Zero: Jurnal Sains, Matematika, dan Terapan
E-ISSN: 2580-5754
P-ISSN: 2580-569X
Vol. 3, No. 2, December 2019
pp. 73-79

Application of K-Means Clustering Method for City Grouping on Food Plant Productivity in North Sumatera

Junita Fadillah[1], Sajaratud Dur[2], Ismail Husein[2]
[1]Department of Mathematics, Institut Pertanian Bogor, Indonesia
[2]Department of Mathematics, Universitas Islam Negeri Smatera Utara, Medan

Article Info

Article history:

Received October 23, 2019
Revised November 14, 2019
Accepted December 15, 2019

Keywords:

Food Crops,
Productivity,
K-Means.

ABSTRACT

The development of population increases every year causing food needs to increase, to meet food needs by increasing food crop productivity so that food availability can be sufficient. Food crops consist of rice, corn, green beans, peanuts, cassava, and sweet potatoes. Productivity in each region has different characteristics and therefore it is necessary to group the regions so that solution can be implemented in accordance with each of the characteristics of the region. The purpose of this study is to group districts/cities in North Sumatera Province based on food crop productivity using the k-means clustering method. Clustering k-means is method of grouping non-hierarchical data that attempts to partition existing data into one or more cluster or groups so that data that has the same characteristics are grouped into one same characterstics are grouped into other groups. The result of this study are the formation of 3 city district clusters namely, cluster 1 amounting to 1 regency/city, cluster 2 totaling 7 districts/cities, and cluster 3 totaling 25 districts/cities.

Corresponding Author:
Junita Fadillah,
Department of Mathematics,
Institut Pertanian, Bogor, Indonesia
Email: fadillahjunita@gmail.com

1. INTRODUCTION

The development of the population in Indonesia food is increasing in Indonesia is also increasing. Endurance Food in Indonesia is based on food, food utilization, and Food access needed sufficient for all regions in Indonesia.North Sumatra Province has the largest population on the island Sumatra and the fourth largest in Indonesia. From the results of the 2010 population census, the number of residents in North Sumatra is increasing every year. In 2018 there were 14.46 million people. The rate of population growth in 2010-2018 reached 1.30 percent per year higher than in 2000-2010 at 1.22 percent per year (BPS Population Census Results 2010).

In achieving its goals it is certainly not easy. Each region faces constraints by important factors that influence the achievement of productivity goals. Productivity is the area of harvested land and the production of crops. If this situation continues it will threaten national food availability in line with the increasing population every year. So the government must increase the productivity of food crops in each region so that food availability can be sufficient. To group regencies / cities based on food crop productivity using k-means clustering.

2. RESEARCH METHOD

Descriptive Analysis

Descriptive analysis aims to provide a description of the data into variables. Results can be seen from the mean (mean), maximum value, minimum value, median value, and standard deviation value (Ghozali, 2009).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

The formula looks for the standard deviation using the equation below:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \quad (2)$$

Data Standardization

Usually referred to as standardization of data, standardization of data is carried out when the variable being worked on contains large unit differences. Then it requires the process of standardizing data by transforming data (standardizing the original data before further analysis). Standardization is carried out on relevant variables into the form of Z-scores (Supranto, 2004).

$$Z_i = \frac{x_i - \bar{x}}{s_x} \quad \text{(Johnson dan Wichern, 2007)} \quad (3)$$

Cluster Analysis

Cluster analysis is a multivariate technique whose aim is to classify objects or cases (respondents) into relatively homogeneous groups, commonly called clusters. The objects or cases in each group tend to be similar to each other and are not the same as objects from other clusters. Cluster analysis is also called numerical classification or taxonomy (Supranto, 2004).

Hierarchy Method

The method of grouping two or more objects / data that has the closest similarity and the process is continued to other objects / data that have a second closeness (Rencher, 2002).

According to Machfudhoh (2013) states that the agglomerative method in the clustering hierarchy method is divided into several methods, namely:

1. Single Linkage
2. Complete Linkage
3. Average Linkage
4. Median Method
5. Ward Method
6. Centroid Method

Non-Hierarchy Method

The non-hierarchical method is called the k-means method. This method begins by determining the number of clusters or groups desired (two or three clusters). If the number of clusters is known, then the object of observation is combined into the cluster.

K-Means Clustering

The steps in the k-means clustering method are as follows:

1. Determine the number of clusters / k objects randomly (Madhulata, 2012).
2. Determine the initial centroid value (cluster center point) randomly as many as k cluster.
3. Calculate the distance of each object / data towards the center of the cluster to each cluster, using the Euclidian Distance formula (Nugroho, 2008).

$$d(x,y) = \sqrt{(x-y)'(x-y)}$$
$$= \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \ldots + (x_p-y_p)^2} \quad \text{(Archana Singh, 2013)} \quad (4)$$

The advantage of this method is that the distance between two objects / data will not be disturbed by the existence of new objects which are outliers. However, the distance can be affected by differences in the scale between dimensions where the distance is calculated (Dibya Jyoti Bora, 2014).

4. Allocate data into the minimum cluster center.
5. Iterate / repeat, then determine the position of the center of the new cluster.
6. Then if the cluster center point does not change again, the cluster process is complete, but if there is still data that moves the cluster, it is repeated again to step

Data obtained from the results of the North Sumatra Agricultural Census of the North Sumatra Central Statistics Agency (BPS) in 2018. Data variables used in this problem are 7 variables, namely rice, corn, soybeans, green beans, peanuts, sweet potatoes, and cassava.

3. RESULT AND ANALYSIS

Descriptive Analysis

Before carrying out the clustering-kmeans method, it is necessary to calculate the mean and standard deviation using equations 1 and 2, where the results are as follows:

Table 1. Descriptive Analysis Results

No.	Case	Rice (X_1)	Corn (X_2)	Soy Beans (X_3)	Green Beans (X_4)	Peanuts (X_5)	Sweet Potatoes (X_6)	Cassava (X_7)
1.	Minimum	0.00	0.00	0.00	0.000	0.00	0.00	0.00
2.	Mean	46.93	55.22	6.08	6.245	9.398	143.6	303.9
3.	Maksimum	62.69	73.73	11.80	11.970	17.630	236.3	433.3
4.	Standar d Deviation	11.6863	12.71792	2.95383	5.46673	5.95648	79.73544	73.79171

Data Standardization

If you know the value of the average and standard deviation, then look for the value of data standardization / standardization of data or commonly called the z-score. The formula of data standardization can be seen in equation (3), namely:

The following is the standardization of data from Nias districts / cities:

$$Z_i = \frac{x_i - \bar{x}}{s_X}$$
$$= \frac{44.03 - 46.93}{12.71792}$$
$$= 0.79427$$

And so on until the district / city of Gunung Sitoli.

K-means Clustering

1. Determine the number of k (cluster)

   According to Edmira Rivani (2010) that the number of clusters can be determined by the researchers themselves, but researchers use the R program, then the cluster number is determined as many as 3 clusters.

2. Specifies the initial centroid (cluster center) randomly.

Clusters were randomly formed as many as three clusters, so that there were three initial cluster centers of the k-means clustering method obtained/determined from three observation objects, namely the first cluster center in the form of Sibolga City, the second cluster center in the form of Gunung Sitoli City, and the third cluster center in the form Toba Samosir district. The results can be seen in the following table:

Table 2. Centroid (Cluster Center) Beginning

Variable	Initial Cluster Center Value		
	1	2	3
Rice (X_1)	-4.01575	-0.10517	1.24684
Corn (X_2)	-4.34178	0.16996	0.11177
Soy Beans (X_3)	-2.05845	1.93637	0.17594
Green Beans (X_4)	-1.14228	-1.14228	-1.14228
Peanuts (X_5)	-1.57786	-1.57786	0.85311
Sweet Potatoes (X_6)	-4.11852	-0.45510	0.96728
Cassava (X_7)	-1.80143	-1.80143	1.16187

3. Calculate the distance of each object / data to each cluster center point

Finding the distance value of each object to each center point of the cluster can use the Euclidian Distance formula in equation (4):

Calculation of Nias Regency with the first cluster center point:

$$D(x_i, y_j) = \sqrt{\begin{array}{l}(-0.90184 - (-4.01575)^2 + (0.79427 - (-4.34178))^2 + (-0.19307 - (-2.05845)^2 + \\ (0.86989 - (-1.14228))^2 + (-1.57786 - (-1.5779)^2 + (0.34472 - (-4.11852)) + \\ (0.5458 - (-1.80143)\end{array}}$$

$$= 8.30867$$

And so on do the district / city calculations at the point of the third cluster, then the results of the calculation are:

Table 3. Initial Cluster Center Point Distance Results

No.	Regency / City	Distance C_1	Distance C_2	Distance C_3	C_1	C_2	C_3
1.	Nias	8.30867	3.99303	3.9695			*
2.	Mandailing Natal	8.34406	3.56347	2.9836			*
:	:	:	:	:	:	:	:
33.	Gunung Sitoli	8.0639	0	4.6523		*	

4. Determine the new cluster center point

From the previous results, the next step is to determine the new cluster center to prove whether there are objects or variables that move like the previous step. The Calculation can be concluded through the table below:

Table 4. Final Cluster Center Distance Results

No.	Regency / City	Distance C_1	Distance C_2	Distance C_3	Minimum Distance	C_1	C_2	C_3
1.	Nias	8.431637	3.020428	2.455062	2.45506			*
2.	Mandailing Natal	8.466517	3.467289	1.753456	1.75345			*
:	:	:	:	:	:	:	:	:
33.	Gunung Sitoli	8.190611	2.053065	4.177793	2.053065		*	

From table 4 there are no cluster members that move, so the iteration / loop is stopped. Then the number of cluster members for the regencies / cities in North Sumatra can be obtained as follows:

a. In cluster I only consisted of 1 member, namely Regency / City of Sibolga.
b. Cluster II consists of 7 members, namely Central Tapanuli Regency / City, Labuhan Batu, Labuhan Batu Utara, West Nias, Tanjung Balai, Pematang Siantar, and Gunung Sitoli.
c. Cluster III consists of Nias Regency / City, Mandailing Natal, South Tapanuli, North Tapanuli, Toba Samosir, Asahan, Simalungun, Dairi Karo, Deli Serdang, Langkat, South Nias, Humbang Hasundutan, Pakpak Bharat, Samosir, Serdang Begadai, Batu Bara, Padang Lawas Utara, Padang Lawas, Labuhan Batu Selatan, North Nias, Tebing Tinggi, Medan, Binjai, and Padang Sidempuan.

5. Calculate the average of variables between cluster
Next to find out the average of the variables between clusters using the following equation:

$$X = m + z(s) \quad (5)$$

Information:
X = variable average into the cluster
μ = population
σ / S = standard deviation

Example calculation of the value of the first cluster with rice variables:

$= \left(46.93 + \left(-4.01572\right)\right)\left(11.6863\right)$

$= 0.0007$

Then the results of the calculation of the average variable between clusters are as follows:

Tabel 5. Results of Average Number of Variables into Clusters

Cluster	(x_1)	(x_2)	(x_3)	(x_4)	(x_5)	(x_6)	(x_7)
I	0.0007	0.0016	0.000	0.0005	-0.0007	-0.0111	-0.0346
II	46.7437	59.7701	17.9538	0.0005	0.3424	308.277	29.299
III	47.182	65.0615	11.8789	10.000	4.4965	235.214	150.000

Table 5 shows that there were 3 clusters formed. Each cluster is different from the smallest to the largest unit value in each variable. So it is said to be a low, medium, and high area. Where the average number of the first cluster consists of only one member of the cluster shows the lowest average total productivity compared to other clusters.

The second cluster consists of 7 cluster members with a high average number for the productivity of soybean and cassava food crops, while for medium food crop productivity, namely rice, peanuts, and sweet potatoes, and has the lowest average number for productivity green beans. The third cluster consists of 25 members of the cluster, the average number of high for the productivity of rice, corn,

peanuts, green beans, and sweet potatoes, while for soybeans, and cassava has a moderate average amount.

4. CONCLUSION

From the results and discussion, it can be seen from the food crop productivity data, and it can be concluded that the results of the cluster are formed into three clusters and have the characteristics of each cluster.

Cluster 1 only consists of Sibolga districts/cities with the lowest average number of food crop productivity. The second cluster consists of Central Tapanuli Regency Labuhan Batu, Labuhan Batu Utara, West Nias, and Kota Tanjung Balai, Pematang Siantar, and Gunung Sitoli with a high average number of soybean and cassava food crop productivity, while for crop productivity medium food, namely rice, peanuts, and sweet potatoes, and has the lowest average amount for the productivity of mung beans.

The 3rd Cluster consists of Regency / Nias City, Mandailing Natal, South Tapanuli, North Tapanuli, Toba Samosir, Asahan, Simalungun, Dairi Karo, Deli Serdang, Langkat, South Nias, Humbang Hasundutan, Pakpak Bharat, Samosir, Serdang Begadai, Batu Bara, Padang Lawas Utara, Padang Lawas, Labuhan Batu Selatan, North Nias, Tebing Tinggi, Binjai, Medan and Padang Sidempuan with high average amounts of productivity for rice, corn, green beans, peanuts, and sweet potatoes, while for the productivity of soybeans, and cassava has an average amount.

The above results show that the first cluster only has 1 cluster, namely Sibolga regency/city, which has the lowest food crop productivity, therefore the government must continue to increase food crop productivity so food availability can be sufficient for each region in North Sumatra.

References

[1] A. Y. A. R. Archana Singh. 2013. K-means with Three different Distance Metrics, *International Journal of Computer Applications*, **vol. 67**, no. 10.

[2] Badan Pusat Statistik Republik Indonesia. 2010. Sensus Penduduk 2010. Sumatera Utara: Badan Pusat Statsitik Provinsi Sumatera Utara

[3] D. A. K. G. Dibya Jyoti Bora. 2014. Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab, *International journal of computer science and 108 information technologies*, **vol. 5**, no. 2, pp. 2501-2506.

[4] Ghozali, Imam. 2009. Aplikasi Analisis Multivariate dengan Program SPS. Semarang: UNDIP.

[5] Husein, Ismail H Mawengkang, S Suwilo "Modeling the Transmission of Infectious Disease in a Dynamic Network" Journal of Physics: Conference Series 1255 (1), 012052, 2019.

[6] Husein, Ismail, Herman Mawengkang, Saib Suwilo, and Mardiningsih. "Modelling Infectious Disease in Dynamic Networks Considering Vaccine." Systematic Reviews in Pharmacy 11.2, pp. 261-266, 2020.

[7] Muqdad Irhaeem Kadhim, Ismail Husein. "Pharmaceutical and Biological Application of New Synthetic Compounds of Pyranone, Pyridine, Pyrmidine, Pyrazole and Isoxazole Incorporating on 2-Flouroquinoline Moieties." Systematic Reviews in Pharmacy 11 (2020), 679-684. doi:10.5530/srp.2020.2.98.

[8] Hamidah Nasution, Herlina Jusuf, Evi Ramadhani, Ismail Husein. "Model of Spread of Infectious Diseases." Systematic Reviews in Pharmacy 11 (2020), 685-689. doi:10.5530/srp.2020.2.99.

[9] Husein, Ismail, Dwi Noerjoedianto, Muhammad Sakti, Abeer Hamoodi Jabbar. "Modeling of Epidemic Transmission and Predicting the Spread of Infectious Disease." Systematic Reviews in Pharmacy 11.6 (2020), 188-195. Print. doi:10.31838/srp.2020.6.30

[10] Husein, Ismail, YD Prasetyo, S Suwilo "Upper generalized exponents of two-colored primitive extremal ministrong digraphs"AIP Conference Proceedings 1635 (1), 430-439, 2014

[11] Husein, Ismail. 2017.Filsafat Sains. Medan: Perdana Publishing.

[12] I Husein, RF Sari, H Sumardi, M Furqan, 2017, Matriks dan transformasi linear, Jakarta: Prenada Media Group

[13] S Sitepu, H Mawengkang, I Husein "Optimization model for capacity management and bed scheduling for hospital" IOP Conference Series: Materials Science and Engineering 300 (1), 01,2016.

[14] Syah Rahmad, M K M Nasution, Ismail Husein, Marischa Elveny, "Optimization Tree Based Inference to Customer Behaviors in Dynamic Control System", International Journal of Advanced Science and Technology, pp. 1102 – 1109,2020.

[15] Husein Ismail, Rahmad Syah, "Model of Increasing Experiences Mathematics Learning with Group Method Project", International Journal of Advanced Science and Technology, pp. 1133-1138, 2020.

[16] Syah Rahmad, Mahyuddin K.M Nasution, Ismail Husein, "Dynamic Control Financial Supervision (OJK) for Growth Customer Behavior using KYC System", International Journal of Advanced Science and Technology, pp. 1110 – 1119, 2020.

[17] Muqdad Irhaeem Kadhim, Ismail Husein, Lelya Hilda, Sajaratud Dur, Abeer Hamoodi jabbar. "The Effect for Chloroquines and Hydroxychloroquines as Experimental therapy of Coronavirus-19." Journal of Critical Reviews 7 (2020), 305-309. doi:10.31838/jcr.07.17.43

[18] Hawraa A. Al-Ameer Humood, Ismail Husein, Lelya Hilda, Sajaratud Dur, Muqdad I Kadhim. "Synthesis the seven-ring compounds (oxazepine) from the principles of schiff bases and study the biological activity of them." Journal of Critical Reviews 7 (2020), 292-304. doi:10.31838/jcr.07.17.42