Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/110311

Cluster Analysis Study on Various Cluster Validity Indexes with Various Linkages and Euclidean Distance (Study on Compliant Paying Behavior of Bank X Customers in Indonesia 2021)

Adji Achmad Rinaldo Fernandes[*], *Solimun Solimun, Erlinda Citra Lucki Efendi, Ni Made Ayu Astari Badung and Endang Krisnawati*

Department of Statistics, Brawijaya University, Malang, Indonesia

Received: 6 Oct. 2021, Revised: 12 Nov. 2021, Accepted: 19 Dec. 2021
Published online: 1 Sep. 2022

Abstract: This study aims to examine and explain the differences in the use of various cluster validity indices in the application of KPR customer grouping at Bank X Malang City with various linkages and euclidean distances. The data used in this study are primary data. The variables used in this study are as follows: quality of service, environment, fashion, willingness to pay, and obedient behavior to pay at Bank X. The object of observation is the customer as many as 100 respondents using purposive sampling technique. Data analysis was carried out quantitatively using cluster analysis with 4 cluster validity indices, namely the Silhouette, Krzanowski-Lai, Dunn, and Davies-Bouldin indexes. The results showed that the difference in the validity index did not affect the number of clusters produced. In addition, differences in the linkage method used affect the number of clusters generated as well as the variance within and between clusters. This study also shows that the best method to examine and explain the differences in the use of various cluster validity indices in the application of grouping mortgage customers at Bank X Malang City is the hierarchical clustering technique with the average linkage method with the Davies-Bouldin index value ranging from 0.47 to 0.55. The novelty in this study is to compare 4 validity indices, namely the Silhouette Index, Krzanowski-Lai, Dunn, Davies-Bouldin on various linkage methods, including ward, average, and complete linkage simultaneously.

Keywords: Cluster analysis, Cluster validity, Euclidean distance, Linkage method.

1 Introduction

Along with the development of the times, all areas of life are inseparable from data. Risk assessment, forecasting an event, and even making decisions can use data. Statistics is a science that is useful for processing data so that information is obtained in all fields. The statistics in this study are used to analyze banking cases in Indonesia. The term bank is known as a financial institution that collects and distributes funds from the public. [1]. One of the bank's services is credit. Giving money based on an agreement to pay off debt after a certain period of time, this is stated in Law no. 10 of 1998. Evaluation of the bank in knowing the level of responsibility of the debtor in paying credit, because credit problems are a customer who has a poor compliance behavior so that the customer only delays payment even more than the due date. From these problems, it is necessary to have supervision in credit, one of the statistical analyzes that can be used on credit problems is cluster analysis.

Classification of objects into relatively homogeneous groups and sub-samples within each group tends to be similar to one another and far different and even tends to be different from objects from other clusters called cluster analysis In cluster analysis, several linkages can be used to form clusters [2]. Clustering in statistical analysis is different from modeling. Clustering aims to deal with the problem of high heterogeneity, while flexible modeling (nonparametric regression) captures various forms of relationship patterns [3][4][5][6]. This study uses the Euclidean distance in measuring the distance between objects, then in determining the number of research clusters using a cluster validity index. In this study, we want to know whether there is indeed a difference in the use of the cluster validity index with the linkage method in clustering Bank X customers.

[*] Corresponding author e-mail: fernandes@staff.ub.ac.id

20 rice cultures from the Regional Agricultural Research Station (RARS), Pattambi under Kerala Agricultural University, Kerala. The main observations taken for this study were plant height (cm), tillers per plant, fine grain per panicle, husks per panicle, grain yield (kg/ha), and straw yield (kg/ha). Hierarchical and non-hierarchical clustering techniques were performed using Euclidean distance as a measure of similarity. There are seven different classification methods used for comparative studies. The results of this study are hierarchical clustering technique, average linkage was found to be the best with Davies-Bouldin index values ranging from 0.47 to 0.55.

2 Literature Review

2.1 *Score Interpretation Criteria*

Measurement of the variables is based on each indicator or question item following the dimensions of the construct built based on the theories and research results. Description of research variables aims to determine the frequency distribution of respondents' answers. The frequency distribution is obtained from the tabulation of the respondent's answer score. The following is the basis for the interpretation of the scores shown in Table 1.

Table 1: Average Score Criteria.

No.	Average Score Criteria	Criteria
1	1.00 – 1.5	Very Low/Very Weak
2	1.5 > - 2.5	Low/Weak
3	2.5 > - 3.5	Moderate
4	3.5 > - 4.5	High/Good
5	4.5 >	Very High/Very Good

Source: Solimun et al. [6]

2.2Cluster Analysis

Cluster analysis (grouping-analysis) is a method of analysis that aims to group objects into several groups, the objects in the group are homogeneous (same) while other group members are heterogeneous (different) [6].

The procedure for group formation in cluster analysis is divided into two, namely hierarchical and non-hierarchical methods. Hierarchical grouping is used when there is no information on the number of clusters. The main principle of the hierarchical method is to group objects that have something in common with one group. Meanwhile, the non-hierarchical method is used when information about the number of clusters is known or determined [7].

2.3 Hierarchical Method

In the hierarchy method grouping begins with grouping two or more objects that have the same thing. Then, the process is continued by forwarding it to another object that has a second closeness. And so on so that we get a tree in which there is a hierarchy or level from the most similar to the different [7]. This tree can provide more clarity in the grouping process or what is commonly known as a dendrogram.

According to Johnson and Wichern [7] in the method of forming groups in the hierarchical method, there are two approaches, namely the agglomerative hierarchical method and divisive hierarchical methods. The Agglomerative Method begins by assuming that each object is a cluster. Then the two objects that have the closest distance are made into one cluster. The process continues so that in the end a cluster consisting of all objects will be formed. The method that is often used is the agglomerative hierarchy method. Several algorithms for the agglomerative hierarchy method used in group formation in this study are the ward, complete, average linkage method.

2.3.1 Ward Linkage Method

The ward linkage method is defined as a grouping method by combining objects, where this method uses the number of squared deviations for each object. This method works by combining two objects when they have a small distance.

$$SSE=\sum_{j=1}^{p}\left(\sum_{i=1}^{n}\varepsilon_{ij}^{2}-\frac{1}{n}\left(\sum_{i=1}^{n}\varepsilon_{ij}^{2}\right)^{2}\right) \tag{1}$$

where:

ε_{ij} : the value for the i object of the j cluster
p : the number of variables
n : the number of respondents in the cluster that was formed

2.3.2 Average Linkage Method

The average linkage method considers the distance between two clusters to be the average distance between all members in one cluster and the other cluster. The distance formula can be written in equation (2).

$$d_{(ij)k} = \frac{\sum_i \sum_j d_{(ij)}}{N_{ij} N_k} \tag{2}$$

where:
$d_{(ij)k}$: the distances between the subsample (ij) and the cluster k
d_{ik} : the distances of subsample i and cluster k
d_{jk} : the distances of sub-sample j and cluster k

2.3.3 Complete Linkage Method

In the Complete linkage method, the distance between clusters is determined by the farthest distance between two objects in different clusters [7].

$$d_{(ij)k} = max\left(d_{ik}, d_{jk}\right) \tag{3}$$

where:
$d_{(ij)k}$: the distances between the subsample (ij) and the cluster k
d_{ik} : the distances of sub-sample i and cluster k
d_{jk} : the distances of sub-sample j and cluster k

2.4 Distance in Cluster Analysis

In this study, the number of clusters used the Euclidean distance. The distance between two points is calculated using the formula (4)

$$d\left(x_i, x_j\right) = \sqrt{\sum_{z=1}^{p}\left(x_{ki} - x_{kj}\right)^2} \tag{4}$$

di mana:
$d\left(x_i, x_j\right)$: the Euclidean distance between the i object and the j object
x_{ki} : the value of the i object in the variable k
x_{kj} : the value of the j object in the variable k
z : the variables to z, $z = 1,2,3, \dots, p$

2.5 Validity Index of Cluster Analysis

a. Silhouette Index

The silhouette validity index is a statistical measure used to solve the problem of determining the optimal number of K clusters which can provide a brief graphic representation of how well each object is located in the cluster. The silhouette index will evaluate the placement of each object in each cluster by comparing the average distance of objects in one cluster and the distance between objects with different clusters [8]. The silhouette method provides information in the form of graphics in determining the optimal number of clusters. Determination of the optimal number of clusters is done by looking at the maximum average value of silhouette *S(i)*. The optimal number of K clusters is an estimate of the price that maximizes the average value of the silhouette validity index *S(i)*.

Assume the data has been grouped into clusters. For each object, let *a(i)* be the average distance of objects to all objects in the same cluster and b (i) is the average minimum distance of object i to all objects in a cluster that is not a cluster member. From the explanation that has been presented, the silhouette validity index can be written with the following equation:

$$S(i) = \frac{b(i) - a(i)}{Max\,\{a(i), b(i)\}} \tag{5}$$

Where:

$a(i)$: the average difference of i objects with all other objects in the same group
$\mathrm{b}(i)$: the minimum value of the mean difference of i objects with all objects in other groups (in the closest group)

The silhouette validity index equation can be written as follows:

$$S_{(i)} = \begin{cases} 1 - \frac{a(i)}{b(i)}, if\ a(i) < b(i) \\ 0, \quad if\ a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, if\ a(i) > b(i) \end{cases} \tag{6}$$

The average $S_{(i)}$ of all objects in a cluster shows how closely the objects in a cluster are similar, which also indicates how precisely the objects have been grouped. The closer $S_{(i)}$ to 1, the better the grouping of objects. Conversely, the closer $S_{(i)}$ it is to -1, the worse the grouping of objects will be. The optimal number of clusters is an estimate of the price that most maximizes the average value of $S_{(i)}$ and if there is one cluster whose members consist of one object, the average value $S_{(i)}$ is 0.

b. *Krzanowski-Lai* Index

The optimal number of clusters is then defined as the k value that maximizes CH (k). The *Krzanowski-Lai* index are determined by:

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right|, \tag{7}$$

Where:

$$DIFF(k) = (k-1)^{2/p} W(k-1) - (k)^{2/p} W(k) \tag{8}$$

Optimization of the value of k is done if KL(k) is maximized, this is done to compare the performance of the three indexes BB, CH, and KL based on the results obtained.

c. *Dunn* Index

The Dunn index was introduced by J. C. Dunn (1973) as a metric for cluster validity. Dunn index is defined as the ratio of the smallest distance between clusters and the largest distance in each data cluster, by comparing d_{min} with d_{max}.
Where:

$$C = \frac{d_{min}}{d_{max}} \tag{9}$$

d_{min}= The smallest distance between objects in different clusters
d_{max}= The largest distance in the cluster

d. *Davies-Bouldin* Index

The Davied-Bouldin index formula can be written as:

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{i \neq j} \left[\frac{d`(c_i) + d`(c_j)}{d(c_i, c_j)} \right] \tag{10}$$

Where:
n = number of groups
d(c_i,c_j) = the distance between groups c_i and c_j
d'(c_k) = the distance in groups c_k

2.6 Operational Definition of Service Quality Variables

Service quality is a way of working for a company that seeks to make continuous quality improvements to the processes, products, and services that the company produces. Service quality is also defined as an effort to meet consumer needs and the accuracy in balancing the expectations of consumers. According to research by Parasuraman et. al., [6] five indicators can measure service quality, namely: 1) Reliability; 2) Responsiveness; 3) Assurance; 4) Empathy; 5) Tangibles.

Reliability is the ability to provide the promised service appropriately and the ability to be trusted, especially in providing services. Responsiveness is the ability to help what consumers need quickly, precisely, and responsively. Assurance is the ability to eliminate customer doubts and make them feel exogenous from dangers and risks. Empathy is the company's ability to understand consumer needs and ease of communication or relationships. Tangibles are the availability of physical facilities, equipment, and others in a company.

2.7 Operational Definition of Environmental Variables

According to Simamora [10], the work environment is the internal/psychological environment of the company and the human resource policies accepted by company employees. According to Carr's research [11], three dimensions can measure the work environment, namely: 1) Physical Work Environment; 2) Temporary Work Environment; 3) Psychological Work Environment.

Physical Work Environment measured through six indicators, namely: lighting, use of color, air circulation, noise, cleanliness, and safety. Temporary Work Environment, measured by two indicators, namely: working hours and rest periods. Psychological Work Environment, measured through three indicators, namely: boredom, fatigue, and work relations.

2.8 Operational Definition of Fashion Variables

Fashion is a model, method, style, or form of habit. Fashion is not only related to clothing styles, but there are also relationships with cosmetic styles, accessories, hairstyles, etc. to support one's appearance.

According to Karlyle; “Fashion is a symbol of the soul. Clothing cannot be separated from the development of human life history and culture. In other words, clothing can be interpreted as a social skin that contains messages and also the way of human life".

The benefits of fashion in everyday life include providing self-confidence for women where psychologically, every woman who looks attractive and comfortable has more confidence than women who look unattractive. Besides, fashion can give its charm, especially in connection with politeness and friendliness there will be an attractive charisma. Fashion can also make you happy because there is a feeling of satisfaction with fashion that is a concern. The indicator of the fashion variable:

2.9 Operational Definition of Willingness to Pay Variable

Willingness to pay is the maximum price of an item that consumers want to buy at a certain time. On the other hand, willingness to pay can be interpreted as the willingness of the community to accept the burden of payment, according to the amount that has been determined. Willingness to pay is important to protect consumers from the dangers of corporate monopoly related to price and product supply quality [12].

2.10 Operational Definition of Compliant Paying Behavior Variable

Customer compliance is the customer has the willingness to fulfill his debt obligations following applicable regulations without any investigation, joint investigation, warning, and application of sanctions both legally and administratively [13], customer actions in fulfilling their debt obligations following regulatory provisions between customers and the leasing party or bank. Based on this theory, it can be concluded that customer compliance is the customer's action in fulfilling their debt obligations according to the previously agreed regulations and is willing to accept sanctions if they do not comply. According to Law No. 6 of 1983, customer compliance can be measured through: timeliness, data accuracy, and sanctions.

3 Methodology

This study uses primary data which is then analyzed further. The variables used in this study are service quality, environment, fashion, willingness to pay, and behavior in paying compliance at Bank X. The instrument of this research used a questionnaire with a Likert scale. Sampling was done by purposive sampling method. The research sample is 100 customers. Selection of a sample of 100 customers because it follows the central limit theory which says that the sampling distribution curve (for a sample size of 30 or more) will center on the value of the population parameter and will have all the characteristics of a normal distribution.

This study uses a quantitative approach, namely descriptive analysis, then cluster analysis using the ward, average, and complete linkage and euclidean distance methods on various cluster validity indices, namely the Sillhouette, Krzanowski-Lai, Dunn, and Davies-Bouldin indexes using software R.

4 Results

The first step is to get clusters for each validity index using several linkage methods. The results of the number of cluster members obtained from various validity indices with various linkage methods can be seen in Table 2.

Table 2: Number of Clusters for Each Index.

Validity Index	Linkage Method	Cluster 1	Cluster 2
Silhouette	Ward	36	64
	Average	42	58
	Complete	52	48
Krzanowski-Lai	Ward	36	64
	Average	42	58
	Complete	52	48
Dunn	Ward	36	64
	Average	42	58
	Complete	52	48
Davies-Bouldin	Ward	36	64
	Average	42	58
	Complete	52	48

After getting the clusters and their members, then the average was looked for to find out the differences in the members of each index and variable. The average results obtained can be seen in Table 3.

Table 3: The Average Cluster Members for Each Index.

Index	Distance	Average									
		X1		X2		X3		Y1		Y2	
		C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
Silhouette	Ward	4.02	3.20	4.05	3.19	4.06	3.22	4.07	3.15	4.00	3.20
	Average	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
	Complete	3.75	3.22	4.01	2.95	3.72	3.32	3.82	3.11	3.84	3.10
Krzanowski-Lai	Ward	4.02	3.20	4.05	3.19	4.06	3.22	4.07	3.15	4.00	3.20
	Average	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
	Complete	3.75	3.22	4.01	2.95	3.72	3.32	3.82	3.11	3.84	3.10
Dunn	Ward	4.02	3.20	4.05	3.19	4.06	3.22	4.07	3.15	4.00	3.20
	Average	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
	Complete	3.75	3.22	4.01	2.95	3.72	3.32	3.82	3.11	3.84	3.10
Davies-Bouldin	Ward	4.02	3.20	4.05	3.19	4.06	3.22	4.07	3.15	4.00	3.20
	Average	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
	Complete	3.75	3.22	4.01	2.95	3.72	3.32	3.82	3.11	3.84	3.10

It can be seen from Table 3., Most of the customers in cluster 1 consider service quality, service quality, environment, fashion, willingness to pay, and compliance behavior of Bank X customers throughout Indonesia as good on all validity indexes with various linkage methods. In cluster 2, most customers consider that service quality, service quality, environment, fashion, willingness to pay, and compliance with paying behavior of Bank X customers throughout Indonesia is sufficient on all validity indices with various linkages. Besides, it can be seen from table 2, which has the highest average for all variables is the ward linkage cluster 1 method, meaning that most customers in cluster 1 obtained from the Ward linkage method show that service quality, environment, fashion, willingness to pay, and compliance behavior is good. After the ward linkage method, the average linkage method and the complete linkage method are followed.

Then choose the validity index and the best linkage method by calculating the variance within groups and between groups, then comparing the results that have the variance in the smallest group and the variance within the largest group is the validity index and the best linkage method. The results of the comparison of the validity index and the linkage method can be seen in Table 4.

Table 4: The Variance Within and Between Groups of Each Index and Linkage Method**.**

Index	Linkage	Variance Within Cluster		Variance Between Cluster
		Cluster 1	Cluster 2	
Silhouette	Ward	13.70	6.51	7.18
	Average	10.83	6.54	4.29
	Complete	8.00	6.56	1.43
Krzanowski-Lai	Ward	13.70	6.51	7.18
	Average	10.83	6.54	4.29
	Complete	8.00	6.56	1.43
Dunn	Ward	13.70	6.51	7.18
	Average	10.83	6.54	4.29
	Complete	8.00	6.56	1.43
Davies-Bouldin	Ward	13.70	6.51	7.18
	Average	10.83	6.54	4.29
	Complete	8.00	6.56	1.43

It can be seen from Table 4., the difference in linkage methods gives different results, it can be seen that the number of cluster members obtained has the same number, besides that the variance between and within groups gives the same results.

The different validity indices in each linkage method give the same results, so it can be concluded in this study that the difference in validity index does not make a difference in the variance within and between clusters. However, the different linkage methods used can affect the number of members of each cluster and give different results on the variance within and between clusters.

The results of this study are consistent with research conducted by Adarsh et al. [15], this study aims to study a comparative study of various clustering techniques and methods in a small-scale sample. The results of this study are hierarchical clustering technique, average linkage was found to be the best with Davies-Bouldin index values ranging from 0.47 to 0.55.

5 Conclusions

The conclusion that can be given is based on the results of the analysis, namely.

1. Cluster analysis can be applied to classify service quality, environment, fashion, willingness to pay, and compliance to pay behavior of bank X customers in Indonesia.
2. The difference in cluster validity index gives the same effect.

The difference in linkage methods gives the results that the number of members of each cluster is different and the differences in the variance between and within clusters.

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] Kasmir, *Analisis Laporan Keuangan*, Jakarta: Raja Grafindo Persada, (2011).

[2] Fernandes, AAR, & Solimun, S. The Effect Of Correlation Between Responses In Bi-Response Nonparametric Regression Using Smoothing Spline For Longitudinal Data. *Communications In Applied Analysis*, **20**, (2016).

[3] Fernandes, A. A. R., & Cahyoningtyas, R. A. Structural equation modelling on Latent Variables to identify farmers satisfaction in East Java using Mixed-Scale Data. *In Journal of Physics: Conference Series*, **1872**, 012022, (2021, May).

[4] Fernandes A.A.R. and Solimun, Multi-responses model in patients suffering from decubitus wound using generalized penalized Spline. *International Journal of PharmTech Research*, **9**, 488–497. (2016).

[5] Fernandes. A. A. R., Astuti. A. B., Amaliana. L., Yanti. I., Arisoesilaningsih. E., and Isaskar. R. Smoothing Spline Nonparametric Path: Application for Green Product and Green Marketing Strategy towards Green Product Purchasing Intention. *In IOP Conference Series: Earth and Environmental Science*. **239**, 012018, (2019, February)

[6] Solimun, Fernandes. A. A. R., and Nurjannah, *Metode Statistika Multivariat Pemodelan Persamaan Struktural (SEM) Pendekatan WarPLS*, Malang: UB Press. (2017).

[7] Johnson. R. A. and Wichern. D. W, *Applied Multivariate Analysis*, Upper Saddle River, NJ: Prentice Hall, (1992).

[8]Aini. F. N, *Clustering Business Process Model Petri Net Dengan Complete Linkage*, (2014).

[9] Parasuraman, A., Valarie A. Zeithaml, and Leonard L. Berry, *Delivering Quality Service : Balancing Customer Perceptions*, (1990).

[10] Simamora. Henry, *Manajemen Sumber Daya Manusia*, Yogyakarta: STIE YKPN, (1997).

[11] Carr. A, *Positive Psychology; The Science of Happiness and Human Strengs*, New York: Brunner Routledge, (2004).

[12] Grace Laumahina and Njo Anastasia, Kesediaan untuk Membayar pada Green Residential, *FINESTA*, **2**, 82-86, (2014).

[13] Gunadi, *Panduan Komprehensif Pajak Penghasilan*, Jakarta: Bee Media Indonesia, (2013).

[14] Fernandes, A. A. R., Solimun, F. U., Aryandani, A., Chairunissa, A., Alifa, A., Krisnawati, E., ... & Rasyidah12, F. L. N., Comparison Of Cluster Validity Index Using Integrated Cluster Analysis With Structural Equation Modelingthe War-Pls Approach, *Journal of Theoretical and Applied Information Technology*, **99**, (2021).

[15] Adarsh, V. S., Joseph, B., & Gopinath, P. P., *Comparison of clustering techniques used in divergence analysis*, (2020).