IOP Conference Series: Materials Science and Engineering

Cluster Analysis of Indonesian Province Based on Household Primary Cooking Fuel Using K-Means

To cite this article: S N Huda 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **185** 012016

Cluster Analysis of Indonesian Province Based on Household Primary Cooking Fuel Using K-Means

S N Huda[1]

[1]Dept. of Informatics, Universitas Islam Indonesia, Yogyakarta, Indonesia
E-mail: [1]sheila@uii.ac.id

Abstract. Each household definitely provides installations for cooking. Kerosene, which is refined from petroleum products once dominated types of primary fuel for cooking in Indonesia, whereas kerosene has an expensive cost and small efficiency. Other household use LPG as their primary cooking fuel. However, LPG supply is also limited. In addition, with a very diverse environments and cultures in Indonesia led to diversity of the installation type of cooking, such as wood-burning stove brazier. The government is also promoting alternative fuels, such as charcoal briquettes, and fuel from biomass. The use of other fuels is part of the diversification of energy that is expected to reduce community dependence on petroleum-based fuels. The use of various fuels in cooking that vary from one region to another reflects the distribution of fuel basic use by household. By knowing the characteristics of each province, the government can take appropriate policies to each province according each character. Therefore, it would be very good if there exist a cluster analysis of all provinces in Indonesia based on the type of primary cooking fuel in household. Cluster analysis is done using K-Means method with K ranging from 2-5. Cluster results are validated using Silhouette Coefficient (SC). The results show that the highest SC achieved from K = 2 with SC value 0.39135818388151. Two clusters reflect provinces in Indonesia, one is a cluster of more traditional provinces and the other is a cluster of more modern provinces. The cluster results are then shown in a map using Google Map API.

1. Introduction

Human can not live without energy, in every activity from transportation, work, to even the most basic and needed activity, eating (cooking), we need energy. Each household definitely provides installations for cooking. Kerosene, which is refined from petroleum products once dominated types of primary fuel for cooking in Indonesia. Whereas the processing of crude oil into kerosene has an expensive cost and small efficiency, in addition to the product can be processed further to produce Avtur (Aviation Turbine) whose efficiency is higher. Indonesian government had rolled out a program to replace kerosene which still had high consumption rate, into LPG (Liquid Petroleum Gas) for cooking.

LPG does have better efficiency compared to kerosene, LPG consumption has a calorific value of 11254.61 Kcal / Kg compared to kerosene amounted to 10478.95 Kcal / Kg, with one liter of Kerosene LPG is equivalent to 0.57 Kg LPG. In addition, carbon emissions from LPG is smaller than the carbon emissions from kerosene. However, LPG supply is also limited. This can be seen from the results of LPG production, in 2000 the production of LPG in Indonesia still met domestic needs, so it did not need to be imported, whereas in 2010 Indonesia needed to import up to 1 million tons of LPG [1]. In the other hand, the utilization of natural gas by pipeline to the household does not require the cost for packaging tubes, but the installation of gas lines currently only covers a small part of settlements in several cities.

In addition, with a very diverse environments and cultures in Indonesia led to diversity of the installation type of cooking. We are still relatively easy to find a wood-burning stove brazier in Central Java. But if we make a survey in Jakarta, stove brazier is hard to be found.

The government is also promoting alternative fuels, such as charcoal briquettes, and fuel from biomass [2]. The use of other fuels is part of the diversification of energy that is expected to reduce community dependence on petroleum-based fuels.

The use of various fuels in cooking that vary from one region to another reflects the distribution of fuel basic use by household. There are some provinces in Indonesia which rely heavily on the use of LPG as the primary cooking fuel. Several other provinces use wood as the dominant primary fuel to cook. By knowing the characteristics of each province, the government can take appropriate policies to each province according each character. Therefore, it would be very good if there exist a cluster analysis of all provinces in Indonesia based on the type of primary cooking fuel in household.

For each province, the main fuel for cooking are calculated from the percentage of family using electricity, LPG, kerosene, charcoal / briquettes, wood, and other fuels. Other fuels here include households who do not cook.

Cluster analysis is done using K-Means method, where K-Means is well known for its simplicity. K-Means method has been used in many cluster analysis, such as in [3], k-means method is used to cluster soil character. In the same field, soil data clustering was also performed by [4] used two different methods, namely K-Means and Fuzzy C-Means uses soil data from Montenegro. The results of the K-Means soil clustering is visually displayed using the Google Map API.

In this paper, K-Means method is used to cluster Indonesian province based on the primary cooking fuel in household. Silhouette Coefficient is used to validate the result.

2. Literature Review

2.1. K-Means Clustering

Cluster analysis is a technique in finding the group in the data set in order for the data residing on one group has a close resemblance, and have a clear distinction with the other groups [5].

K-Means is a basic algorithm that is widely used for the data clustering [6]. Good clustering method will produce high quality clusters with high intra-class similarity (cohesive in one cluster) and low inter-class similarity (a clear difference between groups). The quality of a clustering method depends on the similarity measurement used in the method, its implementation, and the ability to find some hidden patterns. Cluster found using the K-Means algorithm has a convex shape and each cluster is represented by a center / centroid [3].

K-Means is a method for partitioning a set of data to treat the data as an object that has a location and distance to other data. K-Means will partition the object into k classes which are mutually exclusive, so that every object in each cluster is as close as possible to one another, and as far as possible with objects in other clusters. Each cluster is characterized by centroid, or midpoint. K-Means clustering algorithm is one of the most widely used, was first described by [7]. K-Means designed to cluster numerical data where each cluster has a central point, which is called the average. K-Means clustering algorithm is classified as partition or nonhierarchical method [8].

In this algorithm, the number of clusters k is assumed to be static. Then each data is allocated into the nearest cluster. After all data is allocated into k cluster, recalculate the midpoint of each cluster and reallocate each point to the cluster closest to the new midpoint repeatedly until the midpoint of each cluster does not have significant differences or any membership of each data is not changing [9].

To calculate the distance, we use Euclidean Distance as follows.

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - c_{kj})^2} \qquad (1)$$

where:
d_{ik} = distance of i-th data to center of cluster k
m = number of attribute
x_i = i-th data
c_k = center of cluster k

2.2. Silhouette Coefficient

Silhouette coefficient is a method of interpretation for the cluster validation on data objects. This technique provides a brief graphical representation of how well each object is located in its cluster. Silhouette coefficient was first developed by [10].

A silhouette coefficient value of an object, egAj, is in the range between -1 to 1. The closer the object silhouette Aj to 1, the higher the degree of ownership of the object Aj in the cluster. While the value close to -1 (negative value) indicates that an object is likely to be on the wrong cluster, because the object similarity with other objects in the same cluster is low.

Silhouette coefficient test is a popular method that combines in a cluster cohesion and separation between clusters. For each data point, the calculation involves three steps.

1. For each object i in the cluster, calculate the average distance from the object to the entire objects in its own cluster. Call the average distance between the object i to all members of the cluster as ai.
2. For each object i and other cluster that is not cluster where i was in, calculate the average distance from the object to the entire objects in other cluster closest to cluster where i was (cluster neighbors). Call average distance to all objects in the neighboring clusters as bi.
3. For each object i, the silhouette coefficient obtained from:

$$(b_i - a_i)/\max(a_i, b_i) \tag{2}$$

The value of silhouette coefficient resides between -1 and 1. The negative value is not desirable because these values correlated with the case where the ai, as the value of the average distance to other point in the cluster, larger than the value of bi, the smallest average value to points on the other cluster. Silhouette coefficient value is expected to be positive (ai<bi) and ai is expected to be as close as possible to 0, where the silhouette coefficient reaches its maximum value 1 when ai = 0.

The average value of the Silhouette coefficient of the entire cluster will reflect the quality of the clusters. Based on [5] the average value of the Silhouette Coefficient (SC) is interpreted as a strong or weak classification as shown in Table 1.

Table 1. SC Interpretation

SC Value	Interpretation
0.71-1.00	Strong Classification
0.51-0.70	Good Classification
0.26-0.50	Weak Classification
0-0.25	Bad Classification

3. Dataset

The dataset used is the primary cooking fuel percentage of households by province in Indonesia in 2014 which are available from BPS as in Table 2. In Indonesia, we use some type of cooking fuel, such as electricity, LPG, kerosene, charcoal / briquettes, wood, and other fuels.

The clustering system is build using PHP. The dataset has to be normalized first. The data normalization is based on:

$$N_{ij} = (\frac{x_{ij} - \min(x_{1j},\ldots,x_{nj})}{\max(x_{1j},\ldots,x_{nj}) - \min(x_{1j},\ldots,x_{nj})} \tag{3}$$

where
N_{ij} = Normalized value of i-th data in parameter j.
x_{ij} = value of i-th data in parameter j.

K on K-Means will be calculated from K = 2 to K = 5, then the silhouette coefficient test will be conducted. The system will display the most optimal silhouette coefficient from each K. Clustering output will be presented in the form of a cluster map using Google Map API.

Each cluster will be given a different marker, and each marker will be placed in the middle of their respective provinces. Province coordinate data refers to data from Google Map API. As for the representation marker will use a custom marker.

Table 2. Dataset used in experiment

Province	Percentage of Family Based On Province and Primary Cooking Fuel (2014)					
	Electricity	LPG	Kerosene	Charcoal/ Briquettes	Wood	Other
ACEH	0.11	64.61	3.97	0	28.82	2.49
SUMATERA UTARA	0.59	67.92	6.32	0.01	23.8	1.37
SUMATERA BARAT	0.31	23.47	25.54	0	49.22	1.46
RIAU	0.58	69.54	9.7	3.62	14.56	2
JAMBI	0.41	54.67	8.52	2.89	32.66	0.85
SUMATERA SELATAN	0.57	75.07	1.59	0.87	21.6	0.3
BENGKULU	0.37	62.59	1.77	0	34.09	1.19
LAMPUNG	0.27	53.39	0.43	0.03	45.25	0.62
KEP. BANGKA BELITUNG	0.54	56.64	27.53	0.03	14.76	0.49
KEP. RIAU	0.84	67.54	20.13	0.48	5.45	5.57
DKI JAKARTA	0.95	86.76	2.59	0	0.15	9.55
JAWA BARAT	1.06	76.34	0.46	0.02	19.59	2.53
JAWA TENGAH	0.5	63.45	0.25	0.07	33.47	2.25
DI YOGYAKARTA	0.4	54.19	0.57	0.37	33.93	10.54
JAWA TIMUR	0.5	62.77	0.81	0.03	33.56	2.33
BANTEN	0.79	76.75	0.54	0	19.79	2.12
BALI	0.36	62.62	0.98	0.02	28.85	7.18
NUSA TENGGARA BARAT	0.4	36.68	15.9	0	45.43	1.59
NUSA TENGGARA TIMUR	0.77	0.58	18.66	0	79.6	0.39
KALIMANTAN BARAT	0.63	68.3	1.19	0.07	29.42	0.4
KALIMANTAN TENGAH	0.36	16.65	39.41	0.05	42.72	0.82
KALIMANTAN SELATAN	0.16	39.94	24.34	0.03	34.11	1.42
KALIMANTAN TIMUR	0.58	79.94	8.8	0.1	9.04	1.55
SULAWESI UTARA	1.69	40.92	14.66	0.1	40.52	2.1

SULAWESI TENGAH	0.17	7.2	23.8	5.1	62.83	0.91
SULAWESI SELATAN	0.58	71.26	0.57	0.82	26.29	0.49
SULAWESI TENGGARA	0.22	16.28	25.18	2.36	55.18	0.79
GORONTALO	0.24	44.73	9.63	0.03	43.96	1.41
SULAWESI BARAT	0.33	42.72	1.27	1.41	53.92	0.36
MALUKU	0.1	0.48	48.23	0	50.61	0.58

4. Results

We build web based system to cluster provinces in Indonesia based on primary cooking fuel. It will iteratively calculate each cluster member using K = 2 to K = 5 and conduct the Silhouette Coefficient test from each K. The system will only display clustering result from the optimal Silhouette Coefficient result.From the experiment, the system generates the Silhouette Coefficient output as in Table 3.

Table 3. SC Result

K	SC Value
2	0.39135818388151
3	0.34231572024075
4	0.30726059280905
5	0.26847238402292

The optimal Silhouette Coefficient value is resulted from K = 2. The next optimal Silhouette Coefficient value is from K = 3. Therefore, the system will display a cluster membership table and map.

In K = 2, Cluster 1 consists of 22 provinces with characteristic of medium-high use of electricity, medium-high use of LPG, very low use of kerosene, very low use of charcoal / briquettes, very low-medium use of wood, and low-medium use of other fuel. This cluster depicts more modern provinces in Indonesia. Provinces in cluster 1 are Aceh, North Sumatera, Riau, Jambi, South Sumatera, Bengkulu, Lampung, Kep. Bangka Belitung, Kep. Riau, DKI. Jakarta, West Java, Central Java, DI. Yogyakarta, East Java, Banten, Bali, West Kalimantan, East Kalimantan, North Sulawesi, South Sulawesi, Gorontalo, and West Sulawesi.

While Cluster 2 consists of 11 provinces with characteristic of low use of electricity, low use of LPG, medium-high use of kerosene, low-medium use of charcoal/briquettes, high use of wood, and very low use of other fuel. This cluster depicts more traditional provinces in Indonesia. Provinces in cluster 2 are West Sumatera, NTB, NTT, Central Kalimantan, South Kalimantan, Central Sulawesi, South East Sulawesi, Maluku, North Maluku, West Papua, and Papua.

In K = 3, Cluster 1 consists of 5 provinces with characteristic of medium-high use of electricity, high use of LPG, very low use of kerosene, very low use of charcoal/briquettes, low use of wood, and high use of other fuel (other fuel include not cooking). This cluster depicts modern province in Indonesia. Provinces in cluster 1 are Kep. Riau, DKI. Jakarta, DI. Yogyakarta, Bali, and North Sulawesi.

Cluster 2 consists of 17 provinces with characteristic of low use of electricity, high use of LPG, very low use of kerosene, very low use of charcoal/briquettes, low-medium use of wood, very low use of other fuel. Provinces in cluster 2 are Aceh, North Sumatera, Riau, Jambi, South Sumatera, Bengkulu, Lampung, Kep. Bangka Belitung, West Java, Central Java, East Java, Banten, West Kalimantan, East Kalimantan, South Sulawesi, Gorontalo, and West Sulawesi.

Cluster 3 consists of 11 provinces with characteristic of very low use of electricity, very low use of LPG, medium-high use of kerosene, very low use of charcoal/briquettes, high use of wood, very low use of other fuel. This cluster depicts more traditional province in Indonesia. Provinces in cluster 3 are West Sumatera, NTB, NTT, Central Kalimantan, South Kalimantan, Central Sulawesi, South East Sulawesi, Maluku, North Maluku, West Papua, and Papua.

Figure 1 shows map output from the system using K = 2.

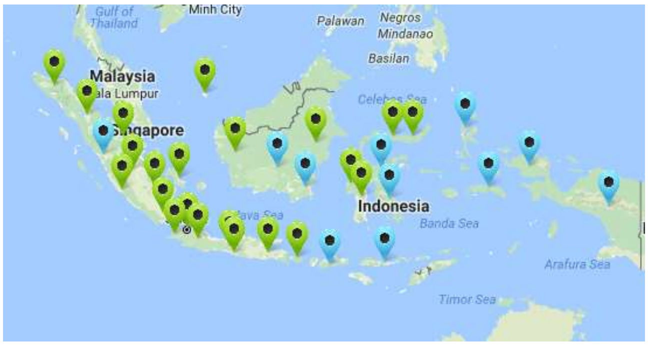

Figure 1. Map Output using K = 2

While figure 2 shows map output using K = 3.

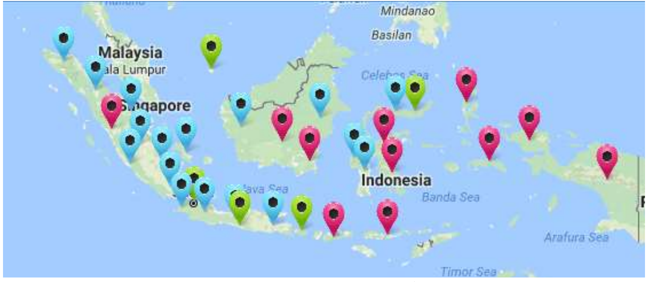

Figure 2. Map Output using K = 3

5. Conclusion

From the above problems it can be concluded that the application of web-based data mining with k-means algorithm can be used to cluster provinces in Indonesia based on primary cooking fuel in household. With this web-based application, it will facilitate government to determine the characteristic of a province in the use of cooking fuel. The application has map output to give easy first glance on the cluster along with the table to give more detail characters. This work can be enhanced using time-series data stream, as each province might have different percentage from each year survey.

Acknowledgment

This work was supported by Universitas Islam Indonesia's internal research funding.

6. References

[1] Kementrian ESDM (Energi dan Sumber Daya Mineral) 2011 Handbook of Energy & Economic Statistics of Indonesia. http://prokum.esdm.go.id/ Publikasi/Handbook%20of%20Energy%20&%20Economic%20Statistics %20of%20Indonesia%20/Handbook %202011.pdf.

[2] Rahmani P, Hartono D M and Kusnoputranto H 2013 Kajian Kelayakan Pemanfaatan Biogas dari Pengolahan Air Limbah untuk Memasak *J. Ilmu Lingkungan* Volume 11 Issue 2: 132-40.

[3] Ashok K D and Kannathasan N 2013 A Study and Characterization of Chemical Properties of

Soil Surface Data Using K-Means Algorithm *Proc. of Pattern Recognition, Informatics and Medical Eng. (PRIME)*

[4] Hot E and Popović-Bugarin V 2015 Soil Data Clustering by Using K-Means and Fuzzy K-Means Algorithm *Proc. of 23rd Telecommunications forum TELFOR 2015*

[5] Kauffman L and Rousseeuw P J 1990 *Finding Group in Data: An Introduction to Cluster Analysis* (New York: Wiley).

[6] Ashok Kumar D, Annie M C L C, and Begum T U S 2012 Computational Time Factor Analysis of K-Means Algorithm on Actual and Transformed Data Clustering *In Proc. of Pattern Recognition, Informatics and Medical Eng. (PRIME),*

[7] Macqueen J 1967 Some Methods for Classification and Analysis of Multivariate Observations *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability* Volume 1 (Berkley C.A.: University of California Press) pp 281-97.

[8] Jain A and Dubes R 1988 *Algorithms for Clustering Data Englewood Cliffs* (NJ: Prentice Hall).

[9] Gan G, Wu J and Ma C 2007 Data Clustering: Theory, Algorithms, and Applications *ASA-SIAM Series on Statistics and Applied Probability* SIAM Philadelphia ASA Alexandria VA

[10] Rousseeuw P J 1987 Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis *J. of Computational and Applied Math.* **20** pp 53-65