BAREKENG: Journal of Mathematics and Its Applications
June 2023 Volume 17 Issue 2 Page 0707–0716
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss2pp0707-0716

K-MEANS CLUSTER COUNT OPTIMIZATION WITH SILHOUETTE INDEX VALIDATION AND DAVIES BOULDIN INDEX (CASE STUDY: COVERAGE OF PREGNANT WOMEN, CHILDBIRTH, AND POSTPARTUM HEALTH SERVICES IN INDONESIA IN 2020)

Iut Tri Utami[1], Fahlevi Suryaningrum[2*], Dwi Ispriyanti[3]

[1,2,3] *Department of Statistics, Faculty of Science and Mathematics, Diponegoro University*
Prof. Soedarto Street, Semarang, 50275, Indonesia

*Corresponding author's e-mail: * fahlevi.surya@gmail.com*

ABSTRACT

Article History:

Received: 23rd November 2022
Revised: 8th April 2023
Accepted: 12th April 2023

Keywords:

Davies Bouldin Index;
K-Means;
Maternal Health Care;
Silhouette Index.

One of the causes of the increasing maternal mortality rate in Indonesia is the declining performance of maternal health services in each Indonesian province. To overcome the decline in performance, namely by determining in advance the provinces that need to be prioritized for services by grouping 34 provinces in Indonesia. This study aims to obtain the best provincial grouping results so that it can prioritize the right provinces. One of the methods that are suitable for grouping provinces is K-Means because it is simple and easy to implement. The disadvantage of K-Means is that it is sensitive to determining the right number of initial clusters, so Silhouette Index and Davies Bouldin Index validation is used to obtain the optimal number of clusters with stable and consistent results. This study used healthcare data for pregnant women, childbirth, and postpartum, with K=2, 3, and 4 as the initial cluster number. K-Means objects are grouped in similarities using Euclidean and Manhattan distances. The result obtained was the optimal number of clusters with K=2 using Manhattan, where the highest Silhouette Index value was 0,658685 and the lowest Davies Bouldin Index was 0,3561214, which met the criteria for determining the optimal cluster.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.

How to cite this article:

I. T. Utami, F. Suryaningrum, and D. Ispriyanti, "*K-MEANS* CLUSTER COUNT OPTIMIZATION WITH *SILHOUETTE INDEX* VALIDATION AND *DAVIES BOULDIN INDEX* (CASE STUDY: COVERAGE OF PREGNANT WOMEN, CHILDBIRTH, AND POSTPARTUM HEALTH SERVICES IN INDONESIA IN 2020)", *BAREKENG: J. Math. & App.*, vol. 17, iss. 2, pp. 0707-0716, June, 2023.

Copyright © 2023 Author(s)
Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

The Sustainable Development Goals (SDGs) are a global action plan agreed upon by world leaders, including Indonesia, to end poverty, reduce inequality and protect the environment. One of the SDGs targets in the health sector that needs to be achieved is to improve the degree of public health as indicated by the decrease in the Maternal Mortality Rate (MMR) [3]. In Indonesia, MMR continues to increase every year, and one of the contributing factors can be seen from the decrease in the percentage of health service performance of pregnant women, childbirth, and postpartum in Indonesian provinces [11]. To overcome this decline in performance, namely by determining in advance the provinces that need to prioritize services by grouping 34 provinces in Indonesia. This study aims to obtain the best provincial grouping results so that they can prioritize the right provinces. One method that is suitable for grouping provinces is cluster analysis, and then the data used is data on health services for pregnant women, childbirth, and postpartum in Indonesia in 2020.

Cluster analysis is a suitable method because it can find out which provinces are high or low clusters in the health services of pregnant women, childbirth, and postpartum by identifying characteristics between 34 provinces in Indonesia. Cluster analysis is divided into two, namely hierarchical and non-hierarchical methods. One of the non-hierarchical methods is *K-Means*. The hierarchy method (agglomerative and divisive) is inefficient and the calculation process is longer if it is used to group large amounts of data compared to *K-Means* [5], so the cluster analysis method that will be used in this study is the *K-Means* method.

K-Means is a non-hierarchical method that can group *n* objects into *k* clusters that have the same characteristics and can be used on numerical data and include simple methods. The disadvantage of *K-Means* is that it is sensitive to determining the most appropriate number of initial *k* clusters because it is generally done randomly, will result in different data groupings, and does not always provide accurate results [8]. The exact and optimal number of *k* clusters can be determined using validation. These validations include the *Silhouette Index* and the *Davies Bouldin Index*. Both validations can see the optimal number of clusters with stable and consistent results. The determination of the number of initial *K* clusters in this study was 2, 3, and 4 by looking for the highest *Silhouette Index* value and the lowest *Davies Bouldin Index* value.

Objects in *K-Means* are grouped by their similarity. Distance measurement plays an important role because it can determine the degree of similarity of data. To measure the degree of similarity, the *Euclidean* and *Manhattan* distances are used. *Euclidean* distances are used very often, but based on research shows that *Manhattan* is better than *Euclidean* in clustering [17]. Therefore, in this study, the *Manhattan* distance was used as a comparison of the two distances.

2. RESEARCH METHODS

Maternal health services are a health effort that concerns the service and maintenance of pregnant women, maternity mothers, and breastfeeding mothers [12]. Pregnant women's health services that have been implemented in Indonesia are antenatal visit services which are pregnancy checks with health workers, giving blood-added tablets to prevent anemia, classes of pregnant women carried out at local government clinics, and providing additional food to pregnant women with chronic lack of energy aimed at overcoming malnutrition. Meanwhile, maternity services are childbirth efforts that are helped by trained health workers and carried out in healthcare facilities. In addition, health services for postpartum mothers, one of which is the provision of vitamin A supplements as early prevention of vitamin A deficiency.

Cluster analysis can group *n* objects based on *p* variables that have relatively similar characteristics among these objects so that the diversity within a cluster is smaller than the diversity between clusters [9]. Cluster analysis can be used in ordinal, interval, and ratio data scales. Cluster analysis is used as a data summarizer by grouping objects based on the similarity of certain characteristics of the object to be studied, which means it is not used to connect or distinguish with samples or other variables. The assumption before conducting cluster analysis is twofold, namely that the sample represents population and multicollinearity [7].

a. Sample Representing Population

Testing of samples representing the population can be done by looking at the degree of adequacy of a sample using the *Kaiser Meyer Olkin* (KMO) test. Test the KMO hypothesis [2]:

Hypothesis

H_0: Sample represents a population

H_1: The sample is not representative of the population

Test Statistics

$$the\ KMO = \frac{\sum_{j=1}^{p}\sum_{l=1}^{p} r_{jl}^2}{\sum_{j=1}^{p}\sum_{l=1}^{p} r_{jl}^2 + \sum_{j=1}^{p}\sum_{l=1}^{p} a_{jl,m}^2} \quad (1)$$

where j = 1,2,3, ..., p and l = 1,2,3, ..., p, for $j \neq l$; r_{jl} : Pearson correlation coefficient between variables j and l; and a_{jl}: partial correlation coefficient between variables j and l by keeping variable m constant.

Test Criteria

A sample is said to be representative of a population of a KMO value greater than 0,5 is obtained.

b. Multicollinearity Test

Multicollinearity is the possibility of a relationship or correlation in a variable. One way of identifying the existence of multicollinearity is to calculate the value of the *Variance Inflation Factor* (VIF) formulated in **Equation 2 [4]**:

$$VIF = \frac{1}{(1 - R^2)} \quad (2)$$

where R^2 is the coefficient of determination of the dependent variable with the independent variable. If the VIF value < 10, then there is no multicollinearity.

Cluster analysis is used to group the similarity of an object in the same cluster, therefore it takes some measure of distance to find out how similar the objects are. For this study, the distance measure used was *Euclidean* distance and *Manhattan* distances.

a. *Euclidean* Distance

The *Euclidean* distance is the root of the sum of the squares of the difference between objects. Formula equation for calculating *Euclidean* distance in **Equation 3 [16]**:

$$d_{euc}(x_i, C_k) = \sqrt{\sum_{j=1}^{p}(x_{ij} - C_{kj})^2}\ ,\ j = 1,2,3,\ldots,p \quad K=2,3,4 \quad (3)$$

where $d_{euc}(x_i, C_k)$ is the *Euclidean* distance between the i-th object, the j-th variable to the center of the cluster (centroid) k-th on the j-th variable; k = 1.2, ... , K; x_{ij} is the value of the i-th object on the j-th variable; C_{kj} is the center of the k-th centroid on the j-th variable; p is the number of observed variables; and K is the number of clusters.

b. *Manhattan* Distance

Manhattan distance is the sum of the absolute difference for each object. *Manhattan* distance is expressed in **Equation 4 [18]**:

$$d_{man}(x_i, C_k) = \sum_{j=1}^{p}\left|x_{ij} - C_{kj}\right|\ ,\ j = 1,2,3,\ldots,p,\ K=2,3,4 \quad (4)$$

where $d_{man}(x_i, C_k)$ is the *Manhattan* distance between the i-th object, the j-th variable to the center of the cluster (centroid) k-th on the j-th variable; k= 1.2, ... , K; x_{ij} is the value of the i-th object on the j-th variable; C_{kj} is the center of the k-th centroid on the j-th variable; p is the number of observed variables; and K is the number of clusters.

K-Means is a non-hierarchical clustering method that seeks to partition data into one or more clusters so that data with the same characteristics is grouped into the same cluster and data with different characteristics is grouped into other clusters. The steps of *K-Means* are [6]:

a. Determining the number of *K*-clusters to be formed;
b. Randomly determine the initial cluster center (centroid);
c. Calculate the distance of each object with each centroid;
d. Grouping each object into the closest centroid, an object will become a member of the *k*-th cluster if the distance of that object to the *k*-th centroid is of the least value when compared to the distance to other centroids;
e. Determine the new centroid by calculating the average of the objects on each cluster with Equation (5):
$$C_{kj} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ij} \tag{5}$$
with $k = 1,2,3,\cdots,K$; $j = 1,2,3,\cdots,p$; C_{kj} is the centroid of the *k*-th cluster of the *j*-th variable; n_k is the number of objects on the *k*-th cluster; and x_{ij} is the value on the *i*-th object on the *j*-th variable;
f. Repeat steps c through e until none of the members of each cluster have changed.

After clustering data into a number of clusters with *K-Means*, a validation process is needed on the cluster. Validation on the cluster is carried out to evaluate the cluster formed by giving it a validity value. This study will be used two validations to determine the optimal number of clusters in *K-Means*, namely by validating the *Silhouette Index* and the *Davies Bouldin Index*.

a. *Silhouette Index* Validation

The *Silhouette* coefficient is formulated in Equation (6):
$$SC = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{6}$$
with $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$, $b(i) = \min d(i,v)$, $a(i) = \frac{1}{n_k - 1}\sum_{h\in Cl_k, h\neq i} d(i,h)$

$b(i)$: Minimum value of the average distance of object *i* with all objects on the other cluster to *v*-th
$a(i)$: average *i*-th object distance with all objects in a cluster

The best grouping is achieved if maximum SC means minimizing the distance in the group ($a(i)$) while maximizing the distance between groups ($b(i)$), the greater the value of the *silhouette* coefficient, the better the quality of a group [13].

b. *Davies Bouldin Index* Validation

Validation of the *Davies Bouldin Index* formulated in Equation 7 [14]:
$$DBI = \frac{1}{K}\sum_{k=1}^{K} R_k \tag{7}$$
with, $R_k = max_{k\neq v}\left(\frac{S_k + S_v}{M_{k,v}}\right)$, $M_{k,v} = d(C_k, C_v)$, $k \neq v$, $S_k = \frac{1}{n_k}\sum_{i=1}^{n_k} d(x_i, C_k)$, and $S_v = \frac{1}{n_v}\sum_{i=1}^{n_v} d(x_i, C_v)$, $k \neq v$

S_k : average of the *i*-th object distances with *k*-th centroid cluster
S_v : average of the *i*-th object distances with *v*-th centroid cluster
$M_{k,v}$: *k*-th cluster centroid distance and *v*-th cluster centroid distance

The smaller the *Davies Bouldin Index* (DBI) value obtained (non-negative ≥ 0), the better the cluster obtained [1].

The last step in grouping provinces in Indonesia based on health services for pregnant women, childbirth, and postpartum, is to interpret or profile the optimal number of clusters. Cluster profiling is used to see the average value of the members of each variable in each cluster, which will then obtain the characteristics of each cluster [15]. Cluster profiling is the stage of interpretation of each cluster that is formed to provide information as an illustration of the nature of the cluster and explain how each cluster can be relevant in each cluster [10].

The type of data used in this study is secondary data obtained from the Indonesian Health Profile in 2020. The data consists of the coverage of health services for pregnant women, childbirth, and postpartum based on 34 provinces in Indonesia in 2020. The research variables used were the percentage of antenatal visits four times (K4) services for pregnant women (X1), the percentage of giving blood-added tablets to pregnant women (X2), the percentage of local government clinics carrying out classes for pregnant women (X3), the percentage of supplementary feeding for pregnant women with Chronic Energy Deficiency (CED) (X4), the percentage of maternity services assisted by trained health workers (X5), the percentage of postpartum maternal health services received vitamin A (X6).

This research was carried out data processing using R software, then the stages of data analysis are:

a. Inputting data on health services for pregnant women, childbirth, and postpartum;
b. Performing a sample assumption test representing a population with *Kaiser Meyer Olkin* (KMO);
c. Conducting a multicollinearity assumption test, with a *Variance Inflation Factor* (VIF) value, if multicollinearity occurs in one of the variables, the main component analysis is carried out, the main component score obtained will be used as input in the next analysis as a substitute for the initial variable data value;
d. Determining the number of clusters (*K*), the *K* values used are *K*=2,3, and 4;
e. Conducting analysis of the *K-Means* method using the *Euclidean* and *Manhattan* distances;
    1) Randomly determining the initial cluster center (centroid);
    2) Calculating the distance of each object with each centroid with *Euclidean* and *Manhattan* distances;
    3) Group each object into the closest centroid;
    4) Defining a new centroid by calculating the average of objects on each cluster;
    5) Repeating steps 1 through 4 until none of the members of each cluster have changed.
f. Calculating the *Silhouette* coefficient value of *K*=2,3, and 4 with *Euclidean* and *Manhattan* distances;
    1) Calculating the average distance of the *i*-th object with all objects in a cluster;
    2) Calculating the average distance of the *i*-th object with all objects on other clusters;
    3) Determining the minimum value of the average distance of the *i*-th object with all objects that are on other clusters;
    4) Calculating *Silhouette* values;
    5) Calculating the *Silhouette* coefficient defined as the average of the *Silhouette* values.
g. Calculating the value of the *Davies Bouldin Index* coefficient from *K*=2,3, and 4 with the *Euclidean* distance and the *Manhattan* distance;
    1.) Calculating the average distance of objects with a centroid of the followed cluster;
    2.) Calculating the centroid distance in a cluster with centroids in another cluster;
    3.) Calculating the ratio to find out the comparison value of the *k*-th and *v*-th clusters;
    4.) Calculating the maximum value of the ratio between clusters;
    5.) Calculating the value of *Davies Bouldin Index.*
h. Evaluating the optimal number of clusters based on the *Silhouette* and *Davies Bouldin Index* coefficient values with *Euclidean* and *Manhattan* distances at *K*=2,3, and 4. The highest *Silhouette* coefficient value and the lowest *Davies Bouldin Index* value will be selected as the optimal number of clusters;
i. Analyzing optimal cluster results and profiling and interpretation of the regional characteristics of each cluster formed from the best groupings.

3. RESULTS AND DISCUSSION

The test results of cluster analysis assumptions based on R software processing are:

a. Sample Representing Population

In this study, the KMO test was not carried out because the data was in the form of a population of health services for pregnant women, childbirth, and postpartum in 34 provinces of Indonesia, so it can be concluded that the data has represented the existing population and the analysis can be continued.

b. Multicollinearity Test

Based on testing the assumption of multicollinearity using R software, it was obtained that the VIF value on the overall variables used in the study was less than 10. The value indicates that each variable does not have multicollinearity. After it is known that the cluster analysis assumptions are met, then further processing is carried out using the *K-Means* method. The results of the final iteration of the *K-Means* method in **Table 1**:

Table 1. ***K-Means* Grouping Results**

The i-th Object	K=2		K=3		K=4	
	Euclidean	***Manhattan***	***Euclidean***	***Manhattan***	***Euclidean***	***Manhattan***
1	1	1	3	1	2	4
2	1	1	3	1	2	4
3	1	1	3	1	2	4
4	1	1	3	3	2	2
5	1	1	1	1	1	4
6	1	1	1	1	1	4
7	1	1	1	1	1	4
8	1	1	1	1	1	4
9	1	1	1	1	1	4
10	1	1	1	1	1	4
11	1	1	1	1	3	4
12	1	1	1	1	3	4
13	1	1	1	1	3	4
14	1	1	3	1	1	4
15	1	1	1	1	3	4
16	1	1	1	1	1	4
17	1	1	1	1	3	4
18	1	1	1	1	1	4
19	2	1	3	3	2	2
20	1	1	3	1	1	4
21	1	1	3	1	2	4
22	1	1	3	1	1	4
23	2	1	3	3	2	2
24	1	1	1	1	1	4
25	1	1	1	1	1	4
26	1	1	3	1	1	4
27	1	1	1	1	1	4
28	1	1	3	1	2	4
29	1	1	3	1	2	4
30	1	1	3	1	2	4
31	2	1	3	3	2	3
32	1	1	3	1	2	4
33	2	2	2	2	4	1
34	2	2	2	2	4	1

Based on Table 1, a grouping of 34 objects was obtained for the number of clusters *K*=2 with *Euclidean*, members of Cluster 1 are 29 objects and Cluster 2 are 5 objects, while with *Manhattan* obtained members of Cluster 1 are 32 objects and Cluster 2 are 2 objects. For *K*=3 with *Euclidean*, members of Cluster 1 are 16 objects, Cluster 2 are 2 objects, and Cluster 3 are 16 objects. Whereas with *Manhattan* in Cluster 1 consists of 28 objects, Cluster 2 are 2 objects, and Cluster 3 are 4 objects. For *K*=4 with *Euclidean*, Cluster 1 has 15 objects, Cluster 2 has 12 objects, Cluster 3 consists of 5 objects, and Cluster 4 consists of 2 objects. Meanwhile, with *Manhattan*, the members of Cluster 1 are 2 objects, Cluster 2 are 3 objects, Cluster 3 are 1 object, and Cluster 4 are 28 objects.

Based on the results of the *K-Means* clustering method using *Euclidean* and *Manhattan* distances for *K*=2, 3, and 4, an evaluation was then carried out based on the validation of the *Silhouette Index* and *Davies Bouldin Index* to determine the most optimal number of clusters. The greater the value of the *silhouette* coefficient, the better the quality of a group, and the smaller the value of the *Davies Bouldin Index* (DBI) obtained (non-negative ≥ 0), the better the cluster obtained. The results of the *K-Means* clustering analysis described in Table 2:

Table 2. *K-Means* Clustering Analysis Results

Number of Clusters	Distance Measures	Validation	
		Silhouette Index	***Davies Bouldin Index***
1	*Euclidean*	0,5756249	0,7514368
2	*Manhattan*	**0,6586850**	**0,3561214**
3	*Euclidean*	0,2953258	1,0498710
4	*Manhattan*	0,4941210	0,7938820
5	*Euclidean*	0,3328544	0,9261220
6	*Manhattan*	0,4586676	0,7307606

Based on Table 2, it can be seen that in the grouping of 34 provinces in Indonesia using the *K-Means* method, the highest value of the *Silhouette Index* was obtained, namely 0,658685, and the *Davies Bouldin Index* obtained the lowest value of 0,3561214. The *Silhouette Index* and *Davies Bouldin Index* values have the same results, so from the results of the *K-Means* analysis, it is concluded that the optimal number of clusters is at K=2 using the *Manhattan* distance.

After the optimal number of clusters is known, the last step in grouping provinces in Indonesia based on health services for pregnant women, childbirth, and postpartum is to interpret or profile the optimal number of clusters. Based on the evaluation of the optimal number of clusters using two validations, the optimal number of clusters was obtained, namely *K*=2 using the *Manhattan* distance. The method gives the result that Cluster 1 consists of 32 provinces and Cluster 2 consists of 2 provinces. The members of each cluster formed are in Table 3:

Table 3. Members of the *K-Means* 2 *Manhattan* Distance Cluster Grouping

Clusters	Cluster Members	Sum
1	Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, DKI Jakarta, West Java, Central Java, DI Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku	32
2	West Papua, Papua	2

The profiling stage will see the characteristics of each cluster formed so that the tendency of each cluster can be seen. The characteristics of the clusters formed in the *K-Means* method can be represented by looking at the average of the members of each of the variables used in the study. The average of each variable in the cluster formed based on the health services of pregnant women, childbirth, and postpartum in Table 4:

Table 4. Average Variables in Pregnant Women, Childbirth, and Postpartum Health Services

Variable	Cluster	
	1	2
Antenatal Visits Four Times (K4) Services for Pregnant Women	80,300	31,000
Administration of Blood Add Tablets to Pregnant Women	80,856	27,550
Local Government Clinic Conducts Classes for Pregnant Women	81,531	7,1500
Supplementary Feeding for Pregnant Women with Chronic Energy Deficiency	94,075	69,650
Maternity Services Assisted by Trained Health Workers	81,234	39,650
Postpartum Maternal Health Services Get Vitamin A	84,622	35,450

Table 4 shows that the average cluster with the highest health services for pregnant women, childbirth, and postpartum is in Cluster 1. Cluster 1 means that the provinces in Cluster 1 have a very good quality of health services for pregnant women, childbirth, and postpartum compared to Cluster 2. Cluster 2 is seen to have a smaller cluster average than Cluster 1. This means that Cluster 2 members are provinces with low quality of health services for pregnant women, childbirth, and postpartum in Indonesia. Based on this, it can be interpreted that West Papua and Papua are provinces that must be paid more attention to by the government in Indonesia because they have low health services for pregnant women, childbirth, and postpartum, especially in the service of antenatal visits four times (K4), giving blood-added tablets, local government clinic carrying out classes for pregnant women, childbirth services, and services for giving vitamin A supplements to postpartum mothers.

4. CONCLUSIONS

The conclusions obtained based on the results of the analysis and discussion that have been carried out are:

1. The results of the grouping of *K-Means* methods from 34 provinces are:
   a. For *K*=2 with *Euclidean* distance, the number of members of Cluster 1 is 29 provinces and Cluster 2 is 5 provinces, while with *Manhattan* in Cluster 1 it consists of 32 provinces and Cluster 2 is 2 provinces;
   b. For *K*=3 with *Euclidean* distance, the number of members of Cluster 1 is 16 provinces, Cluster 2 is 2 provinces, and Cluster 3 is 16 provinces. Whereas with *Manhattan* in Cluster 1 consists of 28 provinces, Cluster 2 is 2 provinces and Cluster 3 is 4 provinces;
   c. For *K*=4 with *Euclidean* distance, Cluster 1 is obtained as many as 15 provinces, Cluster 2 is 12 provinces, Cluster 3 consists of 5 provinces, and Cluster 4 consists of 2 provinces. Meanwhile, with *Manhattan*, the number of members of Cluster 1 is 2 provinces, Cluster 2 is 3 provinces, Cluster 3 is 1 province, and Cluster 4 is 28 provinces.
2. The results of the grouping of 34 provinces in Indonesia using the *K-Means* method obtained the optimal number of clusters at *K*=2 with a size of the *Manhattan* distance. This can be seen in the validation results with the *Silhouette Index* of 0,658685 which is the highest value and the *Davies Bouldin Index* obtained a value of 0,3561214 which is the lowest value. It was obtained that Cluster 1 consisted of 32 provinces and Cluster 2 consisted of 2 provinces. Based on this grouping, it was found that the measurement distance used would affect the cluster results obtained;
3. The profiling results show that the highest cluster average is in Cluster 1, which means that the members of Cluster 1 are provinces with high quality of maternal health services. Cluster 2 has a lower cluster average than Cluster 1, which means that Cluster 2 members are provinces with low maternal health services. It is hoped that the government in Indonesia will pay more attention to the provinces in Cluster 2, namely West Papua and Papua, which are a collection of provinces with low average health services for pregnant women, childbirth, and postpartum compared to Cluster 1, so that the province can improve the quality of maternal health services and can reduce maternal mortality in the coming year.

REFERENCES

[1] A. Bates, *and* J. Kalita, *Counting Clusters in Twitter Posts*. Proceedings of the 2nd International Conference on Information Technology for Competitive Strategies, 2016.

[2] A. Widarjono, *Analisis Statistika Multivariat Terapan Edisi Pertama.* Yogyakarta: UPP STIM YKPN, 2010.

[3] Badan Perencanaan Pembangunan Nasional, “Tujuan Pembangunan Berkelanjutan Sustainable Development Goals Kehidupan Sehat dan Sejahtera”, 2022, https://sdgs.bappenas.go.id/tujuan-3/ [Accessed: 15 January 2022]

[4] D. Gujarati, *Dasar-dasar Ekonometrika Jilid 2.* Jakarta: Erlangga, 2009.

[5] D. Rachmatin, *and* K. Sawitri, “Perbandingan antara Metode Agglomeratif, Metode Divisif dan Metode K-Means dalam Analisis Klaster”, 2019, http://eprints.itenas.ac.id/157/ [Accessed: 20 February 2022].

[6] D. T. Larose, *and* C. D. Larose, *Discovering Knowledge in Data An Introduction to Data Mining Second Edition Wiley Series on Methods and Applications in Data Mining.* New Jersey: John Wiley and Sons, Inc, 2014.

[7] E. Hair, T. Halle, E. Terry-Humen, B. Lavelle, *and* J. Calkins, “Children's School Readiness in the ECLS-K: Predictions to Academic, Health, and Social Outcomes in First Grade”. *Early Childhood Research Quarterly*, vol. 21, no. 4, pp. 431-454, 2006.

[8] E. U. Wahyuningtyas, R. R. Putri, *and* Sutrisno, “Optimasi K-Means untuk Clustering Dosen Berdasarkan Kinerja Akademik Menggunakan Algoritma Genetika Paralel”, *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 8, pp. 2628 – 2635, 2018.

[9] J. F. Hair, R. E. Anderson, R. L. Thatham, *and* W. C. Black, *Multivariate Data Analysis Seventh Edition.* New Jersey: Pearson Education, Inc, 2010.

[10] J. Supranto, *Analisis Multivariat : Arti dan Interpretasi.* Jakarta: PT. Rineka Cipta, 2004.

[11] Kementerian Kesehatan, *Profil Kesehatan Republik Indonesia 2020.* Jakarta: Kementerian Kesehatan, 2020.

[12] L. Fadhurullah, “Gambaran Kualitas Pelayanan Kesehatan Ibu dan Anak”, *Psikoborneo*, vol. 6, no. 1, pp. 81-91, 2018.

[13] L. Vendramin, R. Campello, *and* E. R. Hruschka, “On the Comparison of Relative Clustering Validity Criteria”, *Proceedings of the SIAM International Conference on Data Mining*, vol. 3, no. 4, pp. 733-744, 2009.

[14] M. D. Kartikasari, “Self-Organizing Map Menggunakan Davies Bouldin Index dalam Pengelompokan Wilayah Indonesia Berdasarkan Konsumsi Pangan”, *Jambura J.Math*, vol. 3, no. 2, pp. 187-196, 2021.

[15] N. Pratiwi, *Implementasi K-Means dan K-Medoids Clustering dalam Pengelompokan Unit Usaha Koperasi (Studi Kasus: Unit Usaha Koperasi Terdaftar di Kabupaten Sleman per Tahun Buku 2014).* Yogyakarta: Universitas Islam Indonesia, 2016.

[16] R. A. Johnson, *and* D. W. Wichern, “Applied Multivariate Statistical Analysis”, 2002, http://faculty.smu.edu/tfomby/eco5385/lecture/Scoring%20Measures%20for%20Prediction%20Problems.pdf [Accessed: 10 January 2022].

[17] R. Awasthi, A. K. Tiwari, *and* S. Pathak, “Empirical Evaluation on K-Means Clustering with Effect of Distance Functions for Bank Dataset”, *International Journal of Innovative Technology and Research*, vol. 1, no. 3, pp. 233-235, 2013.

[18] Y. Agusta, “K-Means-Penerapan, Permasalahan, dan Metode Terkait”, 2007, http://www.yudiagusta.file.wordpress.com/2008/03/K-Means.pdf [Accessed: 10 January 2022].