EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI: 10.20885/EKSAKTA.vol2.iss2.art3

Research Article

Hierarchical Clustering Approach for Region Analysis of Contraceptive Users

Dina Tri Utari[1*], **Denesa Salma Hanun**[2]

[1,2] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, Jl. Kaliurang Km 14.5, Yogyakarta

[*] *Corresponding author: dina.t.utari@uii.ac.id*

Received: 19 August 2021; Accepted: 4 September 2021; Published: 9 September 2021

Abstract: Through increasing the use of contraceptives to limit births, the Family Planning (KB) Program is one of the government's efforts to control the rate of population growth. Klaten Districts is one of the regencies in Central Java Province with a relatively high number of births and relatively low coverage of active family planning. This study aimed to determine the grouping of sub-districts and these characteristics in the Klaten Districts in 2020. The method used in this study was a hierarchical cluster analysis method, with the best method being the centroid method. In this study obtained 3 clusters with cluster 1 consisting of 23 sub-districts, cluster 2 consists of 2 sub-districts and cluster 3 with 1 sub-district. The cluster characteristics based on the highest number of users of contraceptive methods are cluster 1 - contraceptives injection, cluster 2- contraception implant, and IUDs in cluster 3.

Keywords: Contraceptive methods, Hierarchical clustering, Centroid method, Characteristics of cluster

Introduction

Population explosion is a critical problem experienced by developing countries, especially in Indonesia. The large number and population growth rate cannot be controlled, resulting in deteriorating health levels, low education levels, and increasing unemployment. One of the government's efforts to suppress population growth and create small, happy and prosperous families is through the Family Planning (KB) Program [1]. One indicator of the achievement of the family planning program is the increasing number of family planning acceptors [2]. Every couple who uses contraception is based on an explicit request for family planning, delaying pregnancy, arranging pregnancy intervals, or not wanting to have more children. Contraceptive devices or methods are one of the efforts to regulate pregnancy. Contraceptive methods are long and short-term. Long-term contraceptive methods (MKJP) include Intra-Uterine Device (IUD), contraceptive implant, male sterilization, and female sterilization. In comparison, those included in Non-MKJP are contraceptive injections, pills, and condoms [3].

Based on data from the Central Java Provincial Health Office in 2020, the number of births in the Klaten Districts is relatively high, with the number of births being 16,061 people. The coverage of active family planning in the Klaten Districts is in a low category.

This study aims to examine the use of contraceptive methods in sub-districts in the Klaten Districts. If the people of Klaten Districts have not maximized the use of contraceptive methods, then family planning cadres in each sub-district can promote this government program again. Therefore, it is hoped that this research can be a recommendation for related parties to design policies to reduce the number of births that are under the characteristics of the use of contraceptive methods in each sub-district.

The method that will be used in this research is the hierarchical clustering method [4]. The use of this method is because the existing objects are less than 100, and it will be more effective in using this method because the object of this study only consists of 26 districts. We use the agglomerative hierarchical cluster method. There are sub-methods such as single linkage, complete linkage, average linkage, centroid method, and ward's method. Of the five methods, the best method for clustering will be determined.

Materials and Methods

Materials

The data used in this study is secondary data obtained from the Central Statistics Agency of Klaten Districts. The data used are related to the number of contraceptives uses in the Klaten Districts in 2020. Table 1 shows the variables used in this study.

Table 1. Variable Definition [5]

Variable	Definition	Unit
IUD (Intra Uterine Device)	A small T-shaped plastic and copper device that is put into the uterus.	People
Pill	A pill containing synthetic versions of female hormones estrogen and progesterone is produced naturally in the ovaries.	People
Condoms	An ultra-thin latex worn on the male genitalia.	People
Male Sterilization	A clinical procedure stops a man's reproductive capacity by severing the vas deferens to prevent pregnancy permanently.	People
Female Sterilization	A clinical procedure stops a woman's fertility by cutting, tying, and placing a ring on the fallopian tube.	People
Contraceptive Injection	The contraceptive injection releases the hormone progestogen into a woman's bloodstream to prevent pregnancy.	People
Contraceptive Implant	The contraceptive device is a small rod inserted under the skin of a woman's upper arm.	People

Furthermore, the variables in Table 1 are used to group the sub-districts in Klaten Districts.

Clustering

Clustering is a process for grouping data into several clusters [6]. Cluster analysis is a technique that aims to identify a group of objects that have specific similar characteristics that can be separated from other object clusters. Therefore, objects in the same cluster are relatively more homogeneous than objects in different clusters.

Clustering Analysis Assumptions

In conducting cluster analysis, there are two critical issues that the researcher must focus on, namely [7, 8]:

1. Representativeness of The Sample

   The sample taken must genuinely represent the existing population. There is no provision regarding the number of representative samples. However, a large enough sample is still needed to carry out the clustering process correctly. To find out whether the sample taken can genuinely represent the existing population, the Kaiser-Meyer-Olkin (KMO) value is needed. KMO is a comparison index of correlation coefficient value to partial correlation.

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \qquad (1)$$

$$a_{ij} = \frac{-r_{ij}}{\sqrt{r_{ij} r_{ij}}} \qquad (2)$$

   where:
   p : number of variables
   r_{ij} : correlation coefficient between variables i and j
   a_{ij}: partial correlation coefficient between variables i and j
   A KMO value of less than 0.5 indicates that the sample taken cannot represent the existing population.

2. Impact of Multicollinearity

   There should be no multicollinearity in the assumption test, namely the linear relationship between the independent variables. The value can see multicollinearity itself of Variance Inflation Factor (VIF).

$$VIF_i = \frac{1}{(1-R_i^2)} \qquad (3)$$

where:
R_i^2 : coefficient of determination
If the VIF value exceeds 10, it can be concluded that there is multicollinearity [9].

Hierarchical Clustering

The formation of hierarchical clusters has properties such as developing a hierarchy or a branching tree-like structure. Start grouping with two or more objects with the closest similarity, which will later be passed on to other objects so that the cluster will form like a tree. The tree has a precise level between objects, from the least similar to the most similar. The hierarchical clustering method is divided into two algorithms, namely divisive and agglomerative.

Agglomerative hierarchical clustering is a bottom-up hierarchical clustering method that combines n clusters into a single cluster. This method places each data object as a separate cluster. The following are several methods of agglomerative hierarchical clustering [6]:

1. Single Linkage Method

   The single-linkage method (also called the nearest-neighbor method) defines the similarity between clusters as the shortest distance from any object in one cluster to any object in the other. The steps of the single linkage method are as follows:

   a) Determine the minimum distance $D = (dx)$.
   b) Calculate the distance between the cluster that has been formed in step 1 with other objects.
   c) From the above algorithm, the distances between (xy) and the other z clusters are calculated by the formula:

   $$d(xy)z = \min\{dxz, dxy\} \qquad (4)$$

   The quantities in dxz and dyz are respectively the shortest distance between cluster x and z and cluster y and z. The results of the clustering can be displayed graphically in the form of a dendrogram or tree diagram. The tree branches represent the number of clusters.

2. Complete Linkage Method

   The complete-linkage method (also known as farthest-neighbor or diameter method) is comparable to the single-linkage algorithm, except that cluster similarity is based on maximum distance between observations in each cluster. The steps in the complete linkage method are the same as the single linkage method. The difference is in calculating the distance between clusters.

   $$d(xy)z = \max\{dxz, dxy\} \qquad (5)$$

   According to [7] dxz and dyz are the distances between objects that are farthest from clusters x and z and clusters y and z.

3. Average Linkage Method

   The average linkage procedure differs from the single-linkage or complete-linkage procedures in that the similarity of any two clusters is the average similarity of all individuals in one cluster with all individuals in another. The formula for the average method is as follows:

   $$d(xy)z = \frac{n_x}{n_x+n_y} d_{xz} + \frac{n_y}{n_x+n_y} d_{yz} \qquad (6)$$

   d_{xz}: distance between cluster x and cluster z
   d_{yz} : distance between cluster y and cluster z
   n_x : number of individuals in cluster x
   n_y : number of individuals in cluster y

4. Centroid Method

   In the centroid method the similarity between two clusters is the distance between the cluster centroids. Cluster centroids are the mean values of the observations on the variables in the cluster variate. Every time a new cluster occurs, the centroid will be recalculated immediately until a fixed cluster is formed. The advantage of this method is that outliers do not have a significant effect compared to other methods. The formed centroid cluster is obtained using the following formula:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} \tag{7}$$

where $N_1 = N_2$ is the number of objects.

5. Ward's Method

The ward's method differs from previous methods in that the similarity between two clusters is not a single measure of similarity, but rather the sum of squares within the cluster summed over all variables. This similarity calculation uses Squared Euclidean between each object as follows:

$$d(xy)z = \frac{(n_x+n_z)d_{xz} + (n_y+n_z)d_{yz} - n_z d_{xy}}{n_x+n_y+n_z} \tag{8}$$

n_x : number of objects in cluster x
n_y : number of objects in cluster y
n_z : number of objects in cluster z
d_{xz}: distance between cluster x and cluster z
d_{yz} : distance between cluster y and cluster z
d_{xy}: distance between cluster x and cluster y

Cophenetic Correlation Coefficient

The cophenetic correlation coefficient is one way that can be used to determine the best grouping method. The cophenetic correlation coefficient is the correlation coefficient between the original elements of the dissimilarity matrix (Euclidean distance matrix) and the elements generated by the dendrogram (cophenetic matrix). The formula for the cophenetic correlation coefficient is as follows [10]:

$$r_{coph} = \frac{\sum_{i<k}(d_{ik}-\bar{d})((d_{c_{ik}}-\bar{d}_c)}{\sqrt{[\sum_{i<k}(d_{ik}-\bar{d})^2][\sum_{i<k}(d_{c_{ik}}-\bar{d}_c)^2]}} \tag{9}$$

where:
r_{coph} : cophenetic correlation coefficient
d_{ik} : Euclidean distance between object i and k
$\bar{d}$: average of d_{ik}
$d_{c_{ik}}$: cophenetic distance between object i and k
$\bar{d}_c$: average of $d_{c_{ik}}$

Results and Discussions

Descriptive Statistics

Descriptive statistics are used to summarize the data in an organized manner by describing the relationship between variables in the data, a sample, or a population. Calculating descriptive statistics is an essential first step when conducting research and should always be done before making inferential statistical comparisons [11]. Descriptive statistics for each variable are presented in Table 2.

Table 2. Descriptive Statistics

Method	Total
IUD (Intra Uterine Device)	12,827
Pill	10,300
Condoms	4,241
Male Sterilization	277
Female Sterilization	7,901
Contraceptive Injection	28,176
Contraceptive Implant	24,906

Based on Table 2., in 2020, the population of Klaten Districs used injection as a contraceptive method the most, namely 28,176 people, followed by implants as many as 24,906 people, and the least used was male sterilization, which was only 277 people.

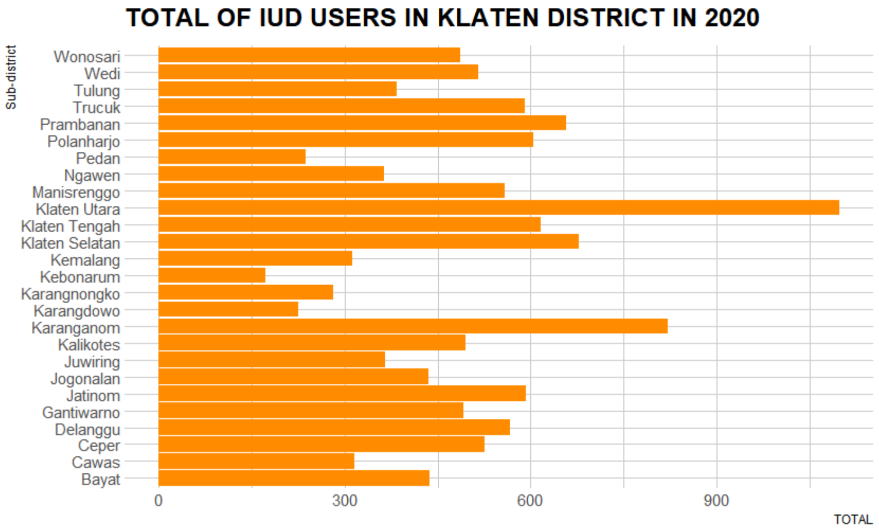

Figure 1. Distribution of IUD user in every sub-district in Klaten districts

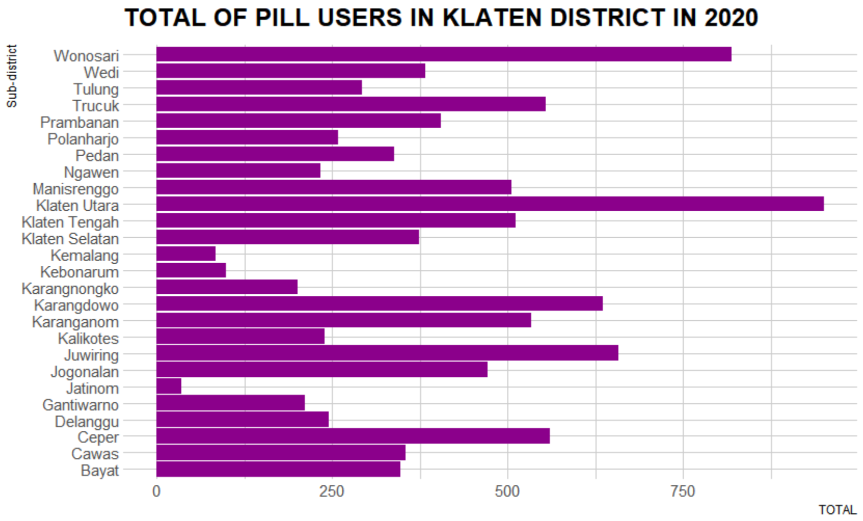

Figure 2. Distribution of pill user in every sub-district in Klaten districts

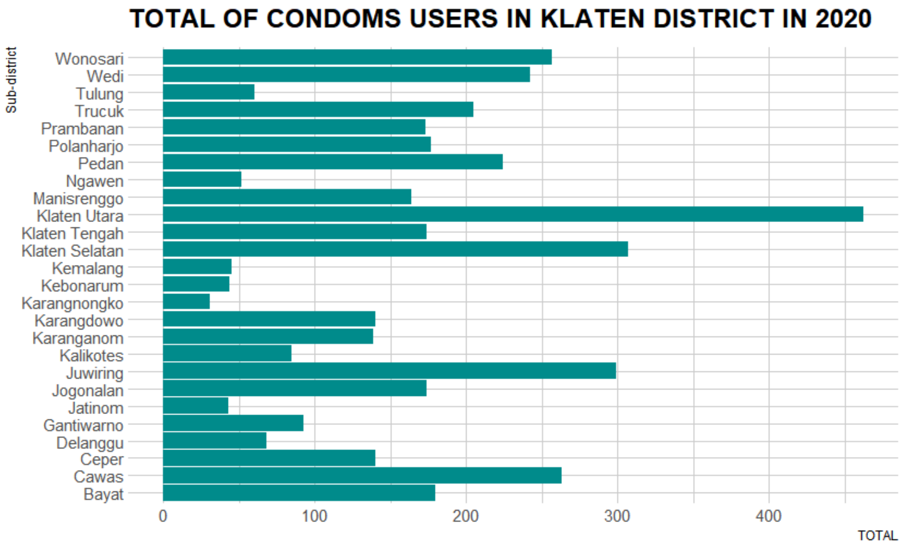

Figure 3. Distribution of condoms user in every sub-district in Klaten districts

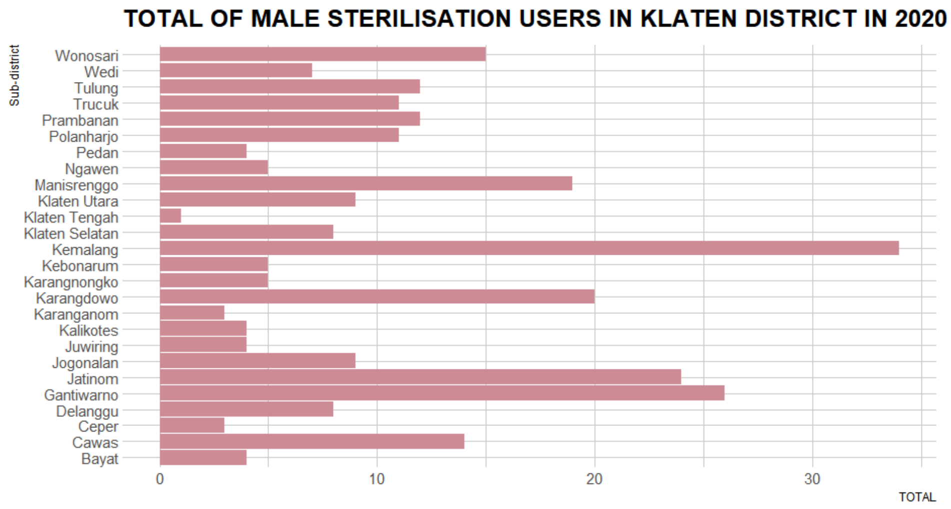

Figure 4. Distribution of male sterilization user in every sub-district in Klaten districts

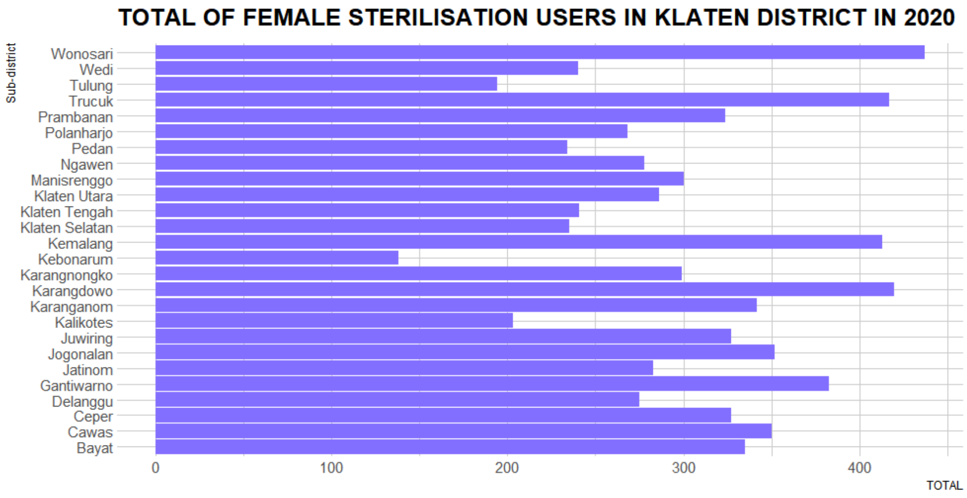

Figure 5. Distribution of female sterilization user in every sub-district in Klaten districts

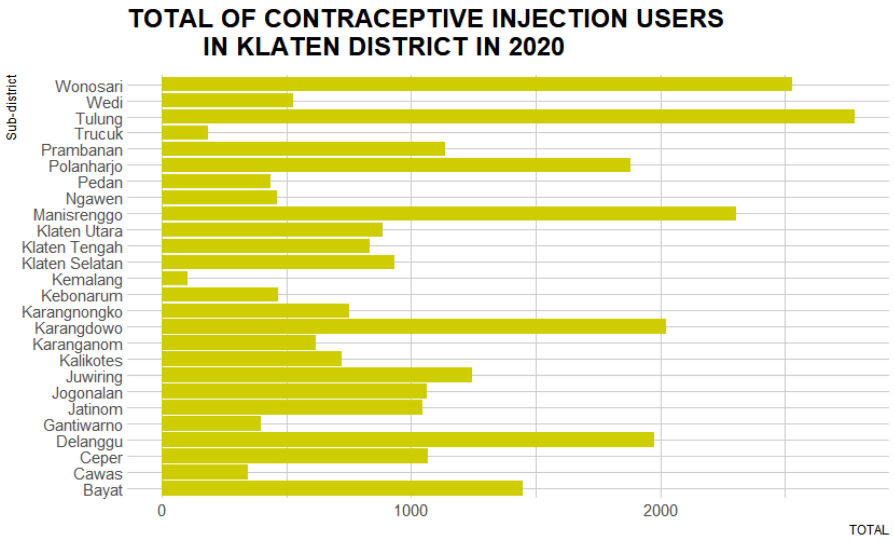

Figure 6. Distribution of contraceptive injection user in every sub-district in Klaten districts

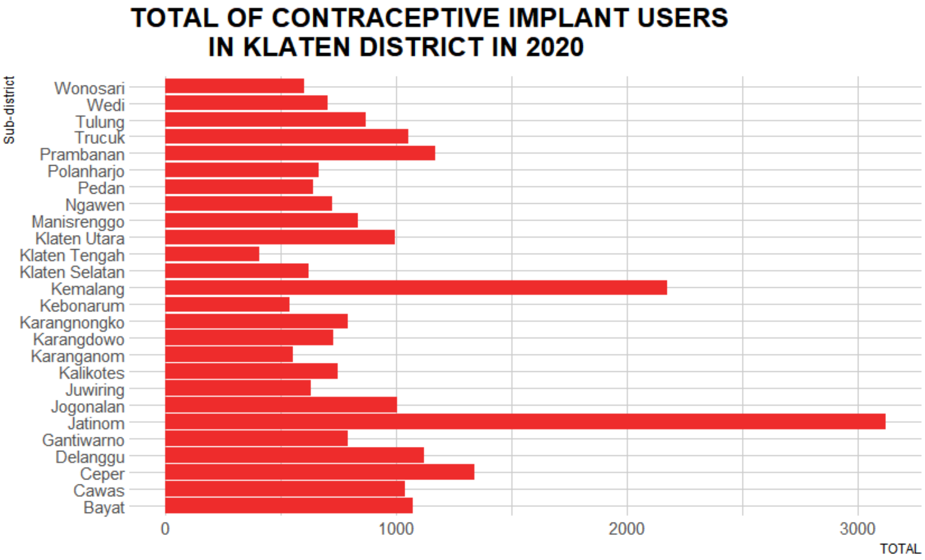

Figure 7. Distribution of contraceptive implant user in every sub-district in Klaten districts

Figure 1. to Figure 7. showing the distribution of each contraceptive method in every sub-district. Klaten Utara is a sub-district where most of the population uses IUDs, pills, and condoms. In contrast, the fewest users for each method are Kebonarum, Jatinom, and Karangnongko. Furthermore, the number of male sterilization users is quite different in each sub-district, but most found in Kemalang and least in Klaten Tengah. Unlike male sterilization, the number of female sterilization users is relatively evenly distributed in each sub-district, recorded highest in Wonosari and lowest in Kebonarum. The contraceptive injection and implant users, the highest population, were found in Tulung and Jatinom, respectively. At the same time, the fewest are found each in Kemalang and Klaten Tengah.

Clustering Assumptions

This study uses population data. Therefore, it is clear that the population data must be representative and there is no need for a KMO test [12]. Before we conducted hierarchical cluster analysis, multicollinearity testing was carried out. According to [13], an indication of multicollinearity is if the VIF value between independent variables is more than 10. Multicollinearity testing is used to determine the size of the distance that can be used. There are two measures of distance, namely Euclidean and Mahala Nobis. The Euclidean distance measure is used if the independent variables do not indicate multicollinearity in them. Meanwhile, the Mahala Nobis distance measure is used if the independent variables indicate multicollinearity in it.

Table 3. VIF Value

Variables	VIF
IUD (IntraUterine Device)	1.632
Pill	6.127
Condoms	3.316
Male Sterilization	2.532
Female Sterilization	2.770
Contraceptive Injection	1.485
Contraceptive Implant	1.944

Based on the calculation of the VIF between the independent variables in Table 3, there is no VIF value of more than 10. Thus, it can be concluded that there is no indication of multicollinearity in the independent variables in the data used.

The Results of Clustering Analysis

We use the value of the cophenetic correlation coefficient to determine the best cluster method. The value of the cophenetic correlation coefficient that is close to 1 indicates that the better the results of the clustering process using this method [14].

Table 4. Cophenetic Correlation Coefficient

Method	Result
Single	0.8167321
Complete	0.4550841
Average	0.8060587
Centroid	0.8282310
Ward's	0.6427581

In Table 4, it can be seen that the value of the cophenetic correlation coefficient in the centroid method is the highest. Therefore, the centroid method is the best cluster method that can be used in this data.

The following is a dendrogram for data on contraceptive users in Klaten Districts by sub-district in 2020 [15].

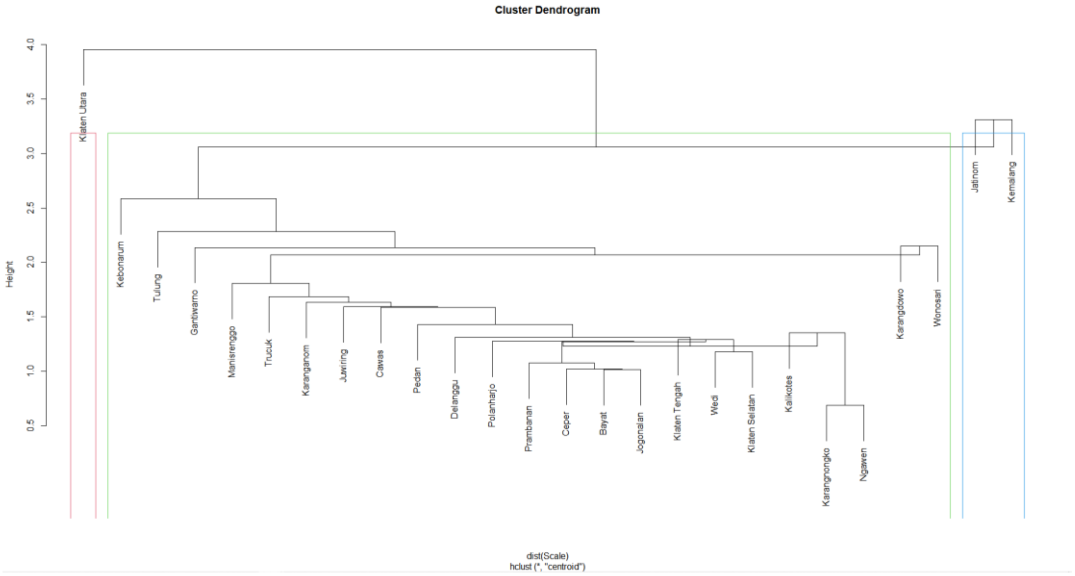

Figure 8. Cluster Dendrogram

Based on Figure 8, it is known that cluster 1 is marked with a green line, a blue line represent cluster 2, and the red one for cluster 3. Then, the researcher has determined the number of classifications of 3 clusters. There are no provisions regarding the number of classifications. Researchers used 3 clusters to determine the sub-districts with the classification level of the population that used each contraceptive method. The classification of cluster results for data on contraceptive users in Klaten Districts by sub-district in 2020 is presented in the Table 5.

Table 5. Result of Clustering

Cluster	Total of members	Member
1	23	Prambanan, Gantiwarno, Wedi, Bayat, Cawas, Trucuk, Klaikotes, Kebonarum, Jogonalan, Manisrenggo, Karangnongko, Ngawen, Ceper, Pedan, Karangdowo, Juwiring, Wonosari, Delanggu, Polanharjo, Karanganom, Tulung, Klaten Selatan, Klaten Tengah
2	2	Jatinom, Kemalang
3	1	Klaten Utara

Profiling Results of Clustering

The clustering process that has been carried out has resulted in three clusters that have each member. In the concept of clustering, each entity is grouped into clusters based on the similarity of its attributes so that each cluster has characteristics or profiles that distinguish one cluster from another. In Table 6 below, a summary of the characteristics of each identified cluster is presented.

Table 6. Average of Cluster Result Characteristics (in units of person)

Cluster	IUD	Pill	Condoms	Male Sterilization	Female Sterilization	Contraceptive Injection	Contraceptive Implant
1	470.7	401.3	160.5	9.1	300.8	1136.4	809.7
2	452.0	59.5	44.0	29.0	348.0	576.0	2644.5
3	1097.0	950.0	462.0	9.0	286.0	887.0	994.0

Table 6 shows that cluster 1 has the highest average contraceptive injection users compared to other clusters. However, users of IUDs, pills, condoms, male sterilization, female sterilization tend to be moderate, and the use of contraceptive implants is the lowest. In contrast to cluster 1, in cluster 2, it was found that contraceptive implants, male sterilization, and female sterilization were the highest, while other contraceptive methods were the lowest. Meanwhile, cluster 3 majority used IUDs, pills, contraceptive implants, contraceptive injections, and condoms. The lowest use is in male sterilization. The distribution map of the clustering results in Klaten District is presented in the following figure.

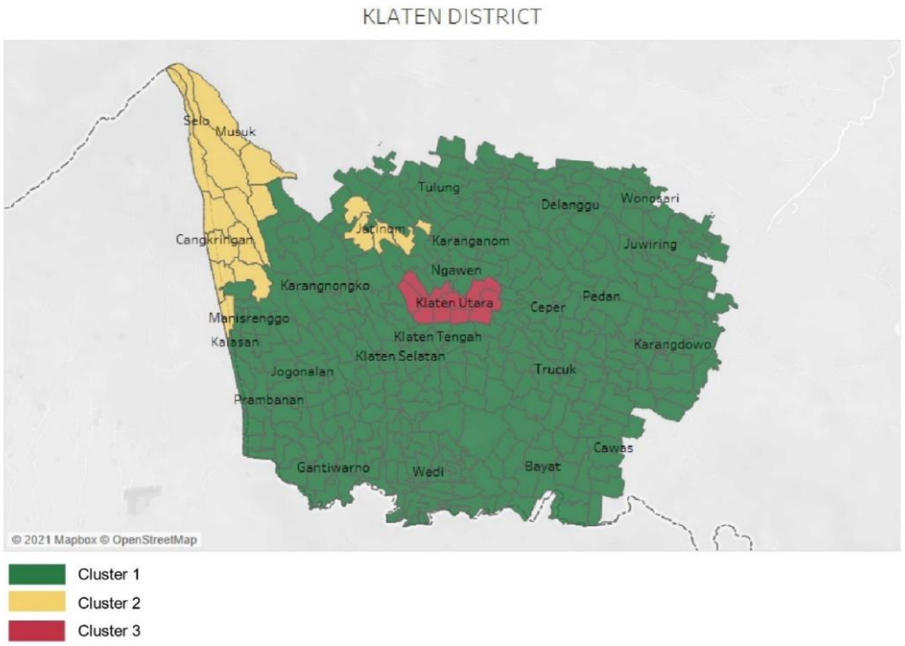

Figure 9. Map of Clustering Result

Conclusion

The best method for the data in this study is the centroid method from several methods in the hierarchical cluster. This study obtained 3 clusters with each characteristic. In cluster 1, the most widely used contraceptive method was a contraceptive injection, and the lowest was a contraceptive implant. In contrast, in cluster 2, the contraceptive implant method is the highest. While in cluster 3, several contraception methods are most widely used, namely IUDs, pills, contraceptive implants, contraceptive injections, and condoms.

References

[1] A. Sulistyawati, Pelayanan Keluarga Berencana, Salemba Medika, Jakarta, 2011.

[2] S. L. Naustion, Faktor-Faktor Yang Mempengaruhi Penggunaan MKJP di Enam Wilayah Indonesia, Pusat Penelitian dan Pengembangan KB, Jakarta, 2011.

[3] B. K. RI, Riset Kesehatan Dasar (RISKESDAS), Balitbang Kemenkes RI, Jakarta, 2013.

[4] P. Shetty and S. Singh, Hierarchical Clustering: A Survey, International Journal of Applied Research 7(4) (2021) 178-181.

[5] NHS, Your Contraception Guide, 17 March 2021. [Online]. Available: https://www.nhs.uk/conditions/contraception/. [Accessed 7 June 2021].

[6] P. N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Pearson Education, Boston, 2006.

[7] J. F. Hair, R. E. Anderson, R. L. Tatham and W. C. Black, Multivariate Data Analysis Fifth Edition, Prentice-Hall, Inc., USA, 1998.

[8] R. A. Johnson and G. K. Bhattacharyya, Statistics Principles & Methods, John Wiley & Sons, USA, 2010.

[9] J. I. Daoud, Multicollinearity and Regression Analysis, Journal of Physics: Conference Series 949 (2017) 1-6.

[10] S. Saracli, N. Dogan and I. Dogan, Comparison of Hierarchical Cluster Analysis Methods by Cophenetic Correlation, Journal of Inequalities and Applications 203(1) (2013) 1-8.

[11] P. Kaur, J. Stoltzfus and V. Yellapu, Descriptive Statistics, International Journal of Academic Medicine 4(1) (2018) 60-63.

[12] N. W. D. Ayuni and I. G. A. M. K. K. Sari, Analysis of Factors that Influencing the Interest of Bali State Polytechnic’s Students in Entrepreneurship, Journal of Physics: Conference Series, 953 (2018) 1-10.

[13] I. Gozhali, Aplikasi Analisis Multivariat dengan Program SPSS, Badan Penerbit Universitas Diponegoro, Semarang, 2001.

[14] P. R. Carvalho, C. S. Munita, A. L. Lapolli, Validity Studies Among Hierarchical Methods of Cluster Analysis Using Cophenetic Correlation Coefficient, International Nuclear Atlantic Conference, (C) (2017)

[15] Z. Zhang, F. Murtagh, S. P. Poucke, S. Lin and P. Lan, Hierarchical Cluster Analysis in Clinical Research with Heterogeneous Study Population: Highlighting Its Visualization With R, Annals of Translational Medicine 5(4) (2017) 1-11.