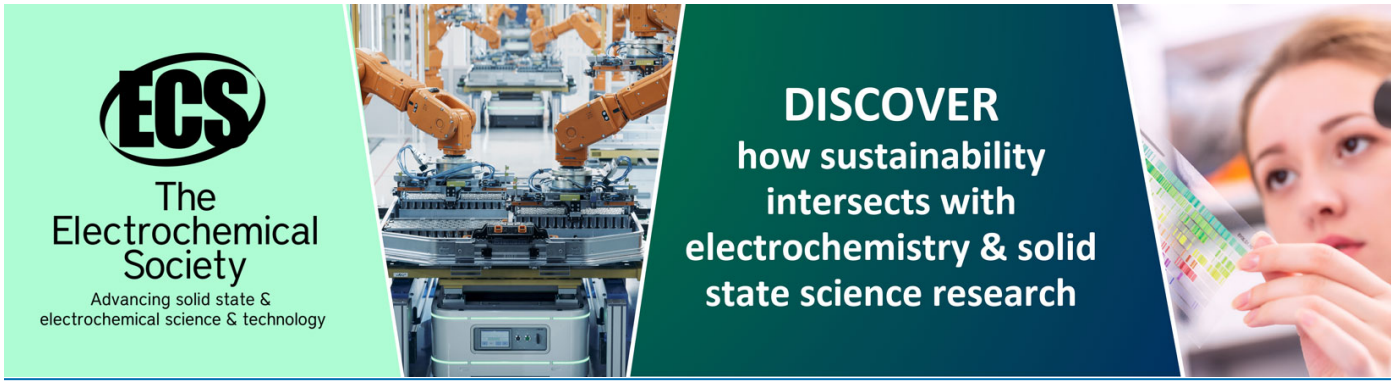
Using K-Means Clustering to Cluster Provinces in Indonesia



IOP Conf. Series: Journal of Physics: Conf. Series **1028** (2018) 012006 doi:10.1088/1742-6596/1028/1/012006

Using K-Means Clustering to Cluster Provinces in Indonesia

Ansari Saleh Ahmar[1,2*], Darmawan Napitupulu[3], Robbi Rahim[4], Rahmat Hidayat[5], Yance Sonatha[6], and Meri Azmi[7]

[1]Department of Statistics, Universitas Negeri Makassar, Makassar,90222, Indonesia
[2]AHMAR Institute, Makassar, 90222, Indonesia
[3]Research Center for Quality System and Testing Technology, Indonesian Institute of Sciences, Jakarta,12710, Indonesia
[4]School of Computer and Communication Engineering, Universiti Malaysia Perlis, Perlis, Malaysia
[5]Department of Information Technology, Politeknik Negeri Padang, Padang,25166, Indonesia
[6]Department of Information Technology, Politeknik Negeri Padang, Padang,25166, Indonesia
[7]Department of Information Technology, Politeknik Negeri Padang, Padang,25166, Indonesia

ansarisaleh@unm.ac.id

Abstract. K-Means Clustering (KMC) is a technique used in performing data groupings. The data classification procedure is based on the degree of membership of each member. The purpose of this study is to group the existing Provinces in Indonesia based on Population Density, School Participation Rate, Human Development Index, and Open Unemployment Rate using K -Means Clustering. The result reveals 5 large clusters in each center in South Sumatra, Lampung, DKI Jakarta, Central Java, and West Kalimantan.

1. Introduction

One of the biggest problems in Indonesia is the problems related to population. Based on U.S. Commission on International Religious Freedom, Indonesia in 2017 is the largest country of the population in the world and is one of the countries that have the largest Muslim population in the world [1]. We cannot be proud if the government cannot anticipate various problems related to the population. Governance in Indonesia is generally divided into 3 levels of government, namely central government, provincial government, and district/city governments. Each level of government has its own authority. Based on the level of government, the management of the region is also stratified according to its level. In helping the government to overcome the population, then one effort that can be done is to group the provinces in Indonesia to be easily classified in overcoming the problems that occur. In this research, there will be provincial grouping process in Indonesia based on population density, school participation rate 13-15, human development index, and open unemployment rate. The selection of these variables is based on the reason that these variables affect the problem of population in Indonesia.

Data mining is one of the processes undertaken to discover patterns and knowledge from large data [2]. In data mining [3]–[5] there are several methods that are often used to cluster data, including the

K-Means clustering [6]. K-Means clustering is an algorithm required as much input k which divides n objects into k cluster so that the level of similarity between members in one cluster is high while the level of resemblance to the members on the other cluster is very low [7]. The level of similarity between members in a cluster is measured by the proximity of the object to the mean value on the cluster or usually referred to as the centroid cluster. The K-Means method is the simplest and most commonly used method of clustering [8]. K-Means are often used because it has the ability to group large amount of data with computation time which is relatively fast and efficient. But the weakness of this method is the weakness in analyzing the distribution of data and depend on the initialization of the centroid. K-Means only looks at the data range to each centroid on each cluster.

2. Method

The data used in this study are population density, school participation rate of 13-15, human development index, and open unemployment rate of a province in Indonesia consisting of 34 provinces. In this study, Indonesian provinces will be grouped based on population density data, school participation rates of 13-15, human development index, and open unemployment rates. This provincial grouping will use the K-Means clustering method. The flowchart is as figure 1.

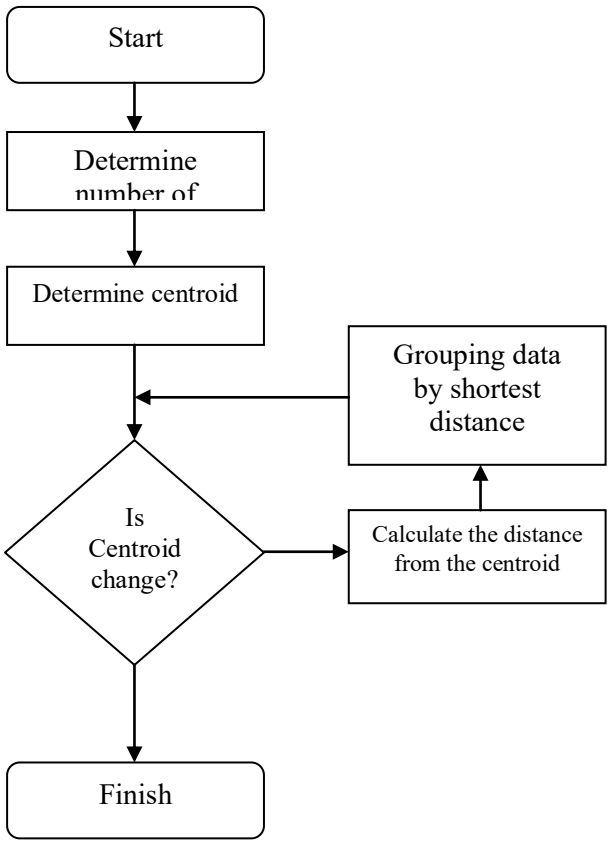


Figure 1 Flowchart of K-Means

3. Results and Discussion

K-Means clustering will be conducted on population density data, school participation rates 13-15, human development index, and open unemployment rate per province. The initial stage of this research is clustering with 5 clusters. The selection of this cluster is based on the presence of 5 major islands in Indonesia. The clustering results are as follows.

Table 1 Summary statistics of data

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
A	34	0	34	9,000	15328,000	711,559	2611,205
B	34	0	34	78,140	99,680	94,431	3,737
C	34	0	34	57,250	78,990	68,577	4,168
D	34	0	34	1,990	9,932	5,980	1,957

where:
A: Population Density (Life/Km^2),
B: School Participation Rate (APS) Age 13-15,
C: Human Development Index,
D: Open Unemployment Rate (TPT).

Table 2 Optimization Summary of data

Repetition	Iteration	Initial within-class variance	Final within-class variance	ln(Determinant(W))
1	**4**	**7172542,880**	**10420,173**	**28,404**
2	3	5859991,276	10420,173	28,404
3	3	6939820,692	10420,173	28,404
4	3	6906406,311	10420,173	28,404
5	3	6660860,968	10420,173	28,404
6	2	6656747,731	10420,173	28,404
7	2	7069908,060	10461,851	28,464
8	2	6830512,605	10420,173	28,404
9	2	5895848,461	10420,173	28,404
10	2	7013780,357	10517,276	28,456

Table 3 Statistics for each iteration of data

Iteration	Within-class variance	Trace(W)	ln(Determinant(W))	Wilks' Lambda
0	7172542,880	208003743,5	35,091	0,632
1	12172,588	353005,0498	28,418	0,001
2	10611,345	307728,9981	28,417	0,001

Table 4 Central objects of data

Class	A	B	C	D
1 (SUMATERA SELATAN)	88,000	93,520	67,460	6,072
2 (LAMPUNG)	234,000	94,240	66,950	5,137
3 (DKI JAKARTA)	15328,000	97,190	78,990	7,230
4 (JAWA TENGAH)	1030,000	95,300	69,490	4,993
5 (KALIMANTAN BARAT)	33,000	91,910	65,590	5,147

Table 5 Distances between the central objects of data

	1 (SUMATERA SELATAN)	2 (LAMPUNG)	3 (DKI JAKARTA)	4 (JAWA TENGAH)	5 (KALIMANTAN BARAT)
1 (SUMATERA SELATAN)	0	146,006	15240,005	942,004	55,063
2 (LAMPUNG)	146,006	0	15094,005	796,005	201,018
3 (DKI JAKARTA)	15240,005	15094,005	0	14298,003	15295,007
4 (JAWA TENGAH)	942,004	796,005	14298,003	0	997,013
5 (KALIMANTAN BARAT)	55,063	201,018	15295,007	997,013	0

Table 6 Results by class of data

Class	1	2	3	4	5
Objects	12	6	1	6	9
Sum of weights	12	6	1	6	9
Within-class variance	313,878	1294,971	0,000	58030,060	263,401
Minimum distance to centroid	1,141	20,544	0,000	18,810	8,360
Average distance to centroid	14,445	31,541	0,000	195,080	14,411
Maximum distance to centroid	35,157	46,543	0,000	330,693	22,824
	ACEH	SUMATERA UTARA	DKI JAKARTA	JAWA BARAT	KALIMANTAN BARAT
	SUMATERA BARAT	LAMPUNG		JAWA TENGAH	KALIMANTAN TENGAH
	RIAU	KEP. RIAU		DI YOGYAKARTA	KALIMANTAN TIMUR
	JAMBI	NUSA TENGGARA BARAT		JAWA TIMUR	KALIMANTAN UTARA
	SUMATERA SELATAN	SULAWESI UTARA		BANTEN	SULAWESI TENGAH
	BENGKULU	SULAWESI SELATAN		BALI	MALUKU
	KEP. BANGKA BELITUNG				MALUKU UTARA
	NUSA TENGGARA TIMUR				PAPUA BARAT
	KALIMANTAN SELATAN				PAPUA
	SULAWESI TENGGARA				
	GORONTALO				
	SULAWESI BARAT				

Table 7 Results by object of data

Observation	Class	Distance to centroid
ACEH	1	6,691
SUMATERA UTARA	2	22,681
SUMATERA BARAT	1	35,157
RIAU	1	16,447
JAMBI	1	21,112
SUMATERA SELATAN	1	1,141
BENGKULU	1	5,947
LAMPUNG	2	20,544
KEP. BANGKA BELITUNG	1	5,576
KEP. RIAU	2	27,881
DKI JAKARTA	3	0,000
JAWA BARAT	4	271,375
JAWA TENGAH	4	18,810
DI YOGYAKARTA	4	125,538
JAWA TIMUR	4	235,684
BANTEN	4	188,380
BALI	4	330,693
NUSA TENGGARA BARAT	2	46,543
NUSA TENGGARA TIMUR	1	16,898
KALIMANTAN BARAT	5	8,360
KALIMANTAN TENGAH	5	9,221
KALIMANTAN SELATAN	1	14,168
KALIMANTAN TIMUR	5	9,779
KALIMANTAN UTARA	5	16,004
SULAWESI UTARA	2	39,784
SULAWESI TENGAH	5	22,347
SULAWESI SELATAN	2	31,812
SULAWESI TENGGARA	1	23,023
GORONTALO	1	12,575
SULAWESI BARAT	1	14,607
MALUKU	5	12,398
MALUKU UTARA	5	11,844
PAPUA BARAT	5	16,918
PAPUA	5	22,824

Based on data clustering on population density data, school participation rates of 13-15, human development index, and open unemployment rate per province using K-Means clustering obtained 5 clusters with each cluster as follows: cluster 1 consists of 12 provinces (Aceh , West Sumatera, Riau, Jambi, South Sumatera, Bengkulu, Bangka Belitung Islands, East Nusa Tenggara, South Kalimantan, Southeast Sulawesi, Gorontalo, West Sulawesi), cluster 2 consists of 6 provinces (North Sumatra, Lampung, Riau Islands, West Nusa Tenggara, North Sulawesi, South Sulawesi), cluster 3 consists of 1

province (DKI Jakarta), cluster 4 consists of 6 provinces (West Java, Central Java, DI Yogyakarta, East Java, Banten, Bali), and cluster 5 consists from 9 provinces (West Kalimantan, Central Kalimantan, East Kalimantan, North Kalimantan, Central Sulawesi, Maluku, North Maluku, West Papua, Papua).

The provincial grouping is based on the proximity of other provinces to the provinces that are central objects: 1 (South Sumatra), 2 (Lampung), 3 (DKI Jakarta), 4 (Central Java), and 5 (West Kalimantan).

4. Conclusion

Based on the clustering resulted from K-Means Clustering, it was found that provincial groupings based on population density, school participation rate of 13-15, human development index and open unemployment rate were 5 clusters centered on South Sumatera, Lampung, DKI Jakarta, Central Java provinces, and West Kalimantan.

References

[1] United States Commission on International Religious Freedom (USCIRF), 2018, Indonesia Chapter - 2018 Annual Report. [Online]. Available: http://www.uscirf.gov/reports-briefs/annual-report-chapters-and-summaries/indonesia-chapter-2018-annual-report.

[2] Virk M and Chauhan V, 2018 Big Data and Shipping-managing vessel performance *JOIV Int. J. Informatics Vis.* **2**, 2 p. 73–75.

[3] Rahim R *et al.*, 2018 C4.5 Classification Data Mining for Inventory Control *Int. J. Eng. Technol.* **7**, 2.3 p. 68–72.

[4] Surahman Viddy A Gaffar A F O Haviluddin and Ahmar A S, 2018 Selection of the best supply chain strategy using fuzzy based decision model *Int. J. Eng. Technol.* **7**, 22 p. 117–121.

[5] Haviluddin Agus F Azhari M and Ahmar A S, 2018 Artificial Neural Network Optimized Approach for Improving Spatial Cluster Quality of Land Value Zone *Int. J. Eng. Technol.* **7**, 2.2 p. 80–83.

[6] Madadipouya K, 2017 A Survey on Data Mining Algorithms and Techniques in Medicine *JOIV Int. J. Informatics Vis.* **1**, 3 p. 61–71.

[7] Dubey A K Gupta U and Jain S, 2018 Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data *Int. J. Adv. Sci. Eng. Inf. Technol.* **8**, 1 p. 18–29.

[8] Sammour M and Othman Z, 2016 An Agglomerative Hierarchical Clustering with Various Distance Measurements for Ground Level Ozone Clustering in Putrajaya, Malaysia *Int. J. Adv. Sci. Eng. Inf. Technol.* **6**, 6 p. 1127–1133.