccurate clustering. Therefore, when using this algorithm people use a number of ways to see if k is the number of matching clusters. A commonly used method is the Elbow-method.

2.3. Data processing

The database provided is extremely large as well as the fact that it is difficult to obtain all the data for many years, due to many objective reasons, the database always contains missing data issue, so data processing is necessary when performing clustering methods. Based on data provided from Thua Thien Hue Electricity, the authors propose three methods in Table 1 based on Article 16 of Circular 19 / Ministry of Industry and Trade [3]:

```
| Methods                  | Advantages                                                                 | Disadvantages                                                                 |
|--------------------------|----------------------------------------------------------------------------|------------------------------------------------------------------------------|
| Delete row and column    | Completely removing data with missing values leads to complete              | Loss of information and data; Performed poorly if the percentage              |
|                          | and highly accurate data structure;                                        | of missing values is high (for example: 30%) compared to all data.            |
|                          | It is better to delete a specific row or column without specific data      |                                                                              |
|                          | information because it does not have a high weight.                       |                                                                              |
| Replace with mean value  | This is a good fix when the size of the dataset is small;                  | Give approximate values with additional variances and deviations;             |
|                          | It can limit data loss resulting in the elimination of rows and columns.  | Poor performance compared to other multi-objective methods.                   |
| Specify by neighbouring  | - Low variance because the data is approximate;                           | - Ineffective when the data is missing too much.                             |
| data                     | - Eliminate data loss by replacing it with adjacent data.                  |                                                                              |

The method chosen for additional use of missing data here is based on Part b of Section 2, Article 16 of the Ministry of Industry and Trade’s Circular 19, 2017, which is “Similarity day: Using similar date data of current week or last week.”

(i) Case 1: The data at P 02 (0:30:00) of Saturday (13/01/2018) is missing.

Solution (see Fig. 3): The program automatically finds the location where the data is missing (13th of January, 2018), then at this missing location is assigned the value of Saturday, 6th of January, 2018 (the value of the previous week). After processing, the value of the additional charge as of P 02 of S13th of January, 2018 is 23.32 kW.

(ii) Case 2: The data at the period P 04 (1:30:00) of Wednesday, 03rd of January, 2018 and the period P 02 (0:30:00) of Friday, 12th of January, 2018 are missing. The solution is presented in Fig. 4.
Figure 4. An example of handling necessary data

- Solution 1: The program automatically finds the location where the data was lost (4th position on 3rd of January, 2018), then the position of missing data is assigned the value of the 4th adjacent to 10th of January, 2018 (value of the following week). If the value of the 4th of 10th of January, 2018 is also lost, the data of the 4th of 3rd of January, 2018 is assigned by the value of the 4th of 17th of January, 2018. After processing, the value of the additional charge as of the period P_04 of Wednesday, 3rd of January, 2018 is 20.4 kW.

- Solution 2: The missing data does not belong to the first 7 days in the data table. The program automatically finds the location where the data was lost (6th place on 12th of January, 2018), then the position of missing data is assigned the value of the 6th that is adjacent to 5th of January, 2018 (value of last week). After processing, the value of the additional charge as of P_02 of Friday 12/01/2018 is 20.78 kW.

3. Application of creating a typical load graph to the distribution grid of Thua Thien Hue power company

3.1. Data of load graph
From the load data provided by Thua Thien Hue Power company, Figure 5 is drawn with the horizontal axis of the time axis and the vertical axis being the power axis.

With the data provided, the paper applies the clustering method as introduced earlier and implementation of the program on Matlab software is displayed as follows.

3.2. Develop typical load graph tool
In order to increase the accuracy of the final results, it is extremely effective to determine the typical load graphs for each type of load, moreover the authors focus on the change the load for each day of the week, thus it is more suitable when creating graphs for the first day of the week, the middle of the week and the weekend, as well as the season including the rainy and sunny seasons. Types of load in the program include: (i) agriculture-forestry-fishery, (ii) Industry-construction, (iii) Trade-services, (iv) Activities, (v) Restaurants -Hotels, and (vi) other activities.
The tool was developed towards building a user interface to address the problem that many people can use this tool of creating typical load graphs in Fig. 6, in order to be convenient for work as well as requiring and its adaptability in many places towards high practical applicability.

**Figure 5.** Load graph for days in 2018

**Figure 6.** The interface of tool
Figure 7. The typical load graph shows the differences between K-means and Fuzzy K-means algorithms.

From Fig. 7, the result shows that the difference between the two methods is small, the peak load of the Fuzzy K-means method is higher than that of K-means. Both methods give fairly quick and accurate results. In the following results, the paper presents the results of clustering of sub-loads by K-means method.

Below are the results of implementing the typical load graph for several load groups with the horizontal axis showing time and vertical axis showing the power consumption (KWh). In particular, the red line represents the peak day load (working days) and the blue line shows the peak day load (days off):

Figure 8. Typical load graph of agriculture, forestry and fishery
In general, the tool created accurately classifies the groups in peak and off working days. Therefore, the red line is always higher than the blue line in Fig. 8, Fig. 9, and Fig. 10.

**Figure 9.** Typical load graph of heavy industry

**Figure 10.** Typical load chart of textile industry group
4. Conclusion
This paper has introduced methods of handling data confidentiality and creating typical load graph for distribution grid based on K-Means and Fuzzy K-Means. The created tool integrated with the proposed method has been successfully built and applied with the actual grid in Thua Thien Hue province and proving the accuracy. The simulation and the calculation results can help design engineers and investors to get an overview of the typical load graph when planning as well as choosing the appropriate plan for the specific load for actual situations. In addition, with the open program, the application of the tool can also be researched and integrated to suit the actual operating characteristics of each distribution grid.

5. References
[1] Vietnam Electricity, "Reducing energy losses: "reason and solution"," 2017.
[2] -, "Junly/2019, evn will submit pilot scheme of the competitive retail electricity market," 2019.
[3] Ministry Industry and Trade, "The circular prescribes the content, method and order of conducting electricity load research," 2017.
[4] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review;" ACM computing surveys (CSUR), vol. 31, no. 3, pp. 264-323, 1999.
[5] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 7, pp. 881-892, 2002.
[6] W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation," Pattern recognition, vol. 40, no. 3, pp. 825-838, 2007.
[7] M. Q. Duong, B. T. M. Le Hong Lam, G. Q. H. Tu, and N. H. Hieu, "A combination of k-mean clustering and elbow technique in mitigating losses of distribution network."

Acknowledgments
The authors would like to thank Thua Thien Hue Power Company for providing data and advice to develop and perfect the typical load graphing tool for the load grid.
This work is funded by the Ministry of Education and Training, Vietnam, code: B2019-DNA-13.