

Analysis Clustering of Electricity Usage Profile Using K-Means Algorithm

¹ Yasirli Amri, ²Amanda Lailatul Fadhilah, ³Fatmawati, ⁴Novi Setiani, ⁵Septia Rani*

^{1,2,3,4,5} Department of Informatics Engineering, Universitas Islam Indonesia, Yogyakarta, Indonesia

E-mail: septia.rani@uii.ac.id

Abstract. Electricity is one of the most important needs for human life in many sectors. Demand for electricity will increase in line with population and economic growth. Adjustment of the amount of electricity production in specified time is important because the cost of storing electricity is expensive. For handling this problem, we need knowledge about the electricity usage pattern of clients. This pattern can be obtained by using clustering techniques. In this paper, clustering is used to obtain the similarity of electricity usage patterns in a specified time. We use K-Means algorithm to employ clustering on the dataset of electricity consumption from 370 clients that collected in a year. Result of this study, we obtained an interesting pattern that there is a big group of clients consume the lowest electric load in spring season, but in another group, the lowest electricity consumption occurred in winter season. From this result, electricity provider can make production planning in specified season based on pattern of electricity usage profile.

Keywords—*electricity usage; clustering; K-Means algorithm*

1. Introduction

Electricity is one of the most important needs of the society. It serves as an economic resource which is the most needed in many activities. Demand for electricity will increase in line with population and investment growth. Electricity usage is an important factor in many sectors, such as in household sector, industrial sector, and also government sector.

Nowadays, the development of information technology has shown how big the role of data that we encounter every day. Existing data will have no mean if it is not treated with the correct methods. Treating the existing data with the correct methods will produce information that can support the work of human beings in all fields. It including with electricity data that became the topic of this paper. By using data mining, which is a study that can process a set of data into valuable information, we can determine the largest or smallest electricity consumption at certain seasons by calculating the used of power (kWh meters).

One of the tasks in data mining that can be used to determine the electricity usage profile by power calculation is clustering. We process the data of electricity usage consist of data power in kWh meter by grouping (clustering) client data using the K-Means algorithm. The intention of this grouping is to analyze the patterns of similarity about electricity consumption made by clients.

The purpose of analysis electricity consumption data is to help determining in which season the electricity at most and the least used. This information can help identifying patterns of clients. In the



future, clients are expected to be aware of the amount of electricity power that they used. In addition, the analysis in this paper can give information to the provider of electricity in order to adjust the production of energy to the actual needs in the field. Adjustment of the amount of electricity production is important because storing electricity is expensive. Thus the amount of production must be adapted to the needs of consumption.

* To whom any correspondence should be addressed.

The rest of the paper is organized as follows. Section 2 describes the literature review. Section 3 describes the experiment. Section 4 describes the results and discussion. Finally, the conclusions of this work are described in section 5.

2. Literature Review

2.1. *K-Means Algorithm*

K-Means is one of the algorithm used for clustering which will split the data into several groups. *K-Means* algorithm is one of the method of data non-hierarchical clustering that can group the data into several clusters based on the similarity of the data. This mechanism enables data which have the same characteristics are grouped into one cluster and data that have different characteristics are grouped in other clusters.

To determine the cluster label of any data, calculated the distance between the data with each cluster centre. There are several ways that can be used to perform distance calculations, such as Euclidean distance, Manhattan distance, and Chebichey distance.

The *K-Means* method aims to minimize the sum of squared distances between all points and the cluster centre. This procedure consists of the following steps, as described in [1].

- Step 1: Select k out of the given n patterns as the initial cluster centres. Assign each of the remaining $n-k$ patterns to one of the k clusters; a pattern is assigned to its closest centre/cluster.
- Step 2: Compute the cluster centres based on the current assignment of patterns.
- Step 3: Assign each of the n patterns to its closest centre/cluster.
- Step 4: If there is no change in the assignment of patterns to clusters during two successive iterations, then stop; else, go to Step 2.

It is obvious in this description that the final clustering will depend on the initial cluster centres chosen and on the value of K . The latter is of the most concern since this requires some prior knowledge of the number of clusters present in the data, which, in practice, is highly unlikely.

K-Means algorithm was chosen because it is simple, easy to implement, widely used in many fields, and the most important is that it has the ability to cluster big dataset. In addition, *K-Means* algorithm is not affected to the order of objects. When using *K-Means* to cluster numeric data, it will be easier to find outlier and noise. The advantage of *K-Means* algorithm is in accordance with the characteristics of electricity data, in which the data is numerical and the size of dataset is quite big.

2.2. *Outlier Detection*

Outliers are data objects with characteristics that are considered to be very different from other data objects within the dataset. Outlier appeared in the form of extreme value to a single variable or combination of variables [2]. Data cleaning with removing outliers is very important. Including outliers in data mining algorithm is sometime will produce inaccurate result because the damaging characteristics of the data processed. Performing outlier detection and remove outliers will make the data more accurate, thus the data mining result will be more accurate.

2.3. *Data Mining on Electricity Consumption*

In the previous study, [3] addressing the segmentation of electricity clients according to social class, contracted power, family size and type of tariff for designing the energy efficiency solutions. [4]

employ support vector machine optimization to predict power load based on time series matrix. This prediction aims to be more easily regulate the use of electricity. In another study, it was mentioned that the genetic algorithm in the improvement of fuzzy clustering is used to predict changes in demand for electricity consumption for seasonal and monthly, especially in developing regions such as in China and Iran [5]. [6] proposed methodology using pattern recognition methodologies to recognize habitual electricity consumption behaviour given the intrinsic characteristics of the family. This approach could be useful to improve small scale forecast, and as a mechanism to enable the provision of tailor-made information to the families.

3. Experiment

3.1. Dataset

We use ElectricityLoadDiagrams2011-2014 dataset, taken from UCI Machine Learning Repository [7]. This dataset contains electricity consumption of 370 instances/clients, collected from 2011 through 2014. There were 140256 attributes and each attribute stating the value of electricity consumption in kW per 15 minutes from 2011-2014 (attribute type is numerical). The example of dataset can be seen in figure 1.

	X	MT_001	MT_002	MT_003	MT_004	MT_005	MT_006	MT_007	MT_008	MT_009	MT_010
36926	2012-01-20 15:30:00	17.7665	24.18208	2.606429	99.5935	48.78049	175.5952	2.826456	259.2593	104.8951	65.5914
36927	2012-01-20 15:45:00	16.49746	24.89331	2.606429	87.39837	42.68293	166.6667	2.826456	259.2593	104.8951	63.8172
36928	2012-01-20 16:00:00	17.7665	24.18208	2.606429	83.33333	43.90244	163.6905	2.261164	265.9933	99.65035	61.29032
36929	2012-01-20 16:15:00	16.49746	24.89331	2.606429	95.52846	42.68293	160.7143	2.826456	262.6263	76.92308	66.66667
36930	2012-01-20 16:30:00	15.22843	25.60455	2.606429	91.46341	39.02439	148.8095	2.826456	235.6902	85.66434	78.49462
36931	2012-01-20 16:45:00	16.49746	25.60455	2.606429	93.49553	39.02439	166.6667	2.261164	249.1582	92.65734	60.21505
36932	2012-01-20 17:00:00	35.53299	24.89331	2.606429	103.6585	39.02439	166.6667	2.261164	259.2593	62.93706	51.6129
36933	2012-01-20 17:15:00	26.64975	25.60455	2.606429	103.6585	39.02439	172.619	2.826456	276.0943	57.69231	50.53763
36934	2012-01-20 17:30:00	16.49746	26.31579	2.606429	115.8537	53.65854	208.3333	3.391747	276.0943	62.93706	74.19355
36935	2012-01-20 17:45:00	32.99492	26.31579	2.606429	115.8537	57.31707	187.5	4.522329	303.0303	59.44056	78.49462
36936	2012-01-20 18:00:00	35.53299	27.73826	2.606429	134.1463	69.5122	247.0238	5.08762	336.7003	73.42657	87.09677
36937	2012-01-20 18:15:00	17.7665	27.73826	2.606429	158.5366	80.4878	288.6905	5.652911	373.7374	90.90909	93.54839
36938	2012-01-20 18:30:00	19.03553	26.31579	2.606429	189.0244	85.36585	312.5	4.522329	390.5724	103.1469	104.3011
36939	2012-01-20 18:45:00	17.7665	26.31579	2.606429	180.8943	91.46341	333.3333	5.08762	427.6094	106.6434	100
36940	2012-01-20 19:00:00	19.03553	26.31579	2.606429	178.8618	92.68293	333.3333	6.218202	427.6094	111.8881	102.1505
36941	2012-01-20 19:15:00	17.7665	29.16074	2.606429	199.187	93.90244	324.4048	5.08762	434.3434	111.8881	106.4516
36942	2012-01-20 19:30:00	19.03553	29.87198	2.606429	225.6058	104.878	327.381	6.218202	491.5825	108.3916	110.7527
36943	2012-01-20 19:45:00	19.03553	31.29445	2.606429	209.3456	109.7561	345.2381	5.652911	474.7475	103.1469	113.9785
36944	2012-01-20 20:00:00	17.7665	27.02703	4.344049	184.9553	103.6585	330.3571	5.652911	430.9764	101.3986	112.9032
36945	2012-01-20 20:15:00	19.03553	32.00569	6.081668	189.0244	100	333.3333	6.218202	424.2424	103.1469	113.9785
36946	2012-01-20 20:30:00	20.30457	30.58321	4.344049	197.1545	104.878	348.2143	7.348785	400.6734	106.6434	93.54839
36947	2012-01-20 20:45:00	19.03553	29.16074	3.475239	195.122	95.12195	330.3571	7.348785	383.3364	94.40559	90.32258
36948	2012-01-20 21:00:00	17.7665	28.4495	3.475239	186.9919	87.80488	330.3571	7.514076	370.3704	83.91608	88.17204
36949	2012-01-20 21:15:00	19.03553	27.73826	3.475239	182.9268	91.46341	315.4762	6.218202	367.0034	82.16783	79.56989
36950	2012-01-20 21:30:00	17.7665	27.02703	3.475239	182.9268	86.58537	315.4762	8.479367	353.3354	80.41958	78.49462
36951	2012-01-20 21:45:00	11.42132	27.73826	5.212858	186.9919	91.46341	324.4048	6.783493	373.7374	87.41259	77.41935

Figure1. Example of dataset.

The original ElectricityLoadDiagrams20112014 dataset can be visualized using matrix D as follows:

$$D = \begin{bmatrix} x_{1,1} & \cdots & x_{1,370} \\ \vdots & \ddots & \vdots \\ x_{140256,1} & \cdots & x_{140256,370} \end{bmatrix}$$

The number of row is 140256 (stating the number of attributes) and the number of column is 370 (stating the number of instances). Data D is transposed so that rows declare instances and columns declare attributes.

$$D' = \begin{bmatrix} x_{1,1} & \cdots & x_{1,140256} \\ \vdots & \ddots & \vdots \\ x_{370,1} & \cdots & x_{370,140256} \end{bmatrix}$$

We perform dimension reduction to the dataset because the number of attribute is too much. Steps of dimension reduction are as follow:

- First, select only certain duration of time. We select the electricity consumption of 370 instances between March 21, 2013 and March 20, 2014.
- From the selected duration of time, we aggregated (summed) the electricity needs for each season. Dataset's information stated that the data is retrieved in Portugal time. Because in Portugal there are four seasons, we perform data aggregation as follows:
 - For each instance, all of the electricity consumption from March 21, 2013 until June 20, 2013 is added to obtain the value of total electricity consumption in spring.
 - For each instance, all of the electricity consumption from June 21, 2013 until September 22, 2013 is added to obtain the value of total electricity consumption in summer.
 - For each instance, all of the electricity consumption from September 23, 2013 until December 20, 2013 is added to obtain the value of total electricity consumption in autumn.
 - For each instance, all of the electricity consumption from December 20, 2013 until March 20, 2014 is added to obtain the value of total electricity consumption in winter.
- Thus the data dimension will be reduced. Each instance will have four attributes, as follows:

$$D_r \text{ (after dimension reduction)} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ \vdots & \vdots & \vdots & \vdots \\ x_{370,1} & x_{370,2} & x_{370,3} & x_{370,4} \end{bmatrix}$$

Description: The first column is the amount of electricity consumption during spring, the second column is the amount of electricity consumption during summer, the third column is the amount of electricity consumption during autumn, while the fourth column is the amount of electricity consumption during winter.

Data D_r will be used for visualization and clustering.

3.2. Tools

In this study we used Rstudio version 3.2.2 [8] with several libraries such as xts package, ggplot, and outliers. We also used Weka version 3.7.4 [9] to perform clustering task.

4. Results and Discussion

4.1. Results

The example of result dataset after the process of dimension reduction can be seen in the figure2.

	row.names	V1	V2	V3	V4
1	MT_001	36017.77	64649.75	98902.28	19529.19
2	MT_002	251569	308256	240322.9	225683.5
3	MT_003	17935.71	16034.75	14913.12	14728.06
4	MT_004	839878	944004.1	981579.3	1212486
5	MT_005	352457.3	427929.3	460042.7	577119.5
6	MT_006	1503557	1546244	1738848	1951997
7	MT_007	31565.86	157960.4	33873.94	45366.87
8	MT_008	2151296	2434896	2236761	2277626
9	MT_009	381304.2	447903.8	516484.3	496896.9
10	MT_010	548790.3	438402.2	535288.2	508466.7
11	MT_011	293096.9	335868.9	350378.5	421040.2
12	MT_012	0	333385.1	1259196	1513834
13	MT_013	602344.9	714411.3	679792.7	652699.8
14	MT_014	356758.8	336899.2	407737.7	473116.8
15	MT_015	58722.62	176699.4	163892.6	448886.9
16	MT_016	254970.4	313702.2	281698.2	336478.9
17	MT_017	393219.2	454653.2	427071.9	449489.7

Figure 2. The data sample after dimension reduction.

Description:

MT_001, MT_002, etc. are the identification number for clients. Column V1 states the total amount of electricity consumption during spring. Column V2 states the total amount of electricity consumption during summer. Column V3 states the total amount of electricity consumption during autumn. Column V4 states the total amount of electricity consumption during winter.

4.2. Data Visualization

The visualization of electricity consumption from 370 clients in each season can be seen in the figure 3. The x-axis is the clients and y-axis is the electricity consumption in kW. Figure 3 shows that the majority of clients use electricity no more than 50 million kW in each season.

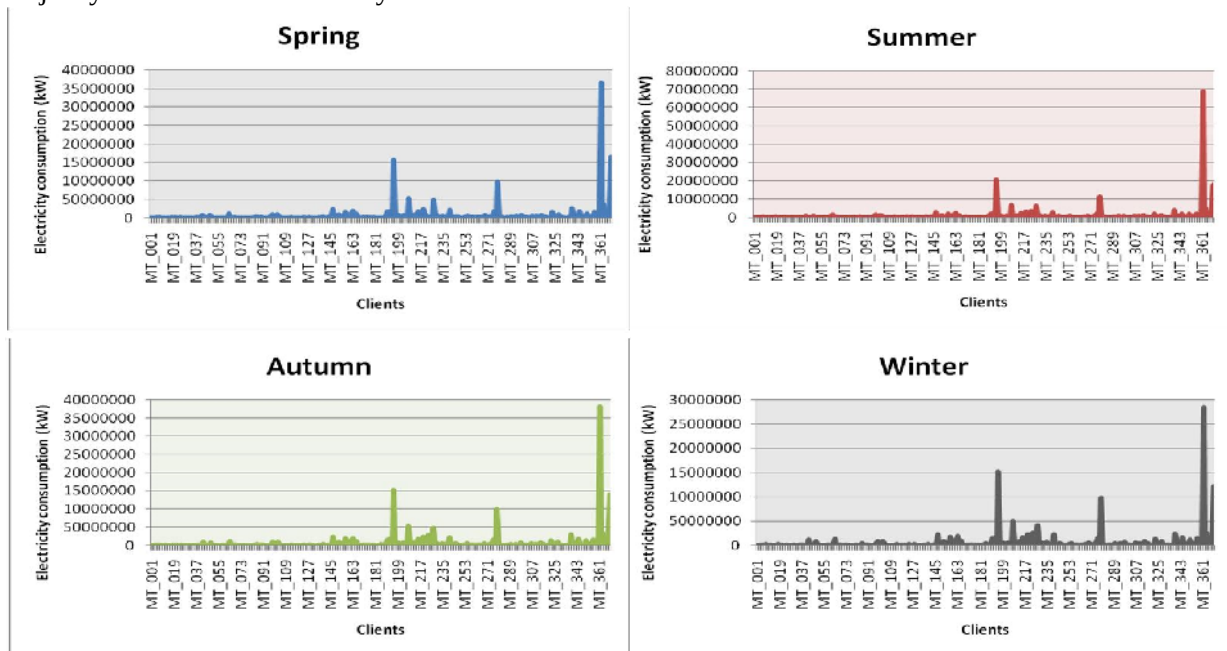


Figure 3. Visualization of electricity consumption in each season.

The comparison graph of electricity consumption between the four seasons can be seen in the figure 4. The x-axis is the clients and y-axis is the electricity consumption in kW. Figure 4 shows that for the majority of clients, the highest amount of electricity consumption is during the summer.

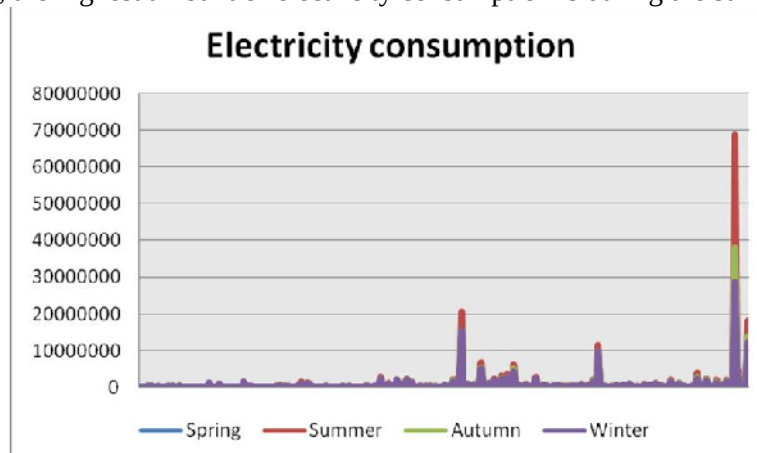


Figure 3. Comparison of electricity consumption between the four seasons.

4.3. K-Means Clustering

For clustering, we use data D_r (as explained in Section 2) as the input. We have four scenarios in the process of clustering, as follows:

- Scenario 1: Using the data D_r without outlier detection and then group the data into 4 clusters.
- Scenario 2: Using the data D_r with outlier detection (remove the outlier) and then group the data into 4 clusters.
- Scenario 3: Using the data D_r without outlier detection and then group the data into 5 clusters.
- Scenario 4: Using the data D_r with outlier detection (remove the outlier) and then group the data into 5 clusters.

For each scenario, we measure the Sum Square Error (SSE) and the number of iteration. SSE describes the value of standard deviation of each cluster to the data centre. The bigger SSE, it means that the degree of data similarity in one cluster is lower. Number of iteration describes the length of clusters formation process.

The clustering result from each scenario can be seen in table 1. From table 1, it can be seen that the largest SSE is obtained when we perform clustering with 4 clusters and outliers are not included (Scenario 2), in which, this scenario also has the most minimal number of iteration, that is 14. The best SSE is obtained when we use 5 clusters without eliminating outliers (Scenario 3).

Table 1. Clustering result.

No	Scenario	SSE	Number of iteration
1	4 clusters without eliminating outliers	0.174	20
2	4 clusters with eliminating outliers	0.752	14
3	5 clusters without eliminating outliers	0.134	21
4	5 clusters with eliminating outliers	0.509	22

The more detail result from Scenario 3 (5 clusters without eliminating outliers) can be seen in table 2. From table 2, it can be seen that the cluster 0 – 3 had the biggest average use of electricity in summer and the smallest average use of electricity in winter. There is only one distinct cluster, cluster number 4, which had the smallest average use of electricity in spring.

Table 2. Clustering result scenario 3.

Cluster Number	Number of cluster member	$\bar{v1}$	$\bar{v2}$	$\bar{v3}$	$\bar{v4}$
0	3	139.584.280,44	165.888.048,23	129.645.729,66	123.076.512,72
1	17	22.810.155,19	31.293.383,26	24.732.531,25	22.188.694,04
2	47	7.148.526,50	8.987.654,70	7.206.537,18	6.817.580,39
3	1	364.726.900,00	687.392.800,00	380.801.800,00	284.465.900,00
4	302	1.008.178,06	1.245.319,00	1.059.189,86	1.061.011,85

Table description:

- $\bar{v1}$: the average use of electricity in spring (kW)
 $\bar{v2}$: the average use of electricity in summer (kW)
 $\bar{v3}$: the average use of electricity in autumn (kW)
 $\bar{v4}$: the average use of electricity in winter (kW)
 : the largest average use of electricity
 : the smallest average use of electricity

4.4. Discussion

Based on the result from the third scenario, we got five clusters describing below:

1. Group of clients with very high electricity consumption (Cluster 3)
There is 1 client with the highest consumption occurred on the summer season (687.392.800,00 kW) and the lowest consumption occurred on the winter season (284.465.900,00 kW).
2. Group of clients with high electricity consumption (Cluster 0)
There are 3 clients with the highest average of electricity consumption, 165.888.048,23 kW, occurred on the summer season and the lowest average of electricity consumption, 123.076.512,72 kW, occurred on the winter season.
3. Group of clients with medium electricity consumption (Cluster 1)
There are 7 clients with the highest average of electricity consumption, 31.293.383,26 kW, occurred on the summer season and the lowest average of electricity consumption, 22.188.694,04kW, occurred on the winter season.
4. Group of clients with low electricity consumption (Cluster 2)
There are 47 clients with the highest average of electricity consumption, 8.987.654,70 kW, occurred on the summer season and the lowest average of electricity consumption, 6.817.580,39kW, occurred on the winter season.
5. Group of clients with very low electricity consumption (Cluster 4)
There are 302 clients with the highest average of electricity consumption, 1.245.319,00 kW, occurred on the summer season and the lowest average of electricity consumption, 1.008.178,06 kW, occurred on the spring season.

This result shows that the most dominant group is the group of clients with very low electricity consumption (Cluster 4). The average electricity consumption from this group is very contrast to another client groups (Cluster 0, 1, 2, and 3). Another difference between cluster 4 and another cluster is they consume the lowest electricity on the spring season but another client groups consume the lowest electricity on the winter season.

5. Conclusions

Based on the analysis above, we obtained an interesting fact that is the biggest electricity consumption actually occurs in the summer. For some clients, the smallest electricity consumption is during winter. But for the majority of clients, the smallest electricity consumption occurs in the spring. This may be happened due to the differences in geographic location, the differences in economic level, and the differences in the use of electricity as the needs of industrial, household, office. The majority of clients are clustered in the group of the lowest electricity consumption. This group has very different electricity usage characteristic with other groups. There is an additional fact that exist a client who consume electricity almost half of the total electricity consumption.

References

- [1] Murty MN and Devi VS 2011 Pattern recognition, An algorithmic approach (London: Springer)
- [2] Hodge V H and Austin Jim 2004 A survey of outlier detection methodologies *J. Artificial Intelligence Review* **22** 85-126
- [3] Pombeiro H Pina A and Silva C Analyzing residential electricity consumption patterns based on consumer's segmentation *Proc. Int. 1st. Workshop on Information Technology for Energy Applications (Lisbon, Portugal, 6-7 September 2012)* vol 923 ed P Carreira and V Amaral (Portugal: CEUR-WS) pp 29-38
- [4] Dong-xiao N, Yong-li W and Xiao-yong M 2010 Optimization of support vector machine power load forecasting model based on data mining and Lyapunov exponents *J. Central South University of Technology* **17** 406-12

- [5] Azadeh A, Saberi M, Ghaderi S F, Gitiforouz A and Ebrahimipour V 2008 Improved estimation of electricity demand function by integration of fuzzy system and data mining approach *J. Energy Conversion and Management* **49** 2156-77
- [6] Abreu J M, Pereira F C and Ferrao P 2012 Using pattern recognition to identify habitual behavior in residential electricity consumption *J. Energy and Building* **49** 479–87
- [7] <https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014>
- [8] <https://www.rstudio.com/products/RStudio/>
- [9] <http://www.cs.waikato.ac.nz/~ml/weka/downloading.html>