pp. 45:50

# Presenting a Practical Way to Pre-process the Raw Data of Smart Meters and Calculate the Load Duration Curve

Mahdi Emadaleslami[1] , Mahdi Emadaleslami [2],Mahmoud-Reza Haghifam[3]*

[1,2]Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran. hassan.majidi@modares.ac.ir
[3]Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran, haghifam@modares.ac.ir

## Abstract

In recent years, due to developments in the electricity industry, the use of smart meters has increased worldwide. Smart meters allow the collection of large amounts of microdata on power consumption. The data collected by smart meters can be used in various cases. Since the load duration curve is of great importance in the study of power systems in this paper, the purpose is to obtain the load duration curve of consumer groups using raw smart meter data. The data collected by smart meters reflects the behavior of subscribers, so by categorizing this data, the behavior of subscribers can be categorized, and thus the continuous load curve can be calculated. However, due to challenges such as being raw data collected from smart meters, the presence of anomalous data, and the presence of lost data, the data collected by smart meters need to be pre-processed and corrected. This paper presents an approach for pre-processing and modification of raw data received from smart meters and using them to calculate the load duration curve and other uses. First, the pre-processing and modification of smart meter data on two data sets collected from Alborz Power Distribution Company is done based on the presented method; Then these data are clustered based on the k-means clustering algorithm, and finally, load duration curve is obtained for each cluster.

*Keywords:* Smart Meter, Load Duration Curve, K-Means Clustering Algorithm, Pre-processing
Article history: Received 04-July-2021; Revised 04-July-2021; Accepted 12-July-2021.

## 1. Introduction

The Faham project (National Smart Metering Program in IRAN) was defined as a national project in 2010 with the permission of the Economic Council and the Management and Planning Organization. Initially, an initial survey was conducted for 20 million subscribers to make their electricity meters smart. However, due to the limited resources of the Economic Council and the Management Organization, this project, which had started in the first phase by replacing one million meters and making them smarter; Postponed. Smart electricity meters record data about power consumption, voltage, current, and power factor. Smart meters report energy consumption in real time during the day at short intervals. Smart meters, along with the communication network and data management system, play an essential role in power distribution systems by recording Consumption curve and facilitating two-way information flows. Data measured by smart meters require a secure platform for transmission and efficient protocols for reading and storing. defects in each of these sections cause problems in using valuable smart meter data. Therefore, due to challenges such as raw data collected from smart meters, the presence of anomalous data, and the presence of lost data, various methods have been proposed to examine the analysis of smart meter data and modify the collected data.

In power systems, Load duration curves are formed by arranging daily peak loads descendingly to create the cumulative load model. Obtaining the load duration curve is of great importance due to its wide range of applications in power systems. Applications of load duration curve include network reliability calculations, production and transmission development, production planning, loss calculations, and so on. Since the data recorded by smart meters reflects the behavior of subscribers, if this collected data can be categorized, it means that the behavior of subscribers is categorized. Therefore, due to the importance of the load duration curve and the increasing use of smart meters, in this paper, a method to obtain load duration curve using smart meter data is presented.

In the second part of this article, the research background and art of states are examined. In the third part, preprocessing and modification of raw data collected from smart meters for two sample data sets is performed. In the fourth section, using the prepared and modified data, the continuous load curve is obtained. The fifth section also includes a summary of the topics and conclusions.

## 2. The Literature Review

To understand and classify studies in the field of smart meters, it is better to first examine the stakeholders of this research [1]. For retailers, there are at least four businesses involved in analyzing smart meter data to increase competition in the retail market, which include load forecasting, pricing design to attract more customers, providing good customer service, and detecting abnormalities or violations. One of the applications of smart meters for consumers is individual load forecasting, which can help reduce electricity bills in home energy management systems [2] and implement energy exchange between consumers in peer to peer markets (P2P) [3]. DSOs can use smart meter data to identify topology of distribution networks, optimize the efficiency of distribution systems, and manage load shedding. Smart meter data must be collected and analyzed by data service providers, so that retailers and consumers can maximize their profits or minimize their costs.

According to the classification of using smart meter data, three main types of smart meter data analysis included load analysis, load forecasting and consumption management of smart meter data. Load analysis studies are reviewed from load profile and the anomaly detection. Anomalies detection in smart meter data is sumerized to bad data detection and null(or energy theft detection) detection. Data loss or unusual patterns that result from unforeseen events, failures to gather information, or failure to communicate or log in, are all examples of bad data. Probabilistic, statistical, and machine learning methods can be used to detect bad data [4].

The process of load profiling involves comparing power consumption patterns of consumers or load curves. Load profiling is divided into two methods, direct clustering and indirect clustering. Various clustering methods such as K-means, hierarchical clustering and self-organized mapping (SOM) have been implemented directly on smart meter data [7-9].

The second analysis of smart meter data is load forecasting, which is widely used in the electricity industry. During the operation and planning process, energy distribution companies use both short-term and long-term forecasts, while retail electricity providers rely primarily on the predicted consumption of their customers. The value and capability that smart meters provide for load prediction is very high. In a high-cited study, seven methods involving linear regression, artificial neural networks, support vector machines, and other types were discussed [10].

The third analysis of smart meter data is load management, which can be summarized in three tipices: first, providing better and more personalized services to consumers involves improving our understanding of their social-demographic information. The second is to target potential consumers for marketing response programs. The third is related to the implementation of the load response plan, including the price plan for price-based load response and the estimation of the baseline for charge-based load response. In [11], non-negative thin coding is used to extract partial consumption patterns from the main load profile.

## 3. Pre-Processing and Modification of Smart Meter Data

In this section, the purpose of preprocessing is to prepare and modify the raw data collected from smart meters. Any problems with the communication platform and reading and storage protocols may result in anomalous data or lost data, so the raw data of the smart meter needs to be pre-processed and modified first. The data considered in this article are from two contracting companies A and B that have a contract with Alborz Power Distribution Company in the field of smart meters.

### *A) Description of the Received Data*

Smart meter data is received separately from the supporting and contracting companies. The information received from Company A in the form of a CSV file consists of six columns of information, the columns of which include the meter ID or meter body number, obis code, year and month, hour, minute and measured value of the variable, respectively.

The data received by Company B was in two separate Excel files with the names of instantaneous power and hourly consumption, each of which had a set of subscriber information. The instantaneous power file contains three columns with the titles of body number, instantaneous power and time and date, which were distributed in six Excel pages. The hourly consumption file consists of four columns: body number, date, time and amount of power consumption, which are in twelve pages, each page containing one month's information; Has been broadcast.

### *B) Data Extraction and Preparation for Preprocessing*

At this stage, the data need to be extracted and synchronized so that the same algorithms for companies A and B can be implemented in the continuation of the work, ie the preprocessing stage. Therefore, according to the needs of the problem, the three columns, the meter ID, abbreviated idc, time and date, abbreviated DAT and value are selected for unification. In the end, the goal is to summarize the information of companies A and B into a table containing three columns: idc, date, and value.

Since Company A data are in the 15-minute interval, according to the standard definition, problem solving needs to be expressed in the hour interval. So according to Equation 1, the mean is used.

$$V_h = \frac{V_{q1} + V_{q2} + V_{q3} + V_{q4}}{4} \qquad (1)$$

In equation 1, $V_h$ is the value per hour and $V_{q1}$ to $V_{q4}$ is the value from zero to forty-five minutes.

After standardizing the clock, the values of hour, year and month should be merged according to the proposed standard and written in the date column.

In Company B data, to achieve the standard, the instantaneous power file only needs to change the name and location of the columns, but the hourly consumption file data, in addition to the need to change the name and location of the columns, need to merge the clock and date columns to obtain have a date. After performing the mentioned steps, the data of the two files can be combined together and the duplicate information can be deleted and a comprehensive file for company B data can be obtained.

### *C) Checking the format of the collected data*

When saving and outputting data, some data may not be stored and exported as defined. Therefore, the column information for both companies should be checked.

The idc column must be a string with 14 character length. Any data in the idc column other than the format mentioned is considered suspicious and anomalous. In the study of the collected data of Company A, no abnormalities were observed, but in the same study for Company B, 106 irrelevant ID were observed, and finally the information related to these 106 meters was removed from the collected data of Company B.

The information received from the company is from the first of Mehr 98 to the end of Shahrivar 99. According to the standard followed in this article, the date and time are expressed from left to right, year, month, day, hour, minute and second, respectively, year, month and day with the character "-" and hour, minute and second with the character. ":" Are separated. In general, time and date are separated by the character "" (Space). " 1398-07-01 00:00:00 "is an example of time and date. Examining the data collected by companies A and B, no data was found to violate the mentioned format.

### *D) Deleting Incorrigible Data*

Missing values or nan is usually one of the main problems in the field of data collection and there are different solutions to overcome this problem. One way is to delete information about that column. This method is generally used when the number of data is large and deleting some of them does not effect with the analysis process or when that column is insignificant for all data. Another way to deal with missing values is to replace them with real data, which is usually a fixed value or the average of the recorded values. This method is commonly used in dealing with numerical data.

Determining which meters have the ability to modify and replace the nan value is very difficult and depends on the amount of error acceptable for recovering lost data. Given the purpose of this paper, which is to classify loads and obtain load duration curve, it is assumed that meters with less than 2000 hours of lost value can be modified. Accordingly, if the number of lost values for each meter is more than 2000 hours, the information of that meter will be deleted from the continuation of calculations. The result of this study is the remove of 916 meter data from 1599 meter data collected for Company A and the remove of 725 smart meter data from 1680 meter data for Company B.

### *E) Correction of anomaly values*

Anomalous or abnormal values refer to the amount of data recorded by a meter, which is a large deviation from the average of the data, which indicates the normal state of the data; has it. Abnormal values in the process of normalizing or prioning data can cause problems. Therefore, in this paper, these abnormal values are identified and replaced according to Equation 2 with the mean value plus variance.

$$un_{ij} = \mu_i + \sigma_i \qquad (2)$$

In relation 2, the values of $un_{ij}$, $\mu_i$ and $\sigma_i$ are the anomalous values of meter i at time j, the one-year average of meter i and the one-year variance of meter i, respectively.

The algorithm for finding and replacing anomalous values is shown in Figure 1.

### *F) Checking of the Daylight Saving Time Effect*

The effect of changing daylight saving time (DST) is observed in two time intervals in the load curve. One at the time of forward clock (DST start) which causes the amount of consumption to not be recorded at that time and the other at the time of backward clock ( DST end) which causes the amount of consumption to be repeated twice an hour. The best way to counteract the impact of DST is to shift the time data at the start of the DST to avoid missing values and removing duplicate values at the end of the DST. But as a precaution, it is better to check the hours around the change. By examining the effect of DST on the data of companies A and B for company A at 23 and 1 o'clock on 30 and 31 Shahrivar 1398, the values of these hours were set with the values of 0 o'clock on 31 Shahrivar 1398. For Company B, the zero hour values on the first day of Farvardin 1399 have been lost, which was replaced with the 1 hour values on the same day.

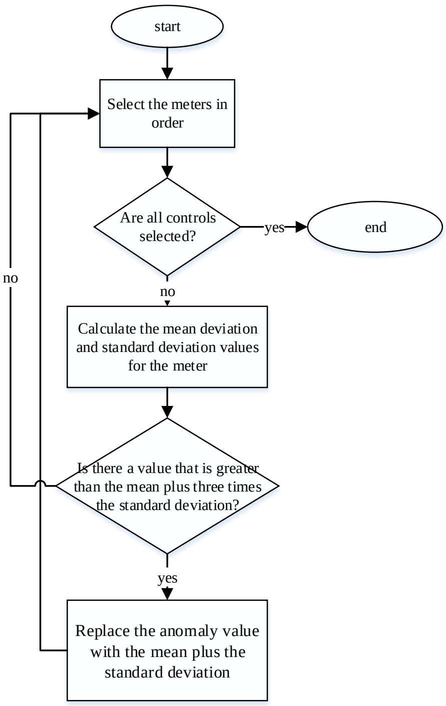
start
Select the meters in order
Are all controls selected?
yes
end
no
no
Calculate the mean deviation and standard deviation values for the meter
Is there a value that is greater than the mean plus three times the standard deviation?
yes
Replace the anomaly value with the mean plus the standard deviation

Fig. 1. Flowchart of correcting anomalous values

### G) *Correcting missing values*

As mentioned before, some of the power in the meter may not be measured or the measured value may not be fully transmitted to the data storage system during transmission due to interruptions and problems with the communication platform. The method of correcting this lost data, similar measurements and using the group average when the lost data is lost instead of the data. In general, the algorithm for correcting lost values consists of two main parts: classifying and quantifying lost values. The algorithm used for clustering in this section is the k-means algorithm. One of the most important inputs of this algorithm is the number of clusters. By increasing the number of clusters, it can be said that the replaced value of the lost value is more reliable, but as the number of clusters increases, the number of members of the category decreases, which may cause the center of the category to have a lost value. Therefore, in selecting the number of clusters, a balance must be struck between the accuracy of the algorithm and the efficiency of the algorithm. According to the purpose and considerations of this paper, the number of clusters for correcting the lost values is considered 5. The flowchart for correcting the lost values by the k-means algorithm is shown in Figure 2.

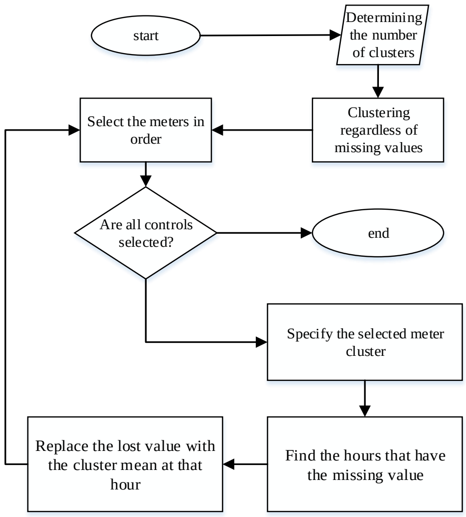
start
Determining the number of clusters
Clustering regardless of missing values
Select the meters in order
Are all controls selected?
end
Specify the selected meter cluster
Find the hours that have the missing value
Replace the lost value with the cluster mean at that hour

Fig. 2. flowchart for correcting missing values.

## 4. Clustering and calculation of load duration curve

The raw data collected by the contractors in the previous stages were prepared and corrected. In this section, these prepared and modified data are used to obtain a continuous load curve for subscriber groups.

### A) *Calculate the optimal number of clusters*

In general, the existing criteria for evaluating clustering and finding the optimal number of clusters can be divided into external and internal categories. External evaluation criteria compare clusters to predefined classes for data and require background knowledge. Internal methods do not require background knowledge and use the statistical knowledge contained in data and clusters to obtain the optimal number of clusters. In this paper, the internal silhouette method is used to calculate the optimal number of clusters [6].

The silhouette criterion for the number of clusters 2 to 9 is applied to the modified data of both companies A and B. Figure 3 shows the profile metrics for modified B company data.

Figure 3 shows the horizontal axis of the number of clusters for the k-means clustering algorithm, and the vertical axis is the mean value of the profile. According to Figure 3 and the definition of the profile criterion, the number of batches 5 is suitable for clustering the modified data of Company B.

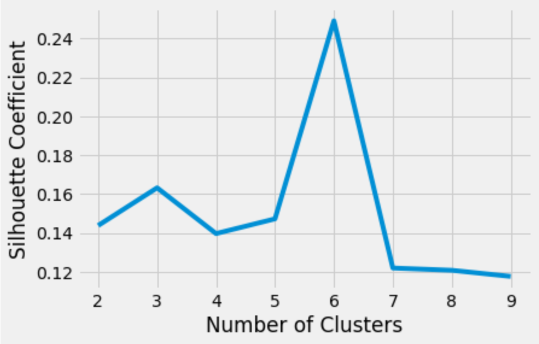

Fig. 3. silhouette criteria on modified B company data.

Similarly, with the implementation of the silhouette method for Company A, it can be seen that batch number 6 is suitable for clustering modified data.

### *B) Calculation of the load duration curve of the center of the clusters*

The load duration curve has many applications in operation, planning, reliability, and other electrical components. Drawing this curve in a one-year period gives a good view of the load consumption and the percentage of peak and off-peak time.

To obtain the load duration curve of the cluster centers, the clustering is repeated with the optimal cluster number calculated from the previous step. Figure 4 is obtained by clustering the modified data of Company A and plotting a continuous load curve for the category centers. In Figure 4, the horizontal axis represents the percentage of load continuity and the vertical axis represents the amount of load in terms of per unit. The center of the categories zero to five is shown in blue, orange, yellow, green, gray, and purple, respectively. These results for Company B are also shown in Figure 6.

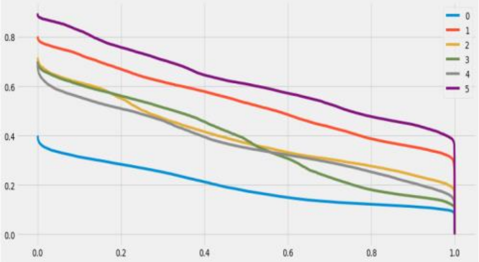

Fig. 4. Load duration curve on modified A company data.

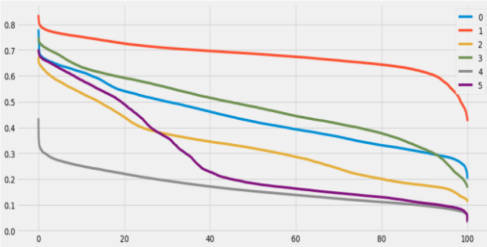

Fig. 5. Load duration curve on modified B company data.

## 5. Conclusion

In this paper, an approach to preprocessing and modifying raw smart meter data for two collected data sets was first performed. It was observed that the received raw data had lost values and anomalous values, which were corrected. Then, these modified data showing the behavior of the subscribers were clustered using the k-means algorithm and the duration curve of the center of the clusters was calculated. The silhouette method was used to find the optimal number of clusters in clustering. The results showed that using this approach, it is possible to obtain a load duration curve from smart meter data.

## References

[1] Y. Wang, Q. Chen, T. Hong and C. Kang, "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges," in IEEE Transactions on Smart Grid, vol. 10, no. 3, pp. 3125-3148, May 2019, doi: 10.1109/TSG.2018.2818167

[2] C. Keerthisinghe, G. Verbič, and A. C. Chapman, "A Fast Technique for Smart Home Management: ADP With Temporal Difference Learning," IEEE Transactions on smart grid, vol. 9, pp. 3291-3303, 2016

[3] A. Pratt, D .Krishnamurthy, M. Ruth, H. Wu, M. Lunacek, and P. Vaynshenk, "Transactive Home Energy Management Systems: The Impact of Their Proliferation on the Electric Grid," IEEE Electrification Magazine, vol. 4, pp. 8-14, 2016

[4] V. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," Artificial Intelligence Review, vol. 22, no. 2, pp. 85–126, 2004.

[5] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity Theft Detection in AMI Using Customers' Consumption Patterns," IEEE Transactions on Smart Grid, vol. 7, pp. 216-226, 2015.

[6] K. Wang, B. Wang, and L. Peng, "CVAP: Validation for Cluster Analyses," Data Science Journal, pp. 0904220071-0904220071, 2009

[7] G. Chicco, "Overview and Performance Assessment of The Clustering Methods for Electrical Load Pattern Grouping," Energy, vol. 42, no. 1, pp. 68–80, 2012

[8] K. Zhou, S. Yang, and C. Shen, "A Review of Electric Load Classification in Smart Grid Environment," Renewable and Sustainable Energy Reviews, vol. 24, pp. 103–110, 2013.

[9] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load Profiling and Its Application to Demand Response: A Review," Tsinghua Science and Technology, vol. 20, no. 2, pp. 117–129, 2015.

[10] R. E. Edwards, J. New, and L. E. Parker, "Predicting Future Hourly Residential Electrical Consumption: A Machine Learning Case Study," Energy and Buildings, vol. 49, pp. 591–603, 2012.

[11] Y. Wang, Q. Chen, C. Kang, Q. Xia, and M. Luo, "Sparse and Redundant Representation-Based Smart Meter Data

Compression and Pattern Extraction," IEEE Transaction. Power Systems, vol. 32, no. 3, pp. 2142–2151, May 2017.