Received: October 29, 2020. Revised: May 22, 2021. Accepted: June 12, 2021. Published: June 24, 2021.

Cluster Analysis in Various Cluster Validity Indexes With Average Linkage Method At Various Distance (Study on Compliant Paying Behavior of Bank X Customers in Indonesia 2021)

Adji Achmad Rinaldo Fernandes[1*], Solimun[2]
Department of Statistics
Brawijaya University
St. Veteran, No. 1, Malang
Indonesia

Abstract: - This study aims to examine and explain the differences in the use of various cluster validity indices and various distances in the application of KPR Bank X Indonesia customer grouping using the average linkage method. The data used in this study are primary. The variables used in this study are as follows: service quality, environment, fashion, willingness to pay, and obedient behavior to pay at Bank X. Data obtained through a questionnaire with a Likert scale. Measurement of variables in primary data using the average score of each item. The sampling technique used was purposive sampling. The object of observation is the customer as many as 100 respondents. Data analysis was performed quantitatively, cluster analysis was carried out using the average linkage method and Euclidean, Manhattan, and Minkowski distances on various cluster validity indices, including Gap Index, Index C, Global Sillhouette, and Goodman-Kruskal in this study used as analysis tools. This research uses R software. The results show that the Gap index, C-Index, Global Sillhouette, and Goodman-Kruskal at the same distance have the same cluster members, but at different distances produce different clusters. The novelty in this study is to compare 4 validity indices at various distance types were Euclidean, Manhattan, and Minkowski.

Key-Words: - Cluster Analysis, Cluster Validity, Euclidean Distance, Average Linkage Method

I. INTRODUCTION

In today's modern era, almost all areas of life are inseparable from data. Data can be used to assess risk, predict an event, and even be used for decision making. Statistics, which is known as the science that is useful for processing data into information, has an important role in all areas of life. One of the fields that still really need statistics is banking. According to Kasmir [1], a bank can be defined as a financial institution whose business activities are collecting funds from the public and channeling these funds back to the public, and providing other bank services.

One of the services provided by banks is credit. According to Law no. 10 of 1998 Credit is a provision of money or claims based on a loan agreement or agreement between a bank and another party that requires the borrower to pay off the debt after a certain time with interest. Before a bank provides credit to a debtor, it is necessary to have an assessment from the bank to measure whether the debtor can fulfill his obligations in credit or not. One of the credit problems is that there are customers who have compliant, neutral, and even non-compliant paying behavior. From these problems, of course, there needs to be supervision in terms of credit, one of the statistical analyzes that can be used in this problem is cluster analysis.

Cluster analysis is used to classify samples into relatively homogeneous subsamples called clusters, the subsamples in each group tend to be similar to each other and differ greatly (not the same) from objects from other clusters. In cluster analysis, several linkages can be used to form clusters. According to Supranto [2], the linkage method consists of single, complete, and average linkage. This study examines the application of the average

linkage method on four types of cluster validity indexes with various distances. Measuring the distance of each linkage in this study using the euclidean, Manhattan, and Minkowski distances. The results of determining the number of clusters with the cluster validity index at different distances will certainly give different results. In this study, we want to examine the effect of using the cluster validity index with various distances on the problem of customer grouping in Bank X.

This study refers to a study conducted by Thant [3] entitled "Euclidean, Manhattan and Minkowski Distance Methods For Clustering Algorithms." The purpose of this study is to implement the inequality matrix for interval scale variables using the Euclidean, Manhattan, and Minkowski distance method. The research method used is the Euclidean, Manhattan, and Minkowski distance method. The result of this study is that the Minkowski distance is a generalization: if q = 2, d is the Euclidean distance, and if q = 1, d is the Manhattan distance.

II. LITERATURE REVIEW

A. *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. The purpose of the research variable description which is part of the descriptive statistical analysis is to determine the frequency distribution of respondents' answers to the distributed questionnaires and to describe in-depth the variables in this study. 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 2.1.

Table 2.1 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. [4]

B. *Cluster Analysis*

Cluster analysis (group 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) [4].

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 [5].

C. *Cluster Analysis*

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 [5]. This tree can provide more clarity in the grouping process or what is commonly known as a dendrogram.

According to Johnson and Wichern [5] 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. One of the agglomerative hierarchy method algorithms used in group formation in this study is the average linkage method.

D. *Average Linkage Method*

In the Average linkage method, the distance between two clusters is considered as the average distance between all members in one cluster and all members of the other clusters. The distance formula can be written in equation (2.1).

$$d_{(ij)k} = \frac{\sum_i \sum_j d_{(ij)}}{N_{ij} N_k} \qquad (2.1)$$

Where:

$d_{(ij)k}$: the distances between the subsample (ij) and the cluster k

d_{ik} : the distance of subsample i and cluster k

d_{jk} : the distance of sub-sample j and cluster k

E. *Distance in Cluster Analysis*

a. Euclidean distance

In this study, the number of clusters used the Euclidean distance. The distance between two points is calculated using the formula (2.2)

$$d(x_i, x_j) = \sqrt{\sum_{z=1}^{p}(x_{ki} - x_{kj})^2} \quad (2.2)$$

where:
$d(x_i, x_j)$: 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 variable to z, $z = 1,2,3, \dots, p$

b. Manhattan distance

In this method, distance measurements are made by calculating the absolute number of differences in objects for each variable. This method is called absolute block or better known as city block distance. Manhattan distance can be formulated as follows:

$$d_{i,k} = \sum_{j=1}^{n}(x_{ij} - x_{kj}) \quad (2.3)$$

Where:
$d_{i,k}$: the distance between object i and j
x_{ij} : the value of object i in variable j
x_{kj} : the value of the k object in the variable j
n : the number of variables observed

c. Minkowski distance

Minkowski Distance is a distance comparison method that is a metric in vector space where a norm is defined as well as a generalization of Euclidean distance and Manhattan distance. The Minkowski distance was discovered [3]. Minkowski distance is a general form of the formula for calculating distance space or the distance between two points. The factor that distinguishes Minkowski Distance from other distance space calculations lies in the exponential value (p) used in the calculation of distance space. In measuring the distance of an object using the Minkowski distance, the p-value is usually 1 or 2.

Formula:

$$d(x, y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p} \quad (2.4)$$

Where:
d = distance between x and y
x = cluster center data
y = attribute center data
i = each data
n = amount of data
x_i = data at the center of the cluster i
y_i = data on each data ith
p = power

F. Validity Index in Cluster Analysis

a. Gap Index

One way to estimate the optimal number of clusters is to use a gap statistic [6]. Suppose that X_{ij} is an observation on the ith object and the j-th variable. Then, do a cluster analysis on the data into k clusters, $C_1, C_2, \dots, C_3$ to C_r are the observations in the r-cluster and n_r is the number of objects in the r-cluster, so that it can be defined as follows:

$$D_r = \sum_{(ik,jk)} d_{ik} \quad (2.5)$$

where D_r is the total distance of all points in the cluster r and d_{ik} is the distance between the i object and the k object.

$$W_k = \sum_{r=1}^{k} \frac{1}{2_{n_r}} D_r \quad (2.6)$$

where W_k is the sum of the squares combined in the cluster.

b. C-Index

This index can be explained as follows:

$$C = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (2.7)$$

Where:
S : the number of distances in all pairs of observed objects from the same group, with ℓ is the number of pairs.
S_{min} : the sum of the ℓ smallest distances if all sample pairs are in different groups.
S_{max} : the sum of the ℓ greatest distances of all pairs .

The smaller C value indicates that a good group will be [7].

c. *Global Sillhouette* Index

To get the Silhouette S (i) index the following formula is used:

$$S(i) = \frac{(b(i) - a(i))}{max\{a(i), b(i)\}} \quad (2.8)$$

where
a(i) : the average difference of the i-object with all other objects in the same group.
b(i) : the minimum value of the mean difference of i-objects with all objects in other groups (in the closest group).

The greatest value from the Global Silhouette Index marks the number of the best groups which are then taken as the optimum group. The Global Silhouette formula is given by:

$$GS_u = \frac{1}{n}\sum_{i=1}^{n} S(i) \quad (2.9)$$

where

S(i) = *Silhouette* of *i* group

n = number of groups

d. *Goodman-Kruskal* (GK) Index

Suppose that the four pairs of all observed objects are (q, r, s, t), where *d(x,y)* is the distance between the object x and y. The four pairs of objects are said to be concordant if they meet the conditions *d(q,r) < d(s,t)*, where q and r are in the same group and s and t are in different groups. Conversely, four pairs of objects are said to be discordant if they meet the following conditions: *d(q,r) < d(s,t)* where q and r are in different groups and s and t are in the same group.

The GK index is calculated from the calculation of the value of the concordant and discordant pairs with the formula:

$$GK = \frac{S_c - S_d}{S_c + S_d} \quad (2.10)$$

di mana

S_c = number of concordant pairs

S_d = number of discordant pairs

Large values indicate the optimum group. All calculated indices can give the optimum number of groups, but each index can give different results. Provides an alternative to choosing the optimum number of groups by combining the group validity index, which can then be selected for the optimum number of groups when the index is the most combined. The step is to calculate the five validity indices, then rank each possible number of groups on each index. The optimum number of groups is obtained at the highest average ranking [7].

G. 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 an effort to fulfill the needs and desires of consumers as well as the accuracy of its delivery in balancing consumer expectations. According to research by Parasuraman [8] five indicators can measure service quality, namely: 1) Reliability; 2) Responsiveness; 3) Assurance; 4) Emphaty; 5) Tangibles.

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

H. Operational Definition of Environmental Variables

According to Simamora [9], the work environment is the internal/psychological environment of the company and the human resource policies accepted by company employees. According to Carr's [10], three dimensions can measure the work environment, namely: 1) Physical Work Environment; 2) Temporary Work Environment; 3) Psychological Work Environment.

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

I. 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:

• Activities

According to KBBI, an activity is an activity. Activities in fashion relate to clothing design work, footwear design and other fashion accessories design, making patterned clothes and accessories, mid-fashion dialogues, and fashion product distributors.

• Interests

The definition of interest according to language (Etymology), is the effort and willingness to learn (learning) and look for something. In terms (terminology), interest is the desire, liking, and willingness to something. Thus, interest is a feeling of wanting and liking something.

• Opinions

In general, the notion of opinion is an opinion, response, view, or the result of a person's thoughts in explaining or addressing a matter but its nature is not objective and the truth is uncertain. Opinions are subjective and everyone may have different opinions about an event or object.

J. Operational Definition of Willingness to Pay Variable

According to Zhao and Kling [11], 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].

According to research by Priambodo and Najib [13], 4 indicators can measure a person's willingness to buy, namely:

1. Options
   Indicators that aim to choose to pay credit on the Bank BNI people's housing credit assessment (KPR).
2. Benefits
   The added value of an expected outcome in using Bank BNI KPR.
3. Sacrifice
   An act of giving up something sincere and sincere moral awareness in using Bank BNI KPR.
4. Be consistent
   Payments made on public housing loans repeatedly from time to time, to be fair and accurate.

K. 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 [11], customer actions in fulfilling their debt obligations following regulatory provisions between customers and the leasing party or bank [14]. 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.

III. METHODOLOGY

The data used in this study are primary. The variables used in this study are as follows: service quality, environment, fashion, willingness to pay, and obedient behavior to pay at Bank X. Data obtained through a questionnaire with a Likert scale. Measurement of variables in primary data using the average score of each item. The sampling technique used was purposive sampling. The object of observation is the customer as many as 100 respondents. 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.

Data analysis was carried out quantitatively, to explain each of the variables studied, a descriptive analysis was carried out first, then cluster analysis

was carried out using the average linkage method and various distances, including euclidean, Manhattan, and Minkowski on various cluster validity indices, including Gap, C- Index, Global Sillhouette, and Goodman-Kruskal in this study were used as analysis tools. This research uses R software.

IV. RESULT

The first step is to get a cluster for each validity index with several distances. The results of the number of cluster members obtained from various validity indices with various distances can be seen in Table 1.

Table 1. Number of Clusters for Each Index

Validity Index	Distance	*Cluster* 1	*Cluster* 2
GAP	*Euclidean*	42	58
	Manhattan	62	38
	Minkowski	42	58
C-index	*Euclidean*	42	58
	Manhattan	62	38
	Minkowski	42	58
Global Sillhouette	*Euclidean*	42	58
	Manhattan	62	38
	Minkowski	42	58
Goodman-Kruskal	*Euclidean*	42	58
	Manhattan	62	38
	Minkowski	42	58

After obtaining the cluster and its members, the average was sought to determine the differences in the members of each index and variable. The average results obtained can be seen in Table 2.

Table 2. The Average Cluster Members for Each Index with Various Distances

Index	Distance	Average									
		X1		X2		X3		Y1		Y2	
		C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
1. GAP	*Euclidean*	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
	Manhattan	3.83	2.96	3.77	3.05	3.84	3.01	3.80	2.95	3.82	2.96
	Minkowski	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
2. C-index	*Euclidean*	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
	Manhattan	3.83	2.96	3.77	3.05	3.84	3.01	3.80	2.95	3.82	2.96
	Minkowski	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
3. Global Sillhouette	*Euclidean*	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
	Manhattan	3.83	2.96	3.77	3.05	3.84	3.01	3.80	2.95	3.82	2.96
	Minkowski	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
4. Goodman-kruskal	*Euclidean*	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
	Manhattan	3.83	2.96	3.77	3.05	3.84	3.01	3.80	2.95	3.82	2.96
	Minkowski	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12

It can be seen from Table 2., most of the customers in cluster 1 consider the quality of service, service quality, environment, fashion, willingness to pay, and compliance behavior of Bank X customers throughout Indonesia to be good on all validity indices with various distance methods. In cluster 2, most of the customers considered that service quality, service quality, environment, fashion, willingness to pay, and compliance with paying behavior of Bank X customers throughout Indonesia were sufficient on all validity indices with various distance methods. As seen from Table 2. It shows that the euclidean and Minkowski distances have the same and higher mean than Manhattan distances.

Then choose the best validity index and distance method by calculating the variance within groups and between groups, then comparing the results that have the variance within the smallest group and the variance within the largest group is the validity index and the best distance method. The

results of the comparison of the validity index and the distance method can be seen in Table 3.

Table 3. The Variance Within and Between Groups of Each Different Index and Distance

Index	Distance	Variance Within *Cluster*		Variance Between *Cluster*
		Cluster 1	*Cluster* 2	
GAP	*Euclidean*	10.83903	6.541533	4.297502
	Manhattan	8.764677	7.139103	1.625575
	Minkowski	10.83903	6.541533	4.297502
C-index	*Euclidean*	10.83903	6.541533	4.297502
	Manhattan	8.764677	7.139103	1.625575
	Minkowski	10.83903	6.541533	4.297502
Global Sillhouette	*Euclidean*	10.83903	6.541533	4.297502
	Manhattan	8.764677	7.139103	1.625575
	Minkowski	10.83903	6.541533	4.297502
Goodman-kruskal	*Euclidean*	10.83903	6.541533	4.297502
	Manhattan	8.764677	7.139103	1.625575
	Minkowski	10.83903	6.541533	4.297502

It can be seen from Table 2., the difference in the distance method used 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 for each of the distance methods used to give the same results, so it can be concluded in this study that the difference in the validity index does not make a difference in the variance within and between clusters. However, differences in the distance method used can affect the number of members of each cluster and give different results on the variance within and between clusters.

V. CONCLUSION

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 application of cluster analysis with different cluster validity indices results in the same number of clusters.
3. 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.

REFERENCES

[1] Kasmir, *Analisis Laporan Keuangan*, Raja Grafindo Persada, 2011.

[2] Supranto, *Analisis Multivariat Arti dan Interpretasi*, Rineka Cipta, 2004.

[3] Thant, A. A., Aye, S. M., & Mandalay, M., *Euclidean, Manhattan and Minkowski Distance Methods For Clustering Algorithms*, 2020.

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

[5] Johnson, R.A. & Wichern, D.W., *Applied Multivariate. Analysis*. Upper Saddle River, Prentice Hall, 1992.

[6] Tibshirani, R., Walther, G., & Hastie, T., Estimating the number of clusters in a data set via the gap statistic, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 63, No. 2, 2001, pp. 411-423.

[7] Bolshakova, N., & Azuaje, F., Improving expression data mining through cluster validation, *In 4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine*, 2003, pp. 19-22.

[8] Parasuraman, A., Zeithaml, V. A., & Berry, L. L., Reassessment of expectations as a comparison standard in measuring service quality: implications for further

research, *Journal of marketing*, Vol. 58, No. 1, 1994, pp. 111-124.
[9] Simamora, H., Manajemen Sumber Daya Manusia, STIE, 1997.
[10] Carr, A., *Positive Psychology: The Science of Happiness and Human Strengths*, Brunner–Routledge, 2004.
[11] Gunadi, *Panduan Komprehensif Pajak Penghasilan*, Bee Media Indonesia, 2013.
[12] Grace Laumahina dan Njo Anastasia, Kesediaan untuk Membayar pada Green Residential, *FINESTA*, Vol. 2, No. 1, 2014, pp. 82-86.
[13] Priambono, L.H., & Najib. M., Analisis Kesediaan Membayar (Willingness to Pay) Sayuran Organik dan Faktor-Faktor yang Mempengaruhinya, *Jurnal Manajemen dan Organisasi*, Vol. 5, No. 1, pp. 1-14.
[14] Rahayu, S. K., *Perpajakan Indonesia: konsep dan aspek formal*, Graha Ilmu, 2010.