Cluster Analysis Using the Hierarki Method
For Grouping Sub-Districts in The District
Steps Based on Health Indicators

Khairun Nisa¹, Rina Filia Sari², Hendra Cipta¹, Ismail Husein³
¹Department of Mathematics, Universitas Islam Negeri Maulana Malik Ibrahim Malang
³Department of Mathematics, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia

ABSTRACT
Cluster analysis is a method used to group objects based on similarity of characteristics they have. Cluster analysis using the hierarchy method is a method with a grouping process that is used in stages. Health is a condition where a person is not sick, has no complaints, and can carry out daily activities. To find out information about the level of health in Langkat Regency, it is necessary to use the grouping method. The grouping was carried out in 23 districts in Langkat Regency. The purpose of this study is to classify sub-districts in Langkat Regency which have similar characteristics based on health indicators through the square euclidian distance is used to measure the similarity between object pairs and the ward method. From the results of cluster analysis using the ward method.

Article info
Received February 15, 2020
Revised March 14, 2020
Accepted April 24, 2020

Keywords:
Cluster Analysis, Hierarchical Methods, Health, Square Euclidian Distance.

1. INTRODUCTION
Human health is a basic need that is multi-nature so it is important to always pay attention, because health is the first and main asset in human survival. Health is also one part of welfare, because people who are more successful in human development are those who have a high level of health.

To describe the health condition of the community in Langkat Regency, indicators of health status were used including morbidity, birth attendants and life expectancy.

In the 2017 National Social Survey (SENSUS), residents of Langkat District who experienced health complaints amounted to 15.65 percent, 46.24 percent for outpatient treatment to overcome these health complaints, while 53.76 percent did not seek outpatient treatment. This is because, because they do not have medical expenses (2.11 percent), do not have transportation costs (0.07 percent), they treat themselves (79.25 percent), feel unnecessary (17.61 percent) and other reasons (0.96 percent) (BPS, 2018). The low awareness, willingness and ability to live a healthy life in Langkat District has resulted in a decreased level of health in Langkat District, this is evidenced by the large number of people experiencing health complaints. Based on the description.
2. **RESEARCH METHODE**

Cluster Analysis

Cluster analysis is a multivariate analysis (many variables) that functions to group objects or several variables based on their characteristics. In addition, cluster analysis also aims to maximize the similarity of objects in the cluster while also maximizing differences between clusters (Hair, 2009).

The process of processing data so that data sets can be formed into clusters using cluster analysis is as follows (Santo, 2015).

### Setting the Distance Between Data Sizes

The measure used in measuring the similarity between data in cluster analysis is Square euclidian distance (squared euclidean distance).

\[
    d_{ij} = \sum_{k=1}^{p} (y_{ik} - y_{jk})^2 ; i, j = 1, 2, ..., n
\]

(2.1)

The distance between objects can be written in the form of a matrix. The matrix is a collection of numbers (elements or etri in the form of real or complex items) arranged in rows and columns to form a rectangle that is m x n size enclosed in square brackets (Husein, Rina, and Hari 2017), as in the equation (2.2).

\[
    D_{n x n} = \begin{pmatrix}
        d_{11} & d_{12} & \cdots & d_{1n} \\
        d_{21} & d_{22} & \cdots & d_{2n} \\
        \vdots & \vdots & \ddots & \vdots \\
        d_{n1} & d_{n2} & \cdots & d_{nn}
    \end{pmatrix}
\]

(2.2)

Where is the distance between objects to \( i \) and to \( j \) for each \( i, j = 1, 2, ..., n \).

### Conduct Data Standardization Process

In the process of standardizing data, the first thing to consider is whether the data unit has outliers or different data on a large scale (outlier) among the research variables. Detection of outlier data can be done by determining the boundary value that is used as part of outlier data, by changing the score from the initial or raw data into standardized score (z-score), with the result of standard deviation of one and means (average) zero.

\[
    Z = \frac{x_i - \bar{x}}{S_x}
\]

(2.3)

### Clustering Process

Clustering process is a process carried out by two methods, namely the hierarchical and non-hierarchical methods.

#### a. Hierarchy Method

Hierarchical method is a method that is done in stages. In this method will form a certain stage as in the tree structure and can be produced in the form of a dendogram. Dendogram is a visual representation of the stages of the cluster analysis process that is formed which produces the value of the distance coefficient at each stage. The result in the form of a number to the right of the dendogram is the object of research, because there is a line that connects these objects with other objects to form a cluster (Simamora, 2005).

**Single Link Method (Single Linkage Method)**

Single linkage method (the closest distance) or a single link can be done by grouping data based on the shortest distance (Rencher, 2002).

\[
    D(A, B) = \min \{d(y_i, y_j) \text{ for } y_i \text{ in } A \text{ and } y_j \text{ in } B\}
\]

(2.4)

**Complete Link Method (Complete Linkage Method)**

Complete linkage method (long distance) can be done by grouping data based on the longest distance (Rencher, 2002).

\[
    D(A, B) = \max \{d(y_i, y_j) \text{ for } y_i \text{ in } A \text{ and } y_j \text{ in } B\}
\]

(2.5)
Average Link Method (Average Linkage Method)

Average linkage method is a method that is done by grouping data based on the average distance between the whole data (Rencher, 2002).

\[ D(A, B) = \frac{1}{n_An_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j) \]  

(2.6)

Ward Method (Ward’s Method)

Ward’s method is clustering by maximizing the similarity in one cluster and using a complete calculation. At each stage, the distance between the two clusters that can be formed is Sum of Square Error (SSE) in the two smallest clusters combined (Rencher, 2002).

\[ SSE = \sum_{i=1}^{n} (y_i - \overline{y}) (y_i - \overline{y}) \]  

(2.7)

If A, B and AB are clusters, then the sum of the squares of errors in the cluster are:

\[ SSE_A = \sum_{i=1}^{n_A} (y_i - \overline{y}_A) (y_i - \overline{y}_A) \]

\[ SSE_B = \sum_{i=1}^{n_B} (y_i - \overline{y}_B) (y_i - \overline{y}_B) \]

(2.8)

\[ SSE_{AB} = \frac{2n_{AB}}{n_A + n_B} \sum_{i=1}^{n_{AB}} (y_i - \overline{y}_{AB}) (y_i - \overline{y}_{AB}) \]

The ward method can join two clusters A and B which can minimize the increase in Sum of Square Error (SSE). Defined as follows:

\[ I_{AB} = SSE_{AB} - (SSE_A + SSE_B) \]  

(2.9)

It can be seen that the increase in the \( I_{AB} \) in equation (2.9) has the following equivalent form (Sukmawati, 2017):

\[ I_{AB} = SSE_{AB} - (SSE_A + SSE_B) \]

\[ I_{AB} = n_A (\overline{y}_A - \overline{y}_{AB}) (\overline{y}_A - \overline{y}_{AB}) + n_B (\overline{y}_B - \overline{y}_{AB}) (\overline{y}_B - \overline{y}_{AB}) \]

(2.10)

Based on equation (2.10), it is the result of minimizing the increase in equivalent SSE by minimizing distances between objects. \( SSE_A \) and \( SSE_B \) will be zero if A only consists of \( y_i \) and B consists of \( y_j \). In equations (2.9) and equation (2.10) produce equations with formulas that will be used in calculating the distance between objects using the ward method as follows:

\[ I_{ij} = SSE_{ij} = \frac{1}{2} (y_i - y_j) (y_i - y_j) \]

\[ = \frac{1}{2} d^2 (y_i, y_j) \]

\[ = \frac{1}{2} \sum_{k=1}^{n} (y_{ik} - y_{jk})^2 \]  

(2.11)

Central Method (Centroid Method)

Centroid method also called the center point method, where the distance between clusters in the centroid method is the distance between centroids. If a new cluster formation occurs, there will be a recalculation (Rencher, 2002).

\[ D(A, B) = d(\overline{y}_A, \overline{y}_B) \]  

(2.12)
Cluster Analysis Using the Hierarchical Method for Grouping Sub-Districts in The District Steps Based On Health Indicators

After two clusters and join, center of the cluster can be given as follows:

\[
\bar{y}_{AB} = \frac{n_A \bar{y}_A + n_B \bar{y}_B}{n_A + n_B}
\]  

(2.13)

b. Non-Hierarchy Method

The non-mathematical method is also called the method k-means. This method is not the same as the hierarchical method, because the non-hierarchical method starts by determining in advance a number the cluster desired start, then the results of these observational objects merge and form the cluster.

Data Sources and Research Variables

In this study, the data used are secondary data obtained from the Langkat District Health Office in 2018. This study uses three indicators of community health degrees with six variables.

Data analysis

1. Gather references regarding cluster analysis and health indicators
2. Collecting data on health indicators obtained from the District Health Office of Langkat
3. Describe health indicator data
4. Using data standardization
5. Determine a procedure in cluster analysis
   - In this study using a cluster analysis of the hierarchical method using the ward method
6. Conduct cluster analysis results
7. After obtaining the results of the cluster analysis, the next step is to interpret the results of the cluster formed
8. Conclusions and recommendations.

3. RESULT AND ANALYSIS

Data Standardization

In the process of standardizing data, calculations are performed the z-score obtained by using equation (2.3). Here is an example of calculating z-scores for variables \(X_1\) (number of deliveries assisted by health workers) in Bahorok sub-district.

\[
Z_{X_1, Bahorok} = \frac{x_1 - \bar{x}_1}{s_{x_1}} = \frac{839 - 918.2609}{382.2111} = -0.207374612
\]

This process continues until the entire data produces the same unit scale.

Calculating Data Between Sizes

This process is done after standardizing the data, the method used to calculate the size between objects is the square euclidian distance (squared euclidean distance) with the equation formula (2.1). The following is a sample by calculating the size between Bahorok sub-district and Serañat sub-district (objects 1 and 2).

\[
d_{(1,2)} = \left( (-0.207374612 - (-1.4355593693))^2 + (0.265885835 - (0.43908295 - (-1.4355593693))^2
\right.
\]

\[
+ (0.495497392 - (-1.075830657))^2 + (0.438783165 - (0.697964858))^2
\]

\[
= 1.564045197 + 1.39823086 + 1.824773467 + 0.067019842 + 0.336786999
\]

\[
= 5.76358514
\]

This process continues until knowing the overall size of the distance between objects.
Based on the calculation of the size between 23 districts using square euclidian distance (squared euclidean distance), it is known that the closest pair of objects is the Salapian district and the Sawit Seberang district with the closest distance of 0.396188.

**Cluster Analysis Process Using the Ward Method**

The clustering process in the hierarchical method with the ward method is carried out using the two closest objects (districts), where the distance is the closest between the distances of 23 objects (districts) that exist. For example Bahorok sub-district and Serapit sub-district by using equation (2.11).

\[ I_{ij} = SSE_{ij} = \frac{1}{2} \left( y_i - y_j \right)^2 \]

\[ SSE_{(Bahorok, Serapit)} = \frac{1}{2} d^2 \left( y_i, y_j \right) \]

\[ = \frac{1}{2} \sum_{k=1}^{n} \left( y_{ik} - y_{jk} \right)^2 \]

\[ = \frac{1}{2} \left( 5.763588514 \right) \]

\[ = 2.881794 \]

Based on the results of the whole calculation, it is known \( SSE \) smallest is. \( SSE_{(Salapian, S. Seberang)} = 0.198 \)

The above process is carried out to count between the two clusters formed, with each cluster consisting of one object. Clustering method that starts from two or more closest objects into one cluster, then is done by calculating the distance of a cluster with a new object, this process is carried out in stages.

The cluster analysis process using the ward method produces 3 clusters that are formed, including: Cluster 1 shows A consisting of Bahorok, Kuala, Binjai, Wampu, Hinnai, Padang Tualang, Batang Serangan, Gebang, Pangkal Kusui and Besitang sub-districts. Cluster 2, for example B, consists of Serapit, Salapian, Kutambaru, Palm Overseas, Sei Lepan, West Brandan and Pematang Jaya districts. Cluster 3, for example C, consists of the districts of Sei Bingai, Done, Stabat, Secanggang, Tanjung Pura and Balaban.

**Cluster Formation 1**

For the \( SSE \) value the object in cluster 1 is \( A \) (\( SSE_A \)), then:

\[ SSE_A = \sum_{i=1}^{n_A} \left( y_i - \bar{y}_A \right)^2 \]

\[ = \left( y_1 - \bar{y}_A \right)^2 + \left( y_2 - \bar{y}_A \right)^2 + \left( y_3 - \bar{y}_A \right)^2 + \left( y_4 - \bar{y}_A \right)^2 + \left( y_5 - \bar{y}_A \right)^2 + \left( y_6 - \bar{y}_A \right)^2 \]

\[ = 0.032823944 + 0.057448908 + 0.088832153 + 0.170471656 \]

\[ + 0.27216902 + 0.497002414 + \ldots + 1.103911392 \]

\[ = 17.15217241 \]

So, value for \( SSE_A \) of 17.15217241 shows that the 10 districts have similar characteristics of health indicators.

The above process continues until the formation of cluster 2 and cluster 3 where the value of each cluster formation is:

- For cluster 2 is B (\( SSE_B \)) obtained a value of 27.98825302 indicating that of the 7 districts have similar characteristics of health indicators.
- For cluster 3 is C (\( SSE_C \)) obtained value of 41.53542429 shows that from 6 districts have similar characteristics based on health indicators.

Based on the results of cluster analysis it is known that cluster A has similar characteristics based on health indicators in Langkat District with the closest value among other clusters with a distance of 17.15217241. While the cluster C has the farthest distance between the other clusters 41.53542429.

The next step is to carry out the process of forming clusters A and B, A and C, B and C as shown in the calculation below:
Formation the cluster A and B
For SSE values the objects in clusters A and B are \( AB \) (\( SSE_{AB} \)). then:

\[
SSE_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \overline{y}_{AB}) (y_i - \overline{y}_{AB})
\]

\[
= \left( (y_1 - \overline{y}_{AB})^2 + (y_2 - \overline{y}_{AB})^2 + (y_3 - \overline{y}_{AB})^2 + (y_4 - \overline{y}_{AB})^2 + (y_5 - \overline{y}_{AB})^2 + (y_6 - \overline{y}_{AB})^2 + \ldots + (y_{144} - \overline{y}_{AB})^2 \right)
\]

\[
= 0.00968907 + 0.001348328 + 0.329994915 + 0.018626115 + 0.038345018
\]

\[
= 56.27292492
\]

So, value for \( SSE_{AB} \) of 56.27292492 shows that from 17 sub-districts have similar characteristics based on health indicators.

\[
I_{AB} = SSE_{AB} - (SSE_{A} + SSE_{B})
\]

\[
= 56.27292492 - (17.15217241 + 27.98825302)
\]

\[
= 11.13249948
\]

So, total value \( I_{AB} \) of 11.13249948 shows that from 17 districts have a maximum value.

The above process continues until the formation of clusters A and C and clusters B and C where the value of each cluster formation is:

- For clusters A and C, values are obtained \( SSE_{AC} \) of 74,58749046 shows that from 16 districts have similar characteristics based on health indicators. So, the total value for \( I_{AC} \) of 15.89989376 shows that of 16 districts have a maximum value.

- For clusters B and C obtained values \( SSE_{BC} \) of 114.7749548 shows that from 13 districts have similar characteristics based on health indicators. So, the total value for \( I_{BC} \) of 45.25127749 shows that of 13 districts have a maximum value.

Based on the results of the analysis above, it is known that the total distance that experiences the farthest value is in 13 districts with a value of 45.25127749 that has similar characteristics based on health indicators while the total distance that experiences the closest value with a distance of 11.13249948 occurs in 17 districts.

### Tabel 1. Profile of each Cluster

<table>
<thead>
<tr>
<th>Centroid value</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
<th>( X_5 )</th>
<th>( X_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>-0.023</td>
<td>0.230</td>
<td>0.388</td>
<td>-0.375</td>
<td>-0.296</td>
<td>-0.081</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>-1.033</td>
<td>-1.197</td>
<td>-1.217</td>
<td>-0.633</td>
<td>0.130</td>
<td>-0.215</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>2.709</td>
<td>2012</td>
<td>2.274</td>
<td>2.378</td>
<td>0.585</td>
<td>0.662</td>
</tr>
</tbody>
</table>

Based on Table 1, it can be concluded that the amount the cluster generated by using the ward method for grouping districts in Langkat Regency as many as 3 the cluster based on health status, i.e. the cluster with a high degree of health, the cluster with moderate health status, and the cluster with a low degree of health.

Cluster with a high degree of health found at the cluster 2, this can be seen in variables \( X_1, X_2, X_3 \) and \( X_4 \) with the lowest value though \( X_5 \) and \( X_6 \) is the second lowest level.

Cluster with a moderate degree of health at the cluster 1, this can be seen in variables \( X_1, X_2, X_3 \) and \( X_4 \) with the second lowest value as well \( X_6 \) with the lowest value.

Cluster with a low degree of health found in the cluster 3, this can be seen in variables \( X_1, X_2, X_3 \) and \( X_3 \) the tall ones as well are in the second level and \( X_5 \), and \( X_6 \) which is quite high.

### 4. CONCLUSION

From the results of the analysis the cluster it was found that there are 3 subdistrict clusters that have similar characteristics based on health indicators with 6 variables.
a. Cluster 1 consists of 10 districts, namely; Bahorok, Kuala, Binjai, Wampu, Hinai, Padang Tualang, Batang Serangan, Gebang, Pangkalan Susu, and Besitang. Value Result \( SSE \) of 17.15217241.

b. Cluster 2 consists of 7 districts, namely; Serapit, Salapian, Kutambaru, Sei Lepan, Seberang Seberang, West Brandan, and Pematang Jaya. Value Result \( SSE \) of 27.98825302.

c. Cluster 3 consists of 6 districts, namely; Sei Bingai, Done, Stabat, Secanggang, Tanjung Pura, and Babalan. Value Result \( SSE \) of 41.53542429.

Cluster analysis results that have similar characteristics based on health indicators with values the closest of the other clusters is in the first clustering by value \( SSE \) of 17.15217241. While the value that has similar characteristics based on health indicators with the \( SSE \) value farthest among other clusters is found in the third cluster with values \( SSE \) of 41.53542429.
REFERENCES


