Volume 5, Nomor 1, Juni 2021
Numerical: Jurnal Matematika dan Pendidikan Matematika
http://journal.iaimnumetrolampung.ac.id/index.php/numerical
DOI:https://doi.org/10.25217/numerical.v5i1.1358

Application of Cluster Analysis Using Agglomerative Method

M. Rais Ridwan[1]*, Heri Retnawati[2]

[1] STKIP YPUP Makassar, Indonesia
[2] Universitas Negeri Yogyakarta, Indonesia
Correspondence: ✉ mraisridwan@stkip.ypup.ac.id

Article Info

Article History
Received: 30-04-2021
Revised: 26-06-2021
Accepted: 28-06-2021

Keywords:

Agglomerative Method; Average Linkage Method; Cluster Analysis; Standard Deviation Average Ratio; Ward Method.

Abstract

Improving the quality of human resources is the main supporting factor in increasing national productivity in various fields and development sectors. The government's productive investment activities that spur the nation's competitiveness in the global era prioritize Indonesia's education development. This study aims to cluster provinces in Indonesia based on educational indicators using the Agglomerative method consisting of the Average Linkage and Ward methods. Data collection is based on documentation techniques obtained from Statistics Indonesia in 2018. Data analysis used hierarchical cluster analysis consisting of data standardization, determining the size of the similarity or dissimilarity between data, the clustering process with a distance matrix, and seeing the characteristics of the cluster results formed. The second clustering method is by doing the initial grouping and determining the excellent cluster based on the average standard deviation ratio to the standard deviation between groups. Clustering results show the Ward method with the number of collections as many as 4 clusters and produces a ratio with a value of 0.01 smaller than the Average Linkage method. It shows that the cluster analysis method using the Ward method has better group accuracy quality than the Average Linkage method.

INTRODUCTION

Improving the quality of human resources is the main supporting factor in increasing national productivity in various fields and development sectors. The government's productive investment that can spur the nation's competitiveness in the global era prioritizes Indonesia's education development. The government determines education development policies in three policy pillars which are outlined in the mission of education. These policies focus on increasing education services' availability, expanding the affordability of education services, improving the quality of education services, realizing equality in education services, and ensuring certainty of obtaining educational services [1]. The education development policy is an indicator of a measuring tool for success based on a strategic education plan with an educational mission. However, efforts to develop education in Indonesia entering the 21st century face challenges in preparing human resources quality since the existence of education autonomy because not all districts or cities can provide valid and reliable data information to the center.

Another challenge is understanding education indicators and the relationship with accountability for the success of education development programs. Only a few education managers who are in the ranks of the Ministry of Education and Culture or education managers in the Provincial Education Office and District or City Education Office understand these two things [1]. Therefore, the preparation and study of education indicators for the education development program's success is one way for all education managers to understand various educational indicators. It can be used to assess the academic development program that has been implemented.

Indicators are tools that can explain and interpret the relationship between different aspects of education in the education system and between elements of the education system and the social, economic, and cultural systems of the human environment [1]. According to Green (1992), indicators are variables that can show or indicate to users about certain conditions so that they can be used to measure changes. Educational indicators based on education's mission consist of realizing broad, equitable, and equitable access and realizing quality learning [1]. Efforts in learning this access, the government provides access, quality of education services, and equality and certainty in obtaining educational services while for quality learning by considering the percentage of illiteracy rates.

Indicators of the availability of education services consist of the ratio of students per school, the proportion of students per class, the balance of courses per class, the percentage of school libraries, the percentage of school health business rooms, the percentage of computer rooms, and the percentage of laboratories. The indicator for the affordability of education services consists of the level of service at affordable local schools. The education service quality indicator consists of the percentage of teachers authorized to teach, the ratio of students per teacher, the number of graduates, the number of repeats, the dropout rate, and the percentage of classrooms belonging to good classrooms. Next, indicators of equality in access to education services include gender differences in gross enrollment rate (GER), GER gender parity index, and private students' percentage. Indicators of the certainty of obtaining education services consist of the gross enrollment rate, the gross input rate or continuing rate, the survival rate at level V or the survival rate, and the average length of study. Furthermore, education support indicators consist of the budget percentage to gross domestic product (GDP); and the education budget against the state budget (APBN). Another indicator is the percentage of the education budget by origin, population by education level; illiteracy rate; the number of children and adolescents still in school; and population aged 15 years and over according to the main occupation and highest education.

The results of the complete summary of the publication of the Statistics Indonesia related to Portrait of Indonesian Education in 2018 provide information about the achievement of the GER for Early Childhood Education (ECE) for the 3 to 6-year-old age group nationally increased to 37.92%, but still far below the development target of 77.2%. Meanwhile, based on residence, there is a disparity between urban and rural areas where the GER for ECE in urban areas is 48.71% higher than in rural areas, namely 36.14%. The factors that affect the ECE GER in all Indonesian provinces consist of the number of kindergartens, students' ratio per school, and students' ratio to teachers and principals [2].

Another publication of Statistics Indonesia shows that school participation still varies between education levels, which can be seen from the GER value for primary education levels or equivalent. The results of other Statistics Indonesia publications show that school participation still varies between levels of education. It is based on the GER value for elementary education level or equivalent, which exceeds 100%, indicating that the Elementry School level is not only the population in the 7-12 year age group. It is due to several factors that influence this condition: the student-teacher ratio and the number of schools for primary and junior high school levels. Also, aspects of regional characteristics show the significance of GER for all education units [3]. Other research related to the factors that influence GER for junior high school education shows that it is not a factor of socio-economic conditions and accessibility but a motivational factor [4]. The factors that affect the GER in Higher Education for each Province in Indonesia consist of central government expenditure in higher education to GRDP; student lecturer ratio; and population [5].

This study's educational indicators to cluster provinces in Indonesia are based on Statistics Indonesia's data in 2018 using the Average Linkage and Ward methods. The education indicators consist of the Population Literacy Rate aged 15-24 years; Literacy Rate ≥15 years old; Child Gross Participation Rate in Early Childhood Education; Higher Education Gross Enrollment Rate; Net Enrollment Rate Population with the lowest 40% expenditure group at the primary school, junior high school, senior high school level; Number of Villages with School Facilities at Elementary School, Junior High School, Senior High School, and College Level; GER Ratio at the Higher Education Level, as well as the Average Length of Schooling for Population Aged ≥15 Years.

The first use of cluster analysis by Tyron in 1939. The purpose of cluster analysis is to classify individuals who are independent of each other in a group to have the same or similar characteristics. Grouping cluster analysis uses a measure that describes the similarity or closeness between complex data into a simple group structure. This measure is a measure of distance or similarity [6] and a measure of the distance known as Euclid's distance [6].

The cluster analysis method consists of hierarchical and non-hierarchical methods. There is no known number of groups to be obtained in the hierarchical method. Meanwhile, the non-hierarchical method assumes that there are k groups first. The hierarchical method consists of the agglomerative and divisive methods. The agglomerative method consists of the Single Linkage method, Complete Linkage, Average Linkage, Ward's, Centroid, and the Median method [7]. The methods that are included in non-hierarchical methods are the K-means method and the fuzzy method. This study using a hierarchical method consisting of the Average Linkage and Ward methods.

The use of cluster analysis has been widely carried out in various scientific fields such as economics, geography, health, social, and multiple fields. The grouping of districts, districts or cities, and provinces in Indonesia uses cluster analysis based on indexes in the economic, geographic, health, and social fields [8], [9], [10], [11], [12], [13], [14]. The use of cluster analysis by grouping regions based on health indicators, people's welfare indicators, village potentials, macroeconomic indicators, human development indexes, and HIV/AIDS indicators.

Relevant research has been conducted by [14] by comparing cluster analysis with the Average Linkage method and the Ward Linkage Method in a case study of the Human Development Index in South Sulawesi Province. The results showed that the grouping using the Average Linkage method produced the best Dunn index of 0.55 compared to the Ward method of 0.43. Then, it was obtained the number of clusters formed as many as 8 clusters. Also, the number of groups formed is 8 clusters. Then, [8] reconducted research related to cluster analysis using the hierarchical method for grouping districts or cities in East Java-based health indicators. The hierarchical method used is Single Linkage, Complete Linkage, Average Linkage, Ward's, and Centroid based on the validity index, namely RMSSTD (Root Mean Square Standard Deviation). The results showed that the Ward Linkage method is the best method of grouping for the hierarchical method used with the smallest RMSSTD index value of 13.947 and forming clusters of 5 groups.

The following relevant research has been conducted by [10] by analyzing sub-district clusters in Semarang district based on village potential using the Ward and Single Linkage methods. The results showed that the Single Linkage method with R-Squared value is smaller than the Ward method, which shows that the Single Linkage method produces heterogeneous clusters compared to the Ward Linkage method. The subsequent research by [9] conducted a cluster analysis using the Average Linkage method in grouping districts or cities in Central Java Province based on People's Welfare Indicators. The results showed that the process of grouping 35 districts or towns in Central Java province could be formed three groups with groups A, B, and C, each consisting of 28, 2, and 5 districts or cities.

Subsequent research uses cluster analysis with the K-Means method for grouping districts or cities in Maluku province based on the 2014 human development index indicators, namely life expectancy, literacy rate, average years of schooling, and per capita expenditure rate [12]. The results showed that there were three clusters: cluster 1 consisting of Ambon City with a very maximum number compared to the other 2 clusters, cluster 2: MTB, Aru Islands, SBB, SBT, MBD, and Bursel, and cluster 3: Malra, Malteng, Buru, Tual. Research by conducting Cluster Analysis with Outlier Data Using Centroid Linkage and K-Means Clustering for Grouping HIV / AIDS Indicators in Indonesia shows that the Centroid Linkage method has a higher homogeneous level compared to the K-Means method [13]. The comparison of the two methods uses the SW and SB ratios. Furthermore, cluster analysis uses the Average Linkage method, and Ward uses Unit Link life insurance customer data [15]. The results showed that the Average Linkage method had better performance than the Ward method with SB and SW of 0.486 and 0.710.

Based on the indicator study and the article literature above, in this study, this study conducted and compared cluster analysis using the Agglomerative method, namely the Average Linkage method and the Ward method in showing regional clusters in Indonesia based on 14 educational indicators. Determining an exemplary group is based on the average standard deviation ratio in the cluster to the standard deviation between clusters.

METODE

Data sources and Research Variables

The data used in grouping using cluster analysis is provincial data in Indonesia in 2018. The data used is secondary data based on education indicators for all Indonesia provinces obtained from the Statistics Indonesia in 2018. The variables in this study consisted of the population literacy rate variable aged 15-24 years (X_1); literacy rate ≥15 years (X_2); The gross enrollment rate of children attending early childhood education (X_3); Higher Education Gross Enrollment Rate (X_4); Net Enrollment Rate (NER) population of the lowest 40% expenditure group is Elementary School level (X_5); NER population of the lowest 40% expenditure group is junior high school level (X_6); NER population of the lowest 40% expenditure group is high school level (X_7); population of the lowest 40% expenditure group at the Vocational High School level (X_8); Number of villages with primary school facilities (X_9); Number of villages that have junior high school facilities (X_{10}); Number of villages with senior high school facilities (X_{11}), number of villages with higher education facilities (X_{12}), GER ratio at the higher education level (X_{13}), average years of schooling for the population aged ≥15 years (X_{14}).

Research Stages

This study's stages consisted of data standardization, multicollinearity testing, a dendrogram of hierarchical cluster analysis method, and the best method's determination based on the average standard deviation ratio in the cluster to the standard deviation between groups.

1. Standardization of data

The standardization process was carried out for the study variables that had significant differences in unit sizes. Striking unit differences can result in invalid calculations in cluster analysis. Therefore, the standardization process needs to be done by transforming the original data before further analysis. The z-score result transforms p variables $X_1, X_2, \cdots, X_p$ into new variable p variables, $z_1, z_2, \cdots, z_p$ which are uncorrelated using the formula $z_i = u_i^t[x - \bar{x}]$ where u_i is the ith eigenvector obtained from the principal component analysis [16].

2. Multicollinearity testing

The use of data in cluster analysis should not be correlated so that there is no multicollinearity. In the cluster analysis, each variable is given the same weight in the calculation of the distance. If some of the variables are correlated, it will cause an unbalanced weighting. As a result, these conditions will affect the results of the analysis in grouping objects. According to [17], a very high correlation between independent variables would result in a regression model estimator that is biased, unstable, and perhaps far from its predictive value.

Identify the presence or absence of multicollinearity for each research variable based on the variance inflation factor (VIF). If the value of VIF≤10 and the value of tolerance ≥0.10, then the regression is free from multicollinearity conditions [18], [19], [20]. According to [21], a VIF value greater than 10 identifies a severe multicollinearity problem. According to [22], for the VIF≤10 value so that high multicollinearity occurs, the study variables should theoretically not be used in the OLS (Ordinary Least Square) regression model functions as a non-significant variable. Also, multicollinearity conditions can be identified based on a coefficient matrix with the correlation between the independent variables less than 0.5 [23]. In this study, multicollinearity testing used Variance Inflation Factor (VIF).

3. Hierarchy cluster analysis method

The grouping analysis in this study used a hierarchical method consisting of the Average Linkage and Ward methods. In general, according to [24], hierarchical cluster analysis is grouping N objects with the following procedure. The first step, starting with the number of *N* clusters. Each cluster contains a single element and asymmetrical matrix $D = \{d_{jl}\}$ is Euclid's distance using the formula $d_{jl} = \left\{(\boldsymbol{x_l} - \boldsymbol{x_j})'(\boldsymbol{x_l} - \boldsymbol{x_j})\right\}^{\frac{1}{2}} = \sqrt{\sum_{k=1}^{i}(x_{lk} - x_{jk})^2}$, $i = 1,2,\cdots,p$ and $l = 1,2,\cdots,n$. Second, determining the closest cluster pair distance with d_{UV} represents the closest distance to clusters *U* and *V*. Third, combining clusters *U* and *V* by identifying the new cluster formed with *(UV)* and recalculating the new distance matrix. The fourth step, repeating the second step as many as *N*-1 iterations so that all objects are in a single cluster.

The Average Linkage cluster method or the average linkage method is a method with the average distance principle. This cluster method's basic rule is the average distance between observations with grouping starting from the center or pairs of keeping with the average length. According to [25], this method begins with finding another member of $D = (d_{ik})$ and combining the corresponding objects, for example, U and V, to become (UV). Then, the distance between (UV) and another group, namely W, is written in the formula $d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}$, where $N_{(UV)}$ represents the sum of several members in group (UV). The cluster analysis steps using the Average Linkage method consist of checking pairs of adjacent provinces; combining them into one cluster;

calculate the distance between the two regions that merge into one group with another region; the next one combines the clusters most similar to form the second cluster. It is then calculated using formula $d_{(UV)W}$ to create a matrix with a new distance, repeating the second and third steps *N*-1 times, where *N* is the number of provincial objects.

The following cluster analysis method is the Ward method. This clustering method uses complete calculations and maximizes homogeneity within one group. In this method, the distance between two clusters is the squares' sum between the two clusters for all variables [26]. This method tends to be used to combine groups with small numbers. The formula used is $ESS = \sum_{j=1}^{k}\left[\sum_{i=1}^{n_i} x_{ij}^2 - \frac{1}{n_{ij}}\left(\sum_{j=1}^{n_j} x_{ij}^2\right)\right]$, where x_{ij} denotes the *i*-th object's value where $i = 1,2,3,\cdots$ is in the *j*-group; *k* means the number of groups per step, and n_{ij} represents the number of groups *i* in group *j*. The stages in cluster analysis using the Ward method consist of the initial steps taking into account *N* clusters with one province per cluster (all provinces are considered clusters) with *ESS* of zero. Second, the first cluster is formed by selecting two of the N clusters with the smallest *ESS* value. Third, re-identifying the *N-1* cluster clusters to determine two of these clusters, which can minimize heterogeneity so that *N*-1 systematically reduces *N*-clusters. The fourth step repeats the second and third steps until one cluster is obtained or all provinces merge into one cluster.

4. Determination of the best method

The step is to determine the best cluster analysis method by grouping based on distance measurements and then comparing them. The method selection is based on the average standard deviation ratio in the cluster to the standard deviation between groups to produce the best grouping quality. The average standard deviation in the group is written as s_W, represented by the formula $s_W = \frac{1}{c}\sum_{k=1}^{c} s_k$. While the standard deviation between clusters (s_B) is formulated as $s_B = \left[\frac{1}{c-1}\sum_{k=1}^{c}(\bar{x}_k - \bar{x})^2\right]^{\frac{1}{2}}$ where *c* is the number of clusters; s_k is the standard deviation in the *k*-th cluster, and $\bar{x}_k$ represents the average of the *k*-th clusters and $\bar{x}$ is the average of all clusters. Based on the s_W and s_B values with the ratio value, $\frac{s_W}{s_B}$, the smallest, the cluster method used has high homogeneity [27]. The smaller the s_B value and the greater the s_B value, the method has good accuracy.

RESULTS AND DISCUSSION

Descriptive Data Analysis

Data grouping in this study is based on education indicators with the number of provinces in Indonesia. Secondary data collection in 2018 is based on fourteen hands consisting of fourteen provinces as the research sample. Based on the data from the grouping results, an analysis was then carried out to obtain a summary of the results of the descriptive analysis for each education indicator below.

Tabel 1. Results of Descriptive Data Analysis Based on Educational Indicators

Variable	Minimum	Maximum	Mean	Median	Std. Deviation
X_1	88.44	100.00	99.45	99.80	1.97
X_2	76.79	99.87	96.00	97.89	4.57
X_3	13.17	69.80	35.25	32.39	10.46
X_4	13.20	70.60	33.32	33.59	10.65
X_5	76.43	99.53	96.94	98.010	3.99
X_6	50.41	87.94	75.56	77.03	8.09
X_7	33.60	70.44	55.03	54.76	7.99
X_8	264.00	8443.00	2124.62	1550.00	2047.14
X_9	144.00	4696.00	1097.26	772.50	1085.73
X_{10}	59.00	2385.00	491.06	313.50	516.57
X_{11}	24.00	1922.00	306.15	161.00	424.73
X_{12}	12.00	394.00	87.88	55.50	92.70
X_{13}	87.48	195.63	117.77	118.05	20.36
X_{14}	6.66	11.06	8.79	8.84	0.87

Based on the results of descriptive data analysis in Table 1 above, it shows that each research variable has a minimum and maximum data and mean, median value, and standard deviation. Table 1 shows that the median and mean values for each variable or indicator are relatively the same, except for the variables X_8, X_9, X_{10}, X_{11}, and X_{12}. In this case, it shows that the distribution is almost symmetrical. Meanwhile, the minimum and maximum values for each variable are pretty far apart. Therefore, the use of data sizes for the variables in this study has quite a significant difference, so it is necessary to transform the initial data into a z-score.

Multicollinearity testing

The cluster analysis process by calculating the distance gives the same weight to each variable in the study. So that if there are variables that are mutually correlated, it will cause an unbalanced weighting. As a result, these conditions will affect the results of the analysis in object grouping. Therefore, the collinearity testing process is carried out to identify the presence or absence of collinearity between variables. The following shows the results of calculating the VIF value for each research variable in table 2 below.

Data grouping in this study is based on education indicators with the number of provinces in Indonesia. Secondary data collection in 2018 is based on fourteen hands consisting of fourteen provinces as the research sample. Based on the data from the grouping results, an analysis was then carried out to obtain a summary of the results of the descriptive analysis for each education indicator below.

Tabel 2. VIF Value of Research Variable Indicators

Variable	VIF Value
X_1	1.003
X_2	1.014
X_3	1.044
X_4	1.049
X_5	1.013
X_6	1.010
X_7	1.048
X_8	29.145
X_9	72.962
X_{10}	14.909
X_{11}	8.283
X_{12}	6.506
X_{13}	1.012

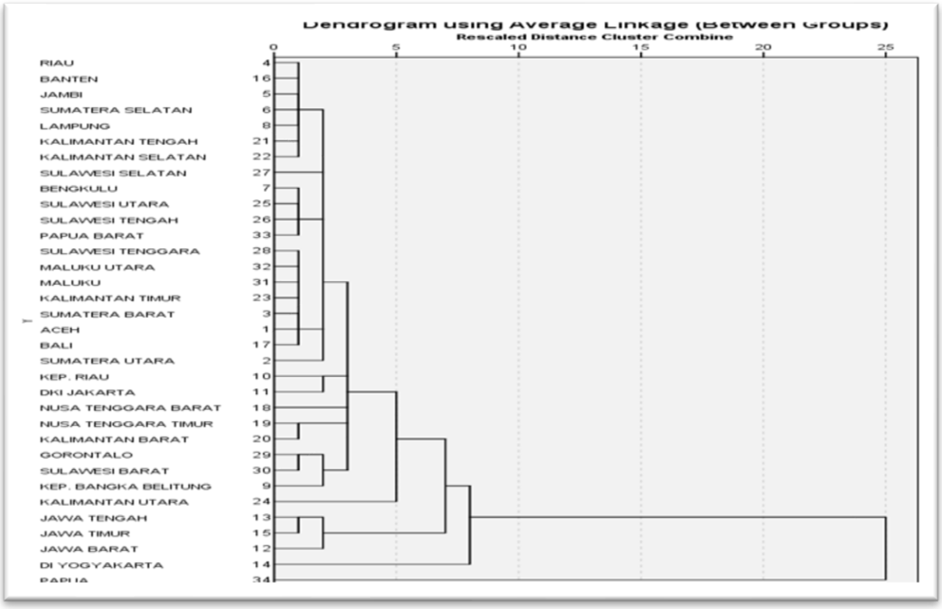
Based on the results of calculating the VIF value in Table 2 above, it shows that there are research variables with a VIF value greater than 10, namely the variables X_8, X_9, and X_{10}. It indicates that these variables indicate multicollinearity [21], [22]. Furthermore, according to [22], the variable with a VIF value is theoretically non-significant, so it is not used in the following analysis. Thus, eliminating variables with a VIF value greater than ten results in 11 research variables for determining the grouping of provinces in Indonesia based on education indicators. The elimination of these indicators or variables consists of the net enrollment rate (NER) of the population of the lowest 40% of the Vocational High School level (X_8), the number of villages that have primary school facilities (X_9), and the number of towns that have junior high school facilities (X_{10}).

Dendrogram of hierarchical cluster analysis method

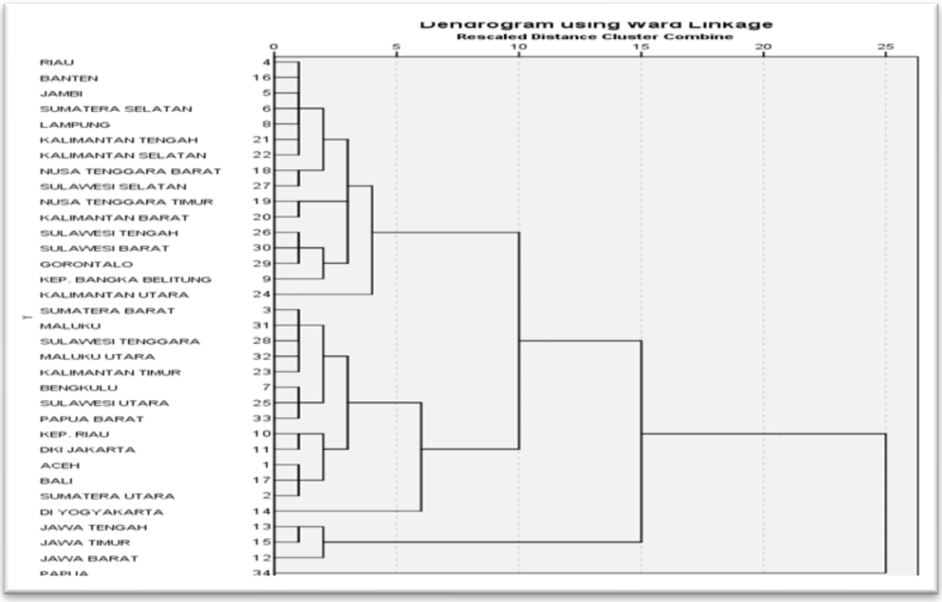
According to Mayr et al. (1953) in [28], the dendrogram is an illustration based on a diagram of relation about the level of similarity. In this conceptual relationship, by combining two data based on the similarities that exist in the data [23]. Merging continues for those that have similarities to other data. According to [29], this merger forms a tree-like appearance, called the agglomerative method. The agglomerative method is a classification method starting from one set of stands and then combining the grouping results with other perspectives into a cluster.

Interpretation of cluster characteristics

The dendrogram results show cluster analysis results by identifying the closest distance between objects as information on grouping objects with similar characteristics—illustration of two objects with the same elements based on two points with the most relative position. The closer the two objects are, the object has the same similarity. However, suppose the object's two ends are further away. In that case, the object is more and more different given the cluster analysis dendrogram using a hierarchical method consisting of four clusters in Figure 1 below.



(a)



(b)

Figure 1. Dendrogram Hierarchy Analysis with The Method (a) Average Linkage (b) Ward

The dendrogram results in Figure 1 above show the output of determining the number of casters and grouping them by the province in Indonesia. Figures 1(a) and 1(b) respectively show the grouping of provinces using the Average Linkage and Ward methods. The results of grouping provinces based on education indicators using the Average Linkage and Ward methods each obtained four clusters.

The results of grouping using the Average Linkage method consist of four clusters. The first cluster consists of the provinces of Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, DKI Jakarta, 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, and West Papua Province. Then, the second cluster consists of the provinces of West Java, Central Java and East Java. As for the third and fourth clusters, respectively, the Special Region Yogyakarta and Papua provinces.

Then, for grouping using the Ward method, it also consists of four clusters. The first cluster consists of the provinces of Aceh, North Sumatra, West Sumatra, Bengkulu, Riau Islands, DKI Jakarta, the Special Region Yogyakarta, Bali, East Kalimantan, North Sulawesi, Southeast Sulawesi, Maluku, North Maluku, and West Papua Province. The second cluster contains sixteen provinces, namely the provinces of Riau, Jambi, South Sumatra, Lampung, Bangka Belitung Islands, Banten, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, North Kalimantan, Central Sulawesi, South Sulawesi, Gorontalo and West Sulawesi. Meanwhile, the third cluster consists of the provinces of West Java, Central Java and East Java and the fourth cluster is only the province of Papua.

Interpretation of cluster characteristics

The determination of the number of clusters and cluster members using the two hierarchical methods above provides information regarding the number of provincial clusters based on education indicators. The next stage is the interpretation of cluster characteristics using each cluster's average for each variable (centroid). The interpretation process can use a centroid cluster [30]—understanding of cluster characteristics using the Average Linkage and Ward methods. In the following, the centroid values for each variable in the first and second clusters are given below.

Tabel 3. Results of the Centroid Value of Average Linkage Method

Variable	First Cluster	Second Cluster
X_1	99.76	99.92
X_2	96.84	94.59
X_3	33.55	47.51
X_4	33.32	25.70
X_5	97.45	98.09
X_6	75.70	78.43
X_7	55.71	50.44
X_{11}	188.55	1556.67
X_{12}	63.93	340.67
X_{13}	119.29	112.29
X_{14}	8.90	8.13

Table 3 above shows that the centroid value's determination is only for the two clusters because the third and fourth clusters each only consist of one object. Based on the centroid value for the first cluster to the variables X_2, X_4, X_7, and X_{13}, they have the highest value compared to the second cluster. It shows that for the provinces in the first grouping compared to the second grouping, it shows that the majority of people aged ≥ 15 years are still literate. Those with the lowest 40% expenditure for Senior High School level are still less participating. However, if it is viewed from the community participation perspective in continuing their studies at higher education institutions, it is greater than the people in the second grouping.

The variables X_1, X_3, X_5, X_6, and X_{12} each have the centroid value for the first cluster, which has the lowest value compared to the second cluster. It shows that the first grouping people lacked participation in starting their children's education in the Early Childhood Education program. However, the communities with the lowest 40% expenditure on primary and junior secondary school levels still have more intense participation when compared to communities in the second grouping. Besides, for school facilities at the tertiary level, the villages' number is greater than the villages in the second cluster. The centroid values for each variable in the first, second, and third clusters are given in below.

Tabel 4. Results of the Centroid Value of Wards Method

Variable	First Cluster	Second Cluster	Third Cluster
X_1	99.23	99.69	99.92
X_2	96.09	95.89	94.59
X_3	35.16	35.35	47.51
X_4	37.17	28.99	25.70
X_5	96.46	97.48	98.09
X_6	77.26	73.65	78.43
X_7	58.56	51.05	50.44
X_{11}	402.11	198.19	1556.67
X_{12}	113.67	58.88	340.67
X_{13}	113.02	123.12	112.29
X_{14}	9.15	8.38	8.13

Table 4 above shows that three clusters only determine the centroid value, with the fourth cluster consisting of only one object. Table 4 shows that the first cluster based on variables X_1, X_3, and X_5 each has the lowest value than the other clusters. It indicates that the community's reading frequency in the first group is higher than the two different groups. However, for the literate society, the frequency was highest compared to the other two clusters. Community participation to include Early Childhood Education is still very low compared to the community in the other two clusters. Likewise, for community participation for the most subordinate 40% expenditure groups at the primary school level.

Meanwhile, the variables X_2, X_4, X_7, and X_{14} each have the highest value among the other two clusters. Another variable shows that the first group's community has higher participation in continuing their studies than the other two clusters. However, for the variable length of schooling, residents over 15 years of age have a more significant percentage of students in the long period in completing their studies.

Then, for the second cluster, the variables X_6, X_{11}, and X_{12} each show the lowest mean compared to the other two clusters' provincial communities. It shows that the community for the lowest 40% expenditure group at the junior high school level has less participation and participation in continuing education in tertiary institutions. Also, several school facilities at the college level are still minimal, and the frequency of people over 15 years of age in the second cluster is more literate.

Furthermore, for the third cluster, the variables X_1, X_3, X_5, X_6, X_{11}, and X_{12} each have the most significant value than the other two clusters. It suggests that the frequency of literate people is higher than the other clusters and also the lack of community participation for the lowest 40% expenditure groups at the primary and junior high school levels. On the other hand, the community in this third grouping is the number of villages with a large number of higher education-level school facilities and community participation to continue their studies at tertiary institutions.

Next, the variables X_2, X_4, X_7, X_{13}, and X_{14} have the lowest average among the other two clusters. It indicates that the number of people aged 15 and over is less than the other two clusters, and the community's participation in the expenditure group for the lowest 40% at the Senior High School level. Then, in this cluster, there is still a lack of community participation to continue their studies. The people in this cluster also have a small number of average years of schooling aged 15 years and over.

Determination of the best method

The determination of the number of clusters and cluster members using the two hierarchical methods above provides information regarding the number of provincial clusters based on education indicators. The next stage is the interpretation of cluster characteristics using each cluster's average for each variable (centroid). The interpretation process can use a centroid cluster [30]—understanding of cluster characteristics using the Average Linkage and Ward methods. In the following, the centroid values for each variable in the first and second clusters are given in below.

Tabel 5. Standard Deviation Value of Average Linkage Method Cluster Analysis

Variable	First Cluster	Second Cluster
X_1	0.30	0.06
X_2	3.10	3.46
X_3	6.45	10.70
X_4	8.32	4.04
X_5	1.76	0.15
X_6	6.94	0.69
X_7	6.89	1.79
X_{11}	140.46	353.20
X_{12}	46.76	77.31
X_{13}	21.53	6.42
X_{14}	0.79	0.42

Table 5 above shows the standard deviation values for the first and second clusters for each research variable. Then, calculating each deviation value in the cluster by finding the square root of the sum of the difference between the standard deviation value and the standard deviation mean for each research variable obtained $\boldsymbol{S_1 = 75.76}$ and $\boldsymbol{S_2 = 193.07}$, respectively. The average standard deviation in the cluster for each cluster is obtained by dividing the total standard deviation value by the number of research variables obtained by $\boldsymbol{\overline{x}_1 = 22.12}$ and $\boldsymbol{\overline{x}_2 = 41.66}$. In contrast, each cluster's mean is obtained by dividing the sum of the average standard deviation in the cluster by the number of clusters to get $\boldsymbol{\overline{x} = 31.89}$.

Then, calculating the standard deviation value for each cluster using the Ward method is given below.

Tabel 6. Standard Deviation Value of Ward Method Cluster Analysis

Variable	First Cluster	Second Cluster	Third Cluster
X_1	2.70	0.37	0.06
X_2	5.44	3.51	3.46
X_3	12.71	7.58	10.70
X_4	11.75	7.42	4.04
X_5	5.18	1.99	0.15
X_6	9.49	5.89	0.69
X_7	8.81	4.56	1.79
X_{11}	561.29	127.79	353.20
X_{12}	116.69	42.17	77.31
X_{13}	13.30	25.58	6.42
X_{14}	0.97	0.50	0.42

Table 6 above shows the standard deviation value for the three clusters for each research variable. The calculation of each deviation value in the cluster by finding the square root of the sum of the difference between the standard deviation values and the mean standard deviation for the eleven research variables $\boldsymbol{S_1 = 304.66}$, $\boldsymbol{S_2 = 68.91}$, and $\boldsymbol{S_3 = 193.07}$. Then, the average standard deviation in the cluster for each cluster is obtained by dividing the total standard deviation value results and the number of research variables for each obtained $\boldsymbol{\overline{x}_1 = 68.03}$, $\boldsymbol{\overline{x}_2 = 20.67}$, and $\boldsymbol{\overline{x}_3 = 41.66}$. Meanwhile, the average for each cluster, by dividing the mean, standard deviation in the cluster, and the number of clusters, is obtained $\boldsymbol{\overline{x} = 43.45}$.

Furthermore, calculating the ratio using the Average Linkage and Ward methods is given in table 7 below.

Tabel 7. Standard Deviation Value of Ward Method Cluster Analysis

Cluster Analysis Method	Number of Clusters	S_W	S_B	Ratio
Average Linkage	4	67.21	63.63	1.06
Ward	4	141.66	10020.48	0.01

Based on table 7 above, the S_W value is obtained by dividing the total standard deviation value results between the cluster and the number of clusters using the Average Linkage and Ward methods of 67.21 and 141.66, respectively. Then, the S_B value is obtained by calculating the comparison of the sum of the squares of the difference in the mean deviations in the cluster, and the mean for each cluster with the number of clusters reduced by one is obtained for the Average Linkage method of 63.63. In contrast, for the Ward Linkage method it is obtained 10020.48.

Furthermore, by calculating the value of the S_W and S_B ratio for the Average Linkage method, it is obtained that it is 1.06. Meanwhile, the S_W and S_B ratio using the Ward Linkage method is 0.01 and smaller than the Average Linkage method. In this case, the Ward method produces a more homogeneous group so that the resulting ratio value is smaller. It means that the Ward method has better group accuracy quality than the Average Linkage method. These results indicate the same thing according to [31], which states that the Ward method is the most optimal method for similarity analysis.

This study discusses the analysis of provincial clusters based on education indicators using the Average Linkage and Ward methods. The results show that the Ward method has better classification accuracy than the Average Linkage method. However, cluster analysis using the Average Linkage method and the Ward method for Respondent Data for Unit Link Life Insurance Customers [15]. It shows the study results that the Average Linkage method has better performance than the Ward method with the respective S_B and S_W ratio values of 0.486. and 0.710. Subsequent research, cluster analysis using the Average Linkage and Ward methods in the case study of the Human Development Index in South Sulawesi Province by [14] using the Dunn index. It obtained grouping using the Average Linkage method resulting in the best Dunn index of 0.55 compared to the Ward method of 0.43.

It shows that determining the best method depends on the use of the research variable indicators and the procedures or stages in the study. Research by [15] has research stages consisting of standardizing data, selecting distance measurements, and implementing the hierarchical method's steps. Meanwhile, the research by [14] has sets consisting of data standardization, determining the size of the similarity or dissimilarity between data, the clustering process with the distance matrix determining the number of clusters and their members, looking at the characteristics of the cluster results formed. In this research, the research stages were carried out by data standardization, multicollinearity testing, hierarchical method cluster analysis dendrogram, interpretation of cluster characteristics, and determination of the best method. However, the results showed that the Ward method has better classification accuracy than the Average Linkage method.

CONCLUSIONS AND SUGGESTIONS

Analysis of the grouping of provinces in Indonesia based on educational indicators uses the Average Linkage and Ward methods. The reduction in the number of variables in this study based on the VIF value consisted of three research variables. They were eliminated for further analysis in the grouping of provinces. Then, the S_W and S_B ratios' acquisition uses the Ward Linkage method of 0.01, which is smaller than the Average Linkage method of 1.05. It shows that the Ward method's grouping analysis produces a more homogeneous group with a smaller ratio value. Thus, the Ward method has better group accuracy quality than the Average Linkage method. Meanwhile, suggestions in research in determining the best agglomerative method depend on the use of research variable indicators and procedures or stages in the study. In this research, the research stages were carried out by data standardization, multicollinearity testing, hierarchical method cluster analysis dendrogram, interpretation of cluster characteristics, and determination of the best method.

REFERENCES

[1] Badan Pusat Statistik, *Potret Pendidikan Indonesia Statistik Pendidikan 2018*. Indonesia: Badan Pusat Statistik, 2018.

[2] Y. Kartakusumah, "Analisis Faktor yang Mempengaruhi Angka Partisipasi Kasar (APK) pada Pendidikan Anak Usia Dini Taman Kanak-Kanak," *S2 Thesis*, 2018.

[3] N. A. Lestari and A. Adji, "Analisis faktor-faktor yang mempengaruhi angka partisipasi sekolah serta angka putus sekolah tingkat Sekolah Dasar dan Sekolah Menengah Pertama," *S2 Thesis*, 2014.

[4] H. R. Rahmatika, "Faktor-faktor yang mempengaruhi Angka Partisipasi Kasar (APK) jenjang pendidikan Sekolah Menengah Pertama pada masyarakat pesisir di Kecamatan Sarang Kabupaten Rembang tahun 2015," *S1 Thesis*, 2016.

[5] S. Habibah, Y. P. Putra, and Y. M. Putra, "Faktor-Faktor yang mempengaruhi angka partisipasi perguruan tinggi pada 32 provinsi di Indonesia tahun 2013-2016," *J. Anggar. dan Keuang. Negara Indones.*, pp. 15–34, 2019.

[6] R. A. Johnson and D. W. Winchern, *Applied multivariate statistical analysis*. New Jersey: Prentice Hall Inc., 1982.

[7] B. Everitt, *Cluster analysis*. London: Heinemann Education Books, 1974.

[8] L. Rahmawati, "Analisis kelompok dengan menggunakan metode hierarki untuk pengelompokan Kabupaten/Kota di Jawa Timur berdasarkan indikator kesehatan," *S1 Thesis*, 2013.

[9] S. Yulianto and K. H. Hidayatullah, "Analisis klaster untuk pengelompokan Kabupaten/Kota di Provinsi Jawa Tengah berdasarkan indikator kesejahteraan rakyat," *J. Stat.*, pp. 56–63, 2014.

[10] A. N. Fathia, "Analisis klaster kecamatan di kabupaten Semarang berdasarkan potensi desa menggunakan metode ward dan single linkage," *J. Gaussian*, pp. 801–810, 2016.

[11] Sukmawati, "Analisis cluster dengan metode hierarki untuk pengelompokkan Kabupaten/Kota di Provinsi Sulawesi Selatan berdasarkan indikator makro ekonomi," *S1 Thesis*, 2017.

[12] W. A. Talakua, Z. A. Leleury, and A. W. Taluta, "Analisis cluster dengan menggunakan metode k-means untuk pengelompokkan Kabupaten/Kota di provinsi maluku berdasarkan indikator indeks pembangunan manusia tahun 2014," *J. Ilmu Mat. dan Terap.*, pp. 119–128, 2017.

[13] R. Silvi, "Analisis cluster dengan data outlier menggunakan centroid linkage dan k-means clustering untuk pengelompokan indikator HIV/AIDS di Indonesia," *MANTIK J. Mat.*, pp. 22–31, 2018.

[14] M. Paramadina, Sudarmin, and M. K. Aidid, "Perbandingan Analisis Cluster Metode Average Linkage dan Metode Ward (Kasus: IPM Provinsi Sulawesi Selatan)," *VARIANSI J. Stat. Its Appl. Teach. Res.*, pp. 22–31, 2019.

[15] S. Laeli, "Analisis cluster dengan average linkage method dan ward's method untuk data responden nasabah asuransi jiwa unit link," *S1 Thesis*, 2016.

[16] J. E. Jackson, *A user's guide to principal components*. New York: John Wiley & Sons Inc., 1991.

[17] A. Farahani, H. Rahiminezhed, A. L. Same, and K. Immannezhed, "A Comparison of Partial Least Square (PLS) and Ordinary Least Square (OLS) regressions in predicting of couples mental health based on their communicational patterns," *Procedia Soc. Behav. Sci.*, pp. 1459–1463, 2010.

[18] I. Ghozali, *Analisis multivariate dengan program SPSS Edisi Ke 4*. Semarang: Badan Penerbit Universitas Diponegoro, 2006.

[19] I. Ghozali, *Aplikasi analisis multivariate dengan program SPSS*. Semarang: Badan Penerbit Universitas Diponegoro, 2011.

[20] D. N. Gujarati and D. C. Porter, *Dasar–dasar ekonometrika*. Jakarta: Salemba Empat, 2012.

[21] T. P. Ryan, *Modern regression methods*. New York: John Wiley & Sons, 1997.

[22] T. Naes and H. Martens, "Comparison of prediction methods for multicollinearity data," *Commun. Stat. Comput.*, pp. 545–576, 1985.

[23] S. Santoso, *Buku latihan SPSS statistik multivariat*. Jakarta: PT Elex Media Komputindo, 2002.

[24] D. Rachmatin, "Aplikasi metode-metode agglomerative dalam analisis klaster pada data tingkat polusi udara," *Infin. J. Ilm. Progr. Stud. Mat. STKIP Siliwangi Bandung*, pp. 133–149, 2014.

[25] R. A. Johnson and D. W. Winchern, *Applied multivariate statistical analysis, Fifth Ed.* New Jersey: Prentice Hall International Inc., 2002.

[26] W. R. Dillon and M. Goldstein, *Multivariate analysis*. New York: John Wiley & Sons, 1984.

[27] W. J. Bunkers, J. R. Miller, and A. T. DeGaetano, "Definition of climate regions in the nothern plains using an objective cluster modification technique," *J. Clim.*, pp. 130–146, 1996.

[28] H. T. Clifford and W. Stepenson, *An Introduction to Numerical Classification*. New York: Academic Press, 1975.

[29] C. J. Krebs, *Ecological Methodology*. New York: Harper and Row Publishers, 1989.

[30] J. F. Hair, R. E. Anderson, R. Thatam, and W. Black, *Multivariate data analysis with readings*. New Jersey: Pearson New International Edition, 1995.

[31] M. Kent and P. Coker, *Vegetation description and analysis: A Practical approach*. London: CRC Press & Belhaven Press, 1997.