**JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE**
VOLUME 1, No 2, October 2022
e-ISSN: 2828-9986
https://ejournal.unib.ac.id/index.php/jsds/index

# Achievement Cluster of Covid-19 Vaccination at the South Bengkulu Health Center Using Agglomerative Hierarchical Clustering

**Nurmaya Sumita[1], Devni Prima Sari[1*]**
[1] Departement of Mathematics, Universitas Negeri Padang, Indonesia

* Corresponding Author: devniprimasari@fmipa.unp.ac.id

**Article Info**

**Article History:**
*Received: October 6, 2022*
*Revised: October 20, 2022*
*Accepted: October 26, 2022*
*Available Online: October 31, 2022*

**Key Words:**
*Agglomerative Hierarchical Clustering*
*Euclidean*
*Covid-19 Vaccination*

**Abstract**

The concerns of many people and the lack of vaccine information are significant obstacles to achieving the Covid-19 vaccination target. The government and health groups must be ready to provide correct vaccine information to reduce public doubts. To evaluate the vaccine implementation, this is necessary to cluster the area regarding the achievement of the vaccination target. Clustering this area can be done using the Agglomerative Hierarchical Clustering method. In this study, clustering was carried out using Covid-19 vaccination data at the South Bengkulu Health Center involving six variables. Three clusters were formed for the clustering process: the first dose of Covid-19 vaccination, the second dose of Covid-19 vaccination, and the first Booster vaccination. Each cluster is represented by low, medium, and high clusters.

## 1. INTRODUCTION

The covid-19 Pandemic appeared first in the People's Republic of China, specifically Wuhan city, at the end of 2019, which caused a stir in people worldwide. The Pandemic resulted in many people being infected with the Covid-19 virus until it caused death. The corona virus that has shocked the world is a new type which was named Severe Acute Respiratory Syndrome Coronavirus2 (SARS-CoV-2) by the World Health Organization (WHO) and the name of the disease is called Coronavirus Disease 2019 (COVID-19) [1]. Thus, the government obliges the community to wear a mask when traveling, going out, and obeying suggestions Restrictions Social Scale Large (PSBB), Lockdown, Social Distancing, and PPKM. Besides, the Indonesian government also recommends that the Public do Covid-19 vaccination. It is done to cut off the chain of the spread of Covid-19 in Indonesia. The number of people's concerns and the lack of information about vaccines is significant obstacles to achieving the Covid-19 vaccination target [2]. The government and public groups must be ready to give information about the correct vaccine to reduce doubts in society. Delivery information could conduct through social media, distributed brochures, and outreach to the public through the device area. Temporarily, the Indonesian government has a target of as many as 208,265,720 residents to do vaccinated. Target dose first for power health, children, age 12-17, elderly, officer public, community vulnerable, and general reached 190,228,123 people. The covid-19 vaccination dose starts on January 13, 2021, with the target being health workers. Then for the target dose, the target second and dose third reached 142,270,154 people and 9,166,818 people.

Covid-19 vaccination in an area can be run through a clustering process to see the level of performance. One of them uses the agglomerative hierarchical clustering method. This method is familiarly used in several previous studies, including implementing Single Linkage and Complete Linkage for groups of districts/cities in East Java Province based on the Human Development Index (HDI) [3]. Next, Ward's method for a group performance of the Mathematics Department of Universitas Tanjungpura is based on the data of the evaluation results of student questionnaires [4] and the grouping of community welfare using Average Linkage [5].

To monitor the assessment of vaccine implementation, the authors are interested in knowing the level of performance of the Covid-19 vaccination, especially at the South Bengkulu Health Center, using the agglomerative hierarchical clustering method. The data used in this study is secondary data provided by the official website of the

Health Service of South Bengkulu. Namely, data on six categories of participants: public servants, health workers, ages 6-11 years, ages 12-17 years, vulnerable people and the general public, and the elderly.

### 1.1 Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering is a method of the data group. Method this by iterative group data by similarities between them for make hierarchy. Following this is steps on the Agglomerative Hierarchical Clustering [6]:

1. State every point as a cluster
2. Count distance proximity
3. Merge couple clusters with minimum distance
4. Produce clusters with cut dendrogram at the appropriate level

A clustering algorithm is built on proximity or similarity Among data objects. The measure used in this study is the Euclidean distance. Euclidean is a method for calculating the distance of two points to find out the relationship between angles and distances [7].

$$d_{jk} = \sqrt{\sum_{i=1}^{n} \left(x_{ij} - y_{ik}\right)^2} \tag{1}$$

where $j$ and $k$ are two objects, $n$ is the number of attributes available in the data object, $X_{ij}$ and $X_{ik}$ are the value of the $i$-th attribute of objects $j$ and $k$.

### 1.2 Function Linkage

Function linkage is a precondition necessary for the analysis of cluster hierarchy. Type function's most common linkage are [6]:

1. Single Linkage

Algorithm Single Linkage is the distance among a couple of clusters determined by two objects closest to different clusters. Single linkage clustering tends to produce elongated clusters, which causes a chain effect. As a result, two clusters with very other properties can be connected due to the presence of noise. However, in clusters far apart from each other the single linkage method works well.

$$D\left(D_l, \left(D_i, D_j\right)\right) = min\left(D(D_l, D_i), D(D_l, D_j)\right) \tag{2}$$

where $D(D_l, D_i)$ is the distance from the neighbor closest to the cluster $D_l$ and $D_i$, and $D\left(D_l, D_j\right)$ is the distance from the nearest neighbor of the cluster $D_l$ and $D_j$.

2. Complete Linkage

The complete linkage algorithm is different from the single linkage grouping. The complete linkage method component uses the furthest distance from a pair of objects to define the distance between clusters. This method is effective in revealing small and compact clusters. The distance measurement in the Complete Linkage method between two groups uses the maximal proximity formula as follows:

$$D\left(D_l, \left(D_i, D_j\right)\right) = max\left(D(D_l, D_i), D(D_l, D_j)\right) \tag{3}$$

where $D(D_l, D_i)$ is the distance from the neighbor closest to the cluster $D_l$ and $D_i$, and $D\left(D_l, D_j\right)$ is the distance from the nearest neighbor of the cluster $D_l$ and $D_j$.

3. Average Linkage

The Average Linkage Algorithm is the distance between two clusters with the average distance between all pairs of data points, each of which comes from a different group. The distance between the new and old clusters is the average distance of $D(D_l, D_i)$ and $D(D_l, D_i)$.

$$D(D_l, (D_i, D_j) = \frac{1}{2}\left(D(D_l, D_i) + D(D_l, D_j)\right) \tag{4}$$

where $D(D_l, D_i)$ is the distance from the neighbor closest to the cluster $D_l$ and $D_i$, and $D\left(D_l, D_j\right)$ is the distance from the nearest neighbor of the cluster $D_l$ and $D_j$.

### 1.3 Dendrogram

Agglomerative Hierarchical clustering or Divisive methods can be illustrated with a two dimensional diagram, known as a dendrogram. The dendrogram depicts the integration at each stage in the analysis [8]. The destination-making dendrogram is for looking for grouping objects with more ease and information. On the dendrogram axis vertical shows distance, whereas the horizontal axis shows object observation.

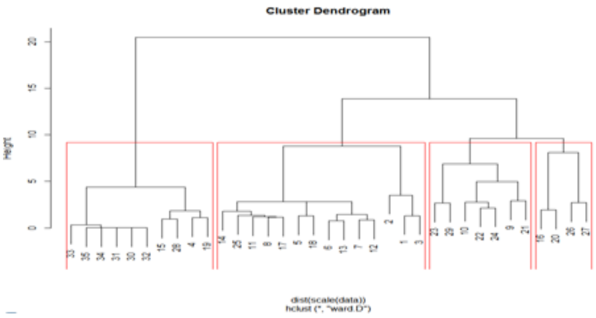



**Figure 1.** Example Dendrogram

### 1.4 Correlation Cophenetic

The cluster validity test used in this study is the cophenetic correlation. Cophenetic correlation is the correlation coefficient between the original elements of the dissimilarity matrix and the elements generated by the cophenetic matrix. According to Dani, et al [9] to calculate the cophenetic correlation coefficient, the following equation can be used:

$$r_{coph} = \frac{\sum_{i<j}^{n}(d_{ij} - \bar{d})(d_{coph\sim ij} - \bar{d}_{coph})}{\sqrt{\left[\sum_{i<j}^{n}(d_{ij} - \bar{d})^2\right]\left[\sum_{i<j}^{n}(d_{coph\sim ij} - \bar{d}_{coph})^2\right]}} \tag{5}$$

where

$r_{coph}$ : correlation coefficient cophenetic

$d_{ij}$ : original distance between $i$-th and $j$-th object

$\bar{d}$ : average of $d_{ij}$

$d_{coph\sim ij}$ : cophenetic distance between $i$-th and $j$-th object

$\bar{d}_{coph}$ : average of $d_{coph\sim ij}$

The value $r_{coph}$ has an interval of -1 to 1. If it is $r_{coph}$ close to 1, the resulting clustering process could say enough good.

## 2. METHOD

The type of research used is applied research using secondary data. In this study, we obtained data from the official website of the Bengkulu Selatan Health Service. The data obtained is for the Covid-19 vaccination at 14 South Bengkulu Health Centers. In a study, this variable free consists of six variables officer health, civil servant, age 6-11 years, age 12-17 years, advanced age, society vulnerable, and the general public. The analytical technique used in a study are :

1. Collect existing data at the Health Office in South Bengkulu Regency.
2. Complete the missing data and complete the blank data.
3. Perform statistical analysis
   a. Calculate the distance using the Euclidean distance method to determine the similarity between objects.
   b. Conduct cluster analysis using the Agglomerative Hierarchical Clustering method with single linkage based on distance variations in the Covid-19 Vaccination data at the South Bengkulu Health Center.
   c. Perform dendrogram cuts
   d. Validate the cluster
   e. Interpreting clustering results.

## 3. RESULTS AND DISCUSSION

Processed data consist of vaccine 1, vaccine 2, and vaccine Boosters first to have six variables: officer health, civil servant, age 6-11 years, age 12-17 years, advanced age, society vulnerable, and the general public. The data in the table will be processed using Agglomerative Hierarchical Clustering Single Linkage with variation distance, that is, Euclidean distance—clustering data processing process conducted with the help RStudio.

1. To do measurement distance similarity with Euclidean Distance:

Counting distance Euclidean with use equation (1). Following this explanation, the calculation of the distance between object one and object two on the vaccine first uses the Euclidean method:

$$d(1,2) = \sqrt{\begin{array}{c}(520-416)^2+(33-75)^2+(494-572)^2\\+(424-788)^2+(449-903)^2+(2464-4390)^2\end{array}} = 654.7757$$

The calculation of the proximity between objects produces a Euclidean distance of 654.7757. For the next object distance measurement is identical to the calculation of objects 1 and 2.Then, the distance between object 1 and object 2 in the second vaccine is calculated as follows:

$$d(1,2) = \sqrt{\begin{array}{c}(373-459)^2+(32-264)^2+(436-836)^2\\+(276-561)^2+(362-909)^2+(1504-3296)^2\end{array}} = 848.4949$$

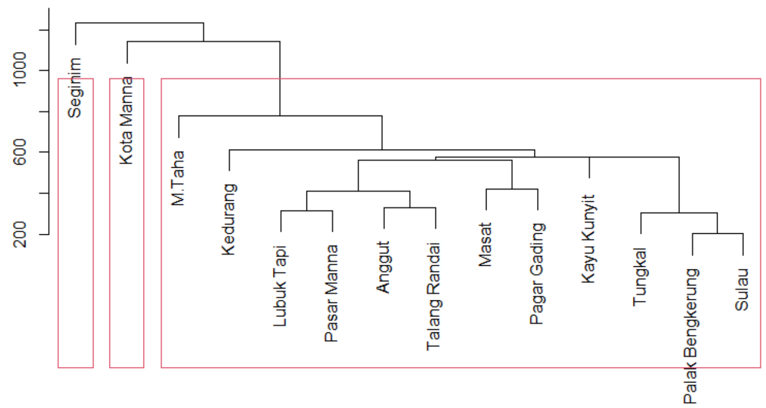
The calculation of the proximity between objects in the second vaccine resulted in a Euclidean distance of 848. 4949. Then, the distance between objects in the booster vaccine was calculated as follows:

$$d(1,2) = \sqrt{\begin{array}{c}(145-155)^2+(25-341)^2+(0-0)^2\\+(166-135)^2+(17-11)^2+(1360-1182)^2\end{array}} = 348.05689$$

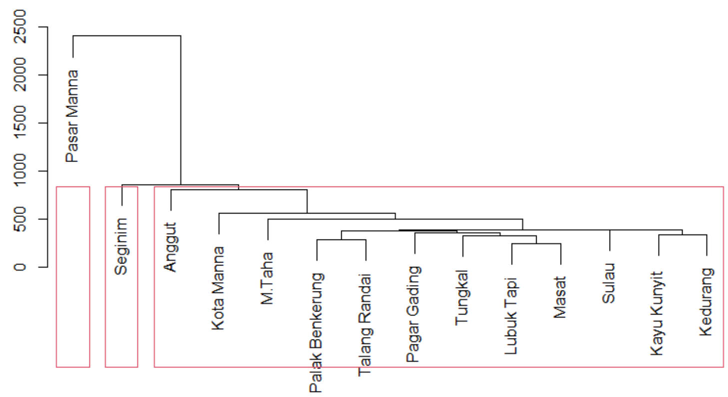
The calculation of proximity between objects in the booster vaccine produces a Euclidean distance of 348, 05689. For measures between other objects, the calculation process is carried out in the same way as the calculation process for object 1 and object 2.

## 2. Agglomerative Hierarchical Clustering Method

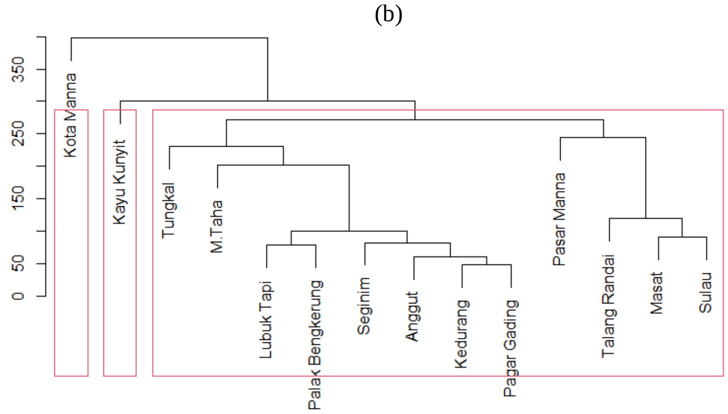
The linkage function used in the Agglomerative Hierarchical Clustering Method in this study is Single Linkage. This linkage is based on bottom-up clustering (agglomeration grouping), at each step combining two clusters containing the closest pair of elements that are not yet included in the same cluster with each other [10]. After measuring the distance, the next step is to cluster data on vaccine 1, vaccine 2, and vaccine Booster using Single Linkage. The dendrogram of vaccination data clustering results at the South Bengkulu Health Center using the Agglomerative Hierarchical Clustering method can be seen in Figure 2.




(a)




(b)




(c)

Figure 2. Dendrogram (a). Vaccine 1, (b). Vaccine 2, and ( c). Vaccine Boosters

Figure 2. shows the results of dendrogram cutting for vaccine 1, vaccine 2 and booster vaccine. Dendrogram cutting using Single Linkage produces 3 clusters

3. Validity Cluster

After getting a solution from the results of grouping Covid-19 vaccination data at the Health Center in South Bengkulu using the Agglomerative Hierarchical Clustering (Single Linkage) method, the next step is to test the validity of the cluster by calculating the cophenetic correlation coefficient. Cophenetic correlation is the correlation coefficient between the elements of the original dissimilarity matrix and the elements produced by the cophenetic matrix. The cophenetic correlation coefficient was calculated using equation (5). In Table 1, you can see the output of the calculation of the cophenetic correlation coefficient using the R Studio software.

**Table 1.** Cophenetic Correlation Coefficient

| Type of Vaccine | Correlation Cophenetic Coefficient |
| --- | --- |
| Coefficient correlation cophenetic 1st vaccine | 0.8104235 |
| Coefficient correlation cophenetic 2nd vaccine | 0.8853514 |
| Coefficient correlation cophenetic booster vaccine | 0.7911208 |

Based on Table 1, it can be concluded that in clustering the second dose of Covid-19 vaccination with Euclidean distance, the most optimal value (almost close to 1) is 0.8853514.

## 4. CONCLUSION

Three clusters were formed for the second dose of Covid-19 vaccination. The first cluster consisted of Anggut, Kayu Kunyit, Kedurang, Manna City, Lubuk Tapi, M. Taha, Masat, Pagar Gading, Palak Benkerung, Sulau, Talang Randai, and Tungkal Health Center. Then, Cluster 2 consists of Pasar Manna Health Center. Finally, the third cluster includes the Seginim Health Center. Based on the average vaccine achievement from the three clusters formed, we can interpret that the first cluster is a Health Center with low attainment of the second vaccination dose. The second cluster is a Health Center with a moderate achievement level of the second vaccination dose. The third cluster is the Health Centers s with high attainment of the second vaccination dose.


## ACKNOWLEDGMENTS

I would like to thank the District Health Office of South Bengkulu for supporting the data for my research.


## REFERENCES


[1] Y. Yuliana, "Corona virus diseases (Covid-19): Sebuah tinjauan literatur," *Wellness and Healthy Magazine*, vol. 2, no. 1, pp. 187–192, Mar. 2020, doi: 10.30604/WELL.95212020.

[2] G. H. Lushington, "Perspective on the COVID-19 Coronavirus Outbreak," *Comb Chem High Throughput Screen*, vol. 23, no. 2, pp. 90–91, Apr. 2020, doi: 10.2174/138620732302200406130010.

[3] Y. N. Indriani, "Perbandingan Jumlah Kelompok Optimal pada Metode Single Linkage dan Complete Linkage dengan Indeks Validitas Silhouette: Studi Kasus pada Data Pembangunan Manusia Jawa Timur," Universitas Negeri Malang, Malang, 2013.

[4] S. Oktavia, N. Mara, and N. S. Intisari, "Pengelompokan Kinerja Dosen Jurusan Matematika FMIPA UNTAN Berdasarkan Penilaian Mahasiswa Menggunakan Metode Ward," *Bimaster : Buletin Ilmiah Matematika, Statistika dan Terapannya*, vol. 2, no. 02, pp. 93–100, Aug. 2013, doi: 10.26418/BBIMST.V2I02.3026.

[5] S. Yulianto and K. H. Hidayatullah, "Analisis Klaster untuk Pengelompokan Kabupaten/Kota di Provinsi Jawa Tengah berdasarkan Indikator Kesejahteraan Rakyat," *Jurnal Statistika*, vol. 2, no. 1, 2014, doi: 10.26714/JSUNIMUS.2.1.2014.

[6] Rui. Xu and D. C. Wunsch, *Clustering*. Wiley-IEEE Press, 2008. Accessed: Oct. 20, 2022. [Online]. Available: https://www.wiley.com/en-us/Clustering-p-9780470276808

[7] Derisma, Firdaus, and Rio Putra Yusya, "Perancangan Ikat Pinggang Elektronik untuk Tunanetra menggunakan Mikrokontroller dan Global Positioning System (GPS) pada Smartphone Android," *Jurnal Teknik Elektro ITP*, vol. 5, no. 2, 2016.

[8] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl, *Cluster Analysis*, 5th ed. John Wiley & Sons, 2011. Accessed: Oct. 20, 2022. [Online]. Available: https://www.wiley.com/en-us/Cluster+Analysis%2C+5th+Edition-p-9780470749913

[9] A. Tri, R. Dani, S. Wahyuningsih, and N. A. Rizki, "Penerapan Hierarchical Clustering Metode Agglomerative pada Data Runtun Waktu," *Jambura Journal of Mathematics*, vol. 1, no. 2, pp. 64–78, Jul. 2019, doi: 10.34312/JJOM.V1I2.2354.

[10] V. Vijaya, S. Sharma, and N. Batra, "Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering," *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, pp. 568–573, Feb. 2019, doi: 10.1109/COMITCON.2019.8862232.