Evaluating the K-Means Analysis in Clustering Area Based on Estates Productivity in Tana Luwu Using Silhouette Index

B Poerwanto 1*  
1Informatics Engineering, Universitas Cokroaminoto Palopo, Indonesia
*e-mail: bobbybp89@gmail.com

Abstract. Tana Luwu is an area in South Sulawesi Province that consists of 4 districts or city, namely Luwu, East Luwu, North Luwu, and Palopo. Those regions have abundant natural resources, one of which is plantation crops. In this manuscript, the first objective was to group the 45 sub-districts in those districts into 2 to 5 clusters based on their productivity on providing plantation crops using K-Means Algorithm. Variables used in this study are coconut, palm, coffee, pepper, cocoa, cloves, and land area. Secondly, after the clusters were formed, those were evaluated using Silhouette Index to find out the most optimum number of clusters. As the result, it was found that the optimum number was two with Silhouette Index 0.8068 while the value of 3 clusters is 0.6468. For the values of 4 clusters and 5 clusters are 0.5029 and 0.5509 respectively.

Keywords: K-Mean Analysis, Estates Productivity, Silhouette Index

1. Introduction

Cluster analysis is one method in machine learning that aims to group objects based on the same properties and characteristics where the groups formed will be homogeneous between members in the same group and heterogeneous between groups [1, 2]. One approach commonly used in cluster analysis is to use the k-means algorithm. The advantage of k-means algorithm is that it is able to group large objects and outliers objects very quickly and the iteration process on this algorithm will stop in local optimum [3].

K-means algorithm is now widely used in various fields, for example Ong [4] has used the k-means algorithm in determining marketing strategies for President University. In addition there is also Niszczoza [5] using k-means in classifying tourists in Eastern Poland. Meanwhile, Bastian [6] applied this algorithm to classify human infectious diseases in Majalengka Regency.

The k-means algorithm in this study will be used for grouping sub-districts in Tana Luwu based on the results of estates crops. Tana-Luwu is also referred to as Bumi Sawerigading, a former area of the Kingdom of Luwu located in the province of South Sulawesi. Administratively, Tana-Luwu is divided into 4 districts and cities, namely Luwu Regency with the capital of Belopa, Palopo City, North Luwu Regency with the capital city of Masamba and East Luwu Regency with the capital city of Malili.

The land in Tana Luwu is known to be quite fertile for the food crops, plantations or vegetables sector. Specifically for estate crops, based on data from the Statistics Agency in Indonesia, the area of North Luwu's plantation area is around 9.05% of the total area. For East Luwu Regency the plantation area is 4.32%, Luwu Regency is 17.23% and for Palopo City is 20.18% of the total area. From the results of grouping sub-districts based on the results of the production of estate crops, later it will be known which sub-district produces the most productive so that it can help to control the distribution of estate...
In this study, the number of clusters will be made as many as 2 to 5 clusters, then the optimum number of clusters will be found by using the Silhouette Index.

2. Material and Method
2.1 Data and Variables
The data used in this study are secondary data from the Central Statistics Agency for Palopo City, Luwu Regency, East Luwu Regency, and North Luwu Regency in 2018 regarding the yield of plantations in Tana Luwu, and the variables used are the production of coconut, oil palm, coffee, pepper, cocoa, cloves, and land area for plantations.

2.2 K-Means Algorithm
Based on Rencher [7], data analysis on the cluster method using the k-means algorithm has the following steps:
a. Determining the number of clusters (k). In this study the number of clusters was as much as 2 to 5.
b. The second step was to determine the centroid randomly.
c. The third step is to calculate the Euclidean distance to the centroid using equation (1) below:

\[ d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2} \]  

(1)

where \( x_{i1}, x_{j1} \) are the two data to be calculated for the distance, \( p \) is the number of dimensions used, and \( d(i, j) \) is the euclidean distance to be measured from the data to the center of the cluster (centroid).
d. Forming initial groups based on calculated Euclidean distances.
e. Determining the new centroid using equations:

\[ C_{m(q)} = \frac{1}{n_m} \sum_{i=1}^{n_m} x_{i(q)} \]

dimana

\( C_{m(q)} \) : the \( m^{th} \) centroid for the \( q^{th} \) variable.
\( m \) : 1, 2, \ldots, \( k \)
\( n_m \) : the number of objects in the \( m^{th} \) group.
\( k \) : number of cluster.
\( q \) : 1, 2, \ldots, \( p \)
\( x_{i(q)} \) : the observation value of the \( i^{th} \) object in the \( q^{th} \) variable.
\( i \) : 1, 2, \ldots, \( n \)

Do iterations by repeating steps c, d, and e until no more data move between groups.

2.3 Silhouette Index Algorithm
This method is quite often used to evaluate the results of cluster analysis that combine cohesion and separation values [8, 9]. The SI algorithm is as follows:
a. Calculating \( a_i^l \):

\[ a_i^l = \frac{1}{m_j-1} \sum_{r \neq i}^{m_j} d(x_i^l, x_r^l), \quad i = 1, 2, \ldots, m_j \]  

(2)

where \( a_i^l \) : average distance of the \( i^{th} \) data to all other data in 1 cluster.
\( m_j \) : the amount of data in the \( j^{th} \) cluster.
\( d(x_i^l, x_r^l) \) : distance of the \( i^{th} \) data with the \( r^{th} \) data in one cluster \( j \).
b. Calculating $b_i$:

$$b_i = \min_{n=1,...,k} \left\{ \frac{1}{m_n} \sum_{r \neq i}^{m_j} d(x_i^r, x^n_r) \right\}, \quad i = 1, 2, 3, \ldots , m_j$$

$b_i$ : the minimum of the average distance for the $i^{th}$ data to all data from other clusters.

c. Calculating SI for the $i^{th}$ data:

$$SI_i^j = \frac{b_i^j - a_i^j}{\max(a_i^j, b_i^j)}$$

d. Calculating the average of SI:

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j$$

e. Calculating the global SI:

$$SI = \frac{1}{k} \sum_{j=1}^{k} SI_j$$

3. Result and Discussion

3.1 Result

There are 45 sub-districts in Tana Luwu that will be the focus for clustering in this study. It is given a general description of the largest plantation production in Tana Luwu, the 5 biggest producing subdistricts for each estates crop in table 1 below.

| Estate Crops | Subdistricts     | Production (Ton) | District       |
|--------------|------------------|------------------|----------------|
| Coconut      | Mappedeceng      | 906,36           | North Luwu     |
|              | Tomoni           | 591,5            | East Luwu      |
|              | Burau            | 515              | East Luwu      |
|              | Bupon            | 433,88           | Luwu           |
|              | Angkona          | 375,55           | East Luwu      |
| Oil Palm     | Tanalili         | 69910,62         | North Luwu     |
|              | Sukamaju         | 63721,89         | North Luwu     |
|              | Baebunta         | 61534,31         | North Luwu     |
|              | Malangke Barat   | 53961,41         | North Luwu     |
|              | Bone-Bone        | 510076,96        | North Luwu     |
| Coffee       | Bastem           | 505,25           | Luwu           |
|              | Seko             | 450,21           | North Luwu     |
|              | Rongkong         | 231              | North Luwu     |
|              | Walenrang Utara  | 140,89           | Luwu           |
|              | Rampi            | 76,64            | North Luwu     |
| Pepper       | Towuti           | 3222,8           | East Luwu      |
|              | Wasuponda        | 429,2            | East Luwu      |
|              | Burau            | 233,33           | East Luwu      |
|              | Malili           | 202,17           | East Luwu      |
|              | Larompong        | 157              | Luwu           |
| Cocoa        | Baebunta         | 6996,45          | North Luwu     |
|              | Sabbang          | 6146,3           | North Luwu     |
|              | Bupon            | 5259,13          | Luwu           |
In table 1, it can be seen that for each estate crop, 5 subdistricts are displayed with the highest number of production. From the 6 crops, 3 of them are subdistricts originating from North Luwu Regency as the biggest producers, namely Mappedeng Subdistrict for coconut plants, Tanalili Subdistrict for palm oil, and Baebunta Subdistrict for cocoa. Then, there are two subdistricts of Luwu Regency as the biggest producers, namely for the Bastem Subdistrict for coffee plants, and the North Walenrang subdistrict for cloves. Specifically for clove plants themselves, the 5 most productive subdistricts are from Luwu Regency.

In implementing the K-Means algorithm, it will be simulated to create 2 to 5 clusters. And, the most optimum will be chosen using the Silhouette Index. As a first step, here are the centroids that are randomly selected for each cluster.

**Table 2. Selection of centroids of each cluster**

| No. of Clusters | Subdistrict | Coconut | Oil Palm | Coffee | Pepper | Cacao | Cloves | Land Area |
|-----------------|-------------|---------|----------|--------|--------|-------|--------|-----------|
| 2               | Larompong Selatan | 179.33 | 0        | 20.16  | 7.81   | 2081.56 | 2975.05 | 8008.55  |
|                 | Angkona     | 375.55 | 28815.35 | 0.72   | 8.32   | 1043.2 | 1.54   | 4960.15  |
| 3               | Larompong Selatan | 179.33 | 0        | 20.16  | 7.81   | 2081.56 | 2975.05 | 8008.55  |
|                 | Angkona     | 375.55 | 28815.35 | 0.72   | 8.32   | 1043.2 | 1.54   | 4960.15  |
|                 | Towuti      | 94     | 910      | 1.46   | 3222.8 | 270.01 | 7      | 3260.16  |
| 4               | Larompong Selatan | 179.33 | 0        | 20.16  | 7.81   | 2081.56 | 2975.05 | 8008.55  |
|                 | Angkona     | 375.55 | 28815.35 | 0.72   | 8.32   | 1043.2 | 1.54   | 4960.15  |
|                 | Towuti      | 94     | 910      | 1.46   | 3222.8 | 270.01 | 7      | 3260.16  |
|                 | Wasuponda   | 69.88  | 6926     | 10.22  | 429.2  | 2509.2 | 0      | 6770.2   |
| 5               | Larompong Selatan | 179.33 | 0        | 20.16  | 7.81   | 2081.56 | 2975.05 | 8008.55  |
|                 | Angkona     | 375.55 | 28815.35 | 0.72   | 8.32   | 1043.2 | 1.54   | 4960.15  |
|                 | Towuti      | 94     | 910      | 1.46   | 3222.8 | 270.01 | 7      | 3260.16  |
|                 | Wasuponda   | 69.88  | 6926     | 10.22  | 429.2  | 2509.2 | 0      | 6770.2   |
|                 | Rampi       | 32.8   | 0        | 76.64  | 10.78  | 154.54 | 0      | 326.54   |

Furthermore, by implementing the K-Means algorithm in the data, several results were obtained. First, the number of iterations needed so that the data did not move to another group, and the determination of members included in the category of each cluster. The results can be seen in table 3 below.

**Table 3. Membership of each cluster**

| No.of Clusters | Cluster | Amount | Maximum Iteration |
|----------------|---------|--------|-------------------|
| 2              | Cluster 1 | 5      | 5                 |
|                | Cluster 2 | 40     |                    |
Next step, using the algorithm of the Silhouette Index for determining the best cluster in grouping estate crops in Tana Luwu. The results can be seen in table 4 below:

| Banyaknya Cluster | Silhouette Index |
|-------------------|------------------|
| 2                 | 0.8068           |
| 3                 | 0.6468           |
| 4                 | 0.5029           |
| 5                 | 0.5509           |

### 3.2 Discussion

In this study, cluster analysis was conducted to see which sub-district is the most stable in producing plantation crops or in other words the considered the most productive. The sub-districts in Tana Luwu would be clustered into 2 to 5 groups and later the Silhouette Index method would be used to determine the optimum number of clusters in this case.

Table 4 shows the SI values for each cluster. SI value is in the interval \([-1, +1]\) and the best if the value is getting closer to 1. So, based on table 4 it can be concluded that the 2 clusters were the optimum because it got the SI value of 0.8068 where this value is higher than the others. These parameters were calculated by using Matlab program where a lot of statistical analysis can be directly used [10]. Based on studies from Subbalakshmi et al [11], this SI method is very good in evaluating the results of cluster analysis, and this method can be modified to be able to handle static and dynamic data. In addition, Table 3 also shows the number of iterations for each cluster selection. For selection 2, 4, and 5 clusters, the number of iterations was 5, while for 3 clusters, the number of iterations was 6. For all experiments, the results for the 5 groups listed in table 5 below are always in the same group for each cluster. The sub-district groups formed by selecting 2 clusters are as follows:

| No | Subdistrict | District     |
|----|-------------|--------------|
| 1  | Baebunta    | North Luwu   |
| 2  | Malangke Barat | North Luwu |
| 3  | Sukamaju    | North Luwu   |
| 4  | Bone-Bone   | North Luwu   |
| 5  | Tanalili    | North Luwu   |

In table 5 it can be seen that there are only 5 subdistricts in cluster 1, and all of these sub-districts are from North Luwu Regency. This is also supported by research by McMahon et al [12] which tested some of the best cocoa producing areas on the island of Sulawesi, and one of them is North Luwu. A description of the differences in each of the districts can be seen visually in Figure 1 below.
Furthermore, for the second cluster which is categorized as less productive, there are 40 subdistricts spread across 3 regencies and 1 city.

Table 6. Subdistricts in cluster 2

| No | Subdistrict      | District   | No | Subdistrict      | District   |
|----|------------------|------------|----|------------------|------------|
| 1  | Wara Selatan     | Palopo     | 21 | Walenrang        | Luwu       |
| 2  | Sendana          | Palopo     | 22 | Walenrang Utara  | Luwu       |
| 3  | Wara             | Palopo     | 23 | Burau            | East Luwu  |
| 4  | Wara Timur       | Palopo     | 24 | Wotu             | East Luwu  |
| 5  | Mungkajang       | Palopo     | 25 | Tomoni           | East Luwu  |
| 6  | Wara Utara       | Palopo     | 26 | Tomoni Timur     | East Luwu  |
| 7  | Bara             | Palopo     | 27 | Angkona          | East Luwu  |
| 8  | Telluwanua       | Palopo     | 28 | Malili           | East Luwu  |
| 9  | Wara Barat       | Palopo     | 29 | Towuti           | East Luwu  |
| 10 | Larompong        | Luwu       | 30 | Nuha             | East Luwu  |
| 11 | Larompong Selatan| Luwu       | 31 | Wasuponda        | East Luwu  |
| 12 | Suli             | Luwu       | 32 | Mangkutana       | East Luwu  |
| 13 | Suli Barat       | Luwu       | 33 | Kalaena          | East Luwu  |
| 14 | Belopa           | Luwu       | 34 | Sambang          | North Luwu |
| 15 | Kamanre          | Luwu       | 35 | Malangke         | North Luwu |
| 16 | Belopa Utara     | Luwu       | 36 | Masamba          | North Luwu |
| 17 | Bajo             | Luwu       | 37 | Mappedeceng      | North Luwu |
| 18 | Bastem           | Luwu       | 38 | Rampi            | North Luwu |
| 19 | Bupon            | Luwu       | 39 | Rongkong         | North Luwu |
| 20 | Bua              | Luwu       | 40 | Seko             | North Luwu |

It can be seen in table 5 that the sub-district groups in cluster 2 where this cluster can be considered a cluster with a production value that is less stable or less productive when compared to the group in cluster 1.

Initially, cluster analysis was widely used for health data [13][14]. Over time, this analysis has begun to be used for other fields, such as education [15]. This study shows that cluster analysis using the k-means algorithm can also be applied to agriculture. So that the mindset of using the cluster analysis method is getting wider. In addition, this research will also have an impact on the socio-economic sector, namely areas that are considered productive in estate crops so that they can be used as centers for purchasing estate crops. This can reduce the cost of transportation and distribution, which has so far
taken both estate crops for resale and for sale outside Tana Luwu. The indirect benefit of this research is that it can be used in policy making grounds to determine what needs to be improved in terms of estate crops for sub-districts that are considered productive, low-income and low-yielding. So that later it is expected to strike a balance between the provider sub-districts and the sub-districts that use them.

4. Conclusion
Based on the analysis that has been done, it can be concluded that cluster 1 which contains the districts with the highest level of productivity, all come from North Luwu namely Baebunta, West Malangke, Sukamaju, Bone-Bone, and Tanalili. Whereas, for the other 40 sub-districts are in cluster 2. In conducting experiments for 2 to 5 clusters, it is known that the average iteration is done 5 times, and only 3 clusters have 6 iterations. As for selecting the optimum number of clusters using the Silhouette Index algorithm it is found that the optimum number of clusters is 2 with a SI value of 0.8068.

Acknowledgement
The authors give great thanks to those who have provided funding assistance, in this case Universitas Cokroaminoto Palopo through an internal research grant scheme.

References
[1] Fränti P, Sieranoja S. How much can k-means be improved by using better initialization and repeats? Pattern Recognit 2019; 93: 95–112.
[2] Poerwanto B, Fa’rifah RY. Analisis cluster k-means dalam pengelompokan kemampuan mahasiswa. Indones J Fundam Sci 2016; 2: 92–96.
[3] Chatterjee S, Mukhopadhyay A. Clustering Ensemble : A Multiobjective Genetic Algorithm based Approach. Procedia Technol 2013; 10: 443–449.
[4] Ong JO. IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING UNTUK MENENTUKAN STRATEGI MARKETING. J Ilm Tek Ind 2013; 12: 10–20.
[5] Borkowska-Niszczota M. Tourism Clusters in Eastern Poland - Analysis of Selected Aspects of the Operation. Procedia - Soc Behav Sci 2015; 213: 957–964.
[6] Bastian A, Sujadi H, Febrianto G. Penerapan Algoritma K-Means Clustering Analysis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka). J Sist Inf 2018; 14: 26–32.
[7] Tan P, Steinbach M, Kumar V. Introduction to Data Mining. Pearson Addison-Wesley, 2006.
[8] Petrovic S. A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters. In: Proceedings of the 11th Nordic Workshop of Secure IT Systems. 2006. pp. 53–64.
[9] Poerwanto B, Farifah RY, Sanusi W, et al. A matlab code to compute prediction of survival trends in patients with DHF A matlab code to compute prediction of survival trends in patients with DHF. J Phys Conf Ser 2018; 1028: 1–7.
[10] Subbalakshmi C, Krishna GR, Mohan SK, et al. A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set. Procedia Comput Sci 2015; 46: 346–353.
[11] Mcmahon PJ, Susilo AW, Parawansa AK, et al. Testing local cacao selections in Sulawesi for resistance to vascular streak dieback. Crop Prot 2018; 109: 24–32.
[12] Prasetyo E. Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab. 2014.
[13] Farifah RY, Poerwanto B. Platelets and Hematocrit in The Survival Model of Dengue Hemorrhagic Fever (DHF) Sufferers in Palopo. Mater Sci Forum 2019; 967: 3–8.
[14] Pasina I, Bayram G, Labib W, et al. Clustering students into groups according to their learning style. MethodsX 2019; 6: 2189–2197.