Journal of Theoretical and Applied Information Technology
15th December 2021. Vol.99. No 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMPARISON OF CLUSTER VALIDITY INDEX AND DISTANCE MEASURES USING INTEGRATED CLUSTER ANALYSIS AND PATH ANALYSIS

AISYAH ALIFA[1] , SOLIMUN[2], MARIA BERNADETHA THERESIA MITAKDA[3], ADJI ACHMAD RINALDO FERNANDES[4], WAEGO ADI NUGROHO[5]

[1,2,3,4]Brawijaya University, Faculty of Mathematics and Natural Sciences, Department of Statistics, Indonesia

E-mail: [1]Aisyhalifa@student.ub.ac.id, [2]solimun@ub.ac.id, [3]dethamitakda@yahoo.com, [4]fernandes@staff.ub.ac.id, [5]waego.hadi.nugroho@gmail.com

ABSTRACT

This study wants to compare the Integrated Cluster Analysis and Path model with various cluster validity indices and distance measures on Character, Capacity, Capital, Collateral, Condition, Intention to Pay Obedience, Punctuality of Payment of Bank X Creditors. The data used in this study are primary data. The variables used in this study are character, capacity, capital, collateral, condition, intention to pay obedience, punctuality of payment bank X creditors. The data were 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 creditor as many as 100 respondents. Data analysis was carried out quantitatively, to explain each of the variables studied, a descriptive analysis was carried out first, then an Integrated Cluster Analysis and path analysis was carried out with the average linkage method on various cluster validity indices, namely Gap, Index C, Global Sillhouette. , and Goodman-Kruskal, as well as three distance measures, namely the Euclidean, Manhattan, and Minkowski distances. This research uses R software. The integrated cluster and path analysis with the Gap Index, Index C, Global Sillhouette, and Goodman-Kruskal with the Manhattan Distance is better than the Gap, Index C, Global Sillhouette, and Goodman-Kruskal with the Euclidean and Minkowski Distance. The novelty in this research is the application of Integrated Cluster Analysis and path model approach to compare 4 cluster validity indices, namely Gap Index, C Index, Global Sillhouette, and Goodman-Kruskal, and three distance measures, namely Euclidean, Manhattan, and Minkowski distances. simultaneously.

Keywords: *Cluster Analysis; Path analysis; Integration Model; Dummy Variable; Cluster Validity Index; Distance Measures*

1. INTRODUCTION

Banks can be defined as financial institutions whose business activities are to collect funds from the public and channel these funds back to the public and provide other banking services (Kasmir, 2011). One of the services provided by banks is credit. Credit is a provision of money or claims based on a loan agreement or agreement between the bank and another party that requires the borrower to pay off the debt after a certain period 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 the existence of debtors who have non-current credit so that which can harm the bank.

Path analysis is used to analyze causal relationships in the model if exogenous variables affect endogenous variables either directly or indirectly. Cluster analysis aims to classify objects into several clusters, where between clusters have different properties. In general, there are two methods in cluster analysis, namely the hierarchical method and the non-hierarchical method. The hierarchical method consists of several methods, namely the Single Linkage method, the Average Linkage method, the Complete Linkage method, the Centroid Linkage method, and the Ward method (Ward's Method).

In cluster analysis, one of the similarity measures used is distance. The distance measure is a measure of similarity, the higher the distance value, the lower the similarity between objects. There are several methods of measuring distances, including

Euclidean, Manhattan/City Block, Mahalanobis, Correlation, Angle-based, Squared Euclidean. This study applies an integrated cluster in Path Analysis with the Euclidean distance measure. The distance measure used can determine the results of the number of clusters formed. Therefore, this study wants to obtain the best distance measure to maximize the measurement of accuracy, sensitivity, and specificity when an integrated cluster is carried out with path analysis method used to determine the relationship between variables that can be measured directly (latent variables).

In this study, researchers will compare the integrated cluster analysis model and Path analysis using the average linkage with different size distances and cluster validity indices. Therefore, distance measures (Euclidean, Manhattan, Minkowski) will be compared with four cluster validity indices, namely statistical gap, goodman kruskal, global silhouette, and index C. Cluster validity test is used to evaluate the results of quantitative Cluster Analysis to produce the optimum group. An optimum group is a group that has a dense distance between individuals in the group and is well isolated from other groups (Ambassador and Jain, 1988).

2. LITERATURE REVIEW

2.1 Cluster Analysis

According to Siswadi and Suharjo (1998), cluster analysis is a multiple variable analysis that aims to group n objects into k clusters with k <n based on p variables, so that each unit object in one cluster has more homogeneous characteristics than the object units in the cluster other. The process of cluster analysis is to classify the data by using two methods, namely the hierarchical method and the non-hierarchical method. In the hierarchical cluster analysis, it is assumed that at first, each object is a separate cluster, then the two closest objects or clusters are combined to form one smaller cluster (Johnson and Wichern, 1992). Hierarchical cluster analysis consists of two methods, namely agglomerative and divisive. In the agglomerative method, each object is considered to be a cluster than between clusters that are close together are combined into one cluster, while the divisive method is initially all objects are in one cluster then the most different properties are separated and form one other cluster (Johnson and Wichern, 1992). The agglomerative method has several algorithms used to form clusters, namely single linkage, complete linkage, and average linkage (Supranto, 2004). In this study, the average linkage method was used. whereas the divisive method initially all objects are in one cluster then the most different properties are separated and form one other cluster (Johnson and Wichern, 1992).

According to Hair et al. (2006), the concept of similarity is important in cluster analysis because the principle of cluster analysis is to group objects that have the same characteristics. The distance measure is a measure of similarity, the higher the distance value, the lower the similarity between objects. This research wants to investigate the application of an integrated cluster in path with a distance measure, namely the Euclidean distance. Euclidean distance is the most commonly used type of distance measurement because it is one of the easiest methods to understand and model. This method is suitable for determining the closest distance between two data. Euclidean distance is the geometric distance between two data objects (Johnson and Winchern, 2002).

2.2 Cluster Analysis Distance Calculation

According to Hair et al. (2006), the concept of similarity is important in cluster analysis because the principle of cluster analysis is to group objects that have the same characteristics. The distance measure is a measure of similarity, the higher the distance value, the lower the similarity between objects. There are several methods of measuring distances, including Euclidean, Manhattan / City Block, Minkowski, Mahalanobis, Correlation, Angle-based, Squared Euclidean.

2.2.1 Euclidean distance

Euclidean distance is the most commonly used type of distance measurement because it is one of the easiest methods to understand and model. Euclidean distance is used to measure the distance from the data object to the center of the cluster. This method is suitable for determining the closest distance between two data. Euclidean distance is the geometric distance between two data objects (Johnson and Winchern, 2002).

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ri} - x_{rj})^2}$$

Where:

$d(x_i, x_j)$: distance between i and J

x_{ri} : variable value r in observation i

x_{rj} : variable value r in observation j

p : lots of data variables

2.2.2 Manhattan distance

Manhattan distance is used to calculate the absolute difference between the coordinates of a

pair of objects. Prasetyo (2012) states that the Manhattan distance is perfect for detecting outliers in the data.

$$d(x_i, x_j) = \sum_{r=1}^{p} |x_{ri} - x_{rj}|$$

Where:

$d(x_i, x_j)$: distance between i and J

x_{ri} : variable value r in an observation i

x_{rj} : variable value r in observation j

p : lots of data variables

2.2.3 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 by Herman Minkowski (1864-1909) (Anonymous, 2008d). Minkowski distance is a general form of the formula for calculating distance space or the distance between two points.

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

Where,

d = distance between x and y

x_i = data at the center of the cluster i

y_i = data on each data i-th

P = power

2.3 Cluster Validity Index

The main problem in Cluster Analysis is the number of groups that the researcher must determine because there is no solid basis for the number of the best groups. The next step is to do a cluster validity test to evaluate the results of the quantitative cluster analysis so that the optimum group is produced. An optimum group is a group that has a dense distance between individuals in the group and is well isolated from other groups (Ambassador and Jain, 1988).

2.3.1 Gap Statistics

Gap Analysis is a measurement method to determine the gap between the performance of a variable and consumer expectations for that variable. Gap Analysis itself is part of the IPA (Importance-Performance Analysis) method. A positive gap (+) will be obtained if the perception score is greater than the expected score, whereas if the expectation score is greater than the perception score, a negative (-) gap will be obtained. The higher the expectation score and the lower the perception score, the bigger the gap. If the total gap is positive, the creditor is considered very satisfied with the company's services. Conversely, if not, the gap is negative, then the creditor is less / not satisfied with the service. The smaller the gap the better. Usually, companies with a good level of service will have a smaller gap (Irawan, 2002). One way to estimate the optimal number of clusters is to use a statistical gap (Tibshirani et al, 2001). Suppose that it is an observation on the ith object and the j-variable. Then a cluster analysis was carried out on the data into k clusters, namely X_{ij} $C1, C_2, \ldots, C_3$ with are observations in the r and the cluster $C_r n_r$ is the number of objects in the r-th cluster, so it can be defined as follows:

$$D_r = \sum_{(ik,jk)} d_{ik} \quad (2.2)$$

Where D_r is the total distance of all points in the cluster r and is the distance between the ith object and the k-th object. d_{ik}

$$W_k = \sum_{r=1}^{k} \frac{1}{2_{n_r}} D_r \quad (2.3)$$

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

2.3.2 Goodman-Kruskal

The Goodman-Kruskal index measures cluster validation internally. The Goodman-Kruskal index finds the concordance and discounting of all possible input parameters. Good clustering is clustering that has many concordant and few concordant (Goodman and Kruskal, 1954). The Goodman-Kruskal index measures the ranking correlation between two sequences A = {a1, a2, .., an} and B = {b1, b2, .., bn} in terms of the number of concordant and discordant pairs in A and B. , aj) and (bi, bj) concordant if they are both ai <aj and bi <bj or ai> aj and bi> bj. Conversely, A and B are discordant if both ai <aj and bi> bj or ai> aj and bi <bj .. or if, for example, the four pairs of all observed objects are (q, r, s, t) with d (x, y) is the distance between the x and y objects. 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 different groups and s and t are in the same group. The Goodman-Kruskal index is calculated from the calculation of the value of the concordant and discordant pairs using the formula:

$$GK = \frac{S_c - S_d}{S_c + S_d}$$

where, S_c = number of concordant pairs

S_d = number of discordant pairs

Large GK values indicate the optimum group (Bolshakova, 2003).

Journal of Theoretical and Applied Information Technology
15th December 2021. Vol.99. No 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.3.3. Silhoutte Global Index

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

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

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:

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

Where

S (i) = Silhoutte group i

n = number of groups

2.3.4. Index C

This index can be explained as follows:

$$C = \frac{S - S_{min}}{S_{max} - S_{min}}$$

Where

S = the sum of distances in all pairs of observed objects from the same group, with ℓ the number of these pairs,

Smin = the number of ℓ the smallest distance if all sample pairs are in different groups.

Smax = the number of ℓ the greatest distance of all pairs.

A small C value indicates a good group (Bolshakova, 2003).

2.4 Path Analysis

In everyday life, there are many relationships between exogenous and endogenous variables through dummy variables in path analysis, namely when exogenous variables affect endogenous variables with the influence of dummy variables.

Lemma 1. Path analysis with dummy variable

View data $(X_i, Y_{1i}, Y_{2i}); i = 1,2, \ldots., n$ which will be modeled with dummy variable path analysis.

$$Z_{Y_{1i}} = \beta_{11} Z_{X_{1i}} + \beta_{12} D_i Z_{X_{2i}} + \varepsilon_{1i} \quad (2.7)$$

$$Z_{Y_{2i}} = \beta_{21} Z_{X_{1i}} + \beta_{22} D_i Z_{X_{2i}} + \beta_{23} Z_{Y_{1i}} + \varepsilon_{2i} \quad (2.8)$$

where :

$Z_{Y_{1i}}$: The i-th value of the standard score of the endogenous variable Y_1 (i=1,2,3,…,n);

$Z_{Y_{2i}}$: The i-th value of the standard score of the endogenous variable Y_2;

Z_{X_i} : The i-th value of the standard score of the exogenous variable;

n : Sample size;

β_{ij} : The j-th value of the coefficient of the effect of the exogenous variable on the i-th endogenous;

ε_i : Value of i-th residual variable

$\varepsilon_i \sim$NIID(0,σ^2);

D_i : The i-th value of the dummy variable.

Proof:

The combination of equations (2.9) and (2.10) forms the following matrix.

$$\boldsymbol{Y}_{2n\times 1} = \boldsymbol{X}_{2n\times 5}\boldsymbol{\beta}_{5\times 1} + \boldsymbol{\varepsilon}_{2n\times 1} \quad (2.9)$$

$$\begin{bmatrix} Z_{Y_{11}} \\ Z_{Y_{12}} \\ Z_{Y_{13}} \\ \vdots \\ Z_{Y_{1n}} \\ Z_{Y_{21}} \\ Z_{Y_{22}} \\ Z_{Y_{21}} \\ \vdots \\ Z_{Y_{2n}} \end{bmatrix} = \begin{bmatrix} X_{Z_{XX_D}} & \mathbf{0}_{n\times 5} \\ \mathbf{0}_{n\times 3} & X_{Z_{XX_DY}} \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \\ \beta_{23} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \vdots \\ \varepsilon_{1n} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \vdots \\ \varepsilon_{2n} \end{bmatrix}$$

where:

$$X_{Z_{XX_D}} = \begin{bmatrix} Z_{X_1} & D_1 \\ Z_{X_2} & D_2 \\ Z_{X_3} & D_3 \\ \vdots & \vdots \\ Z_{X_n} & D_n \end{bmatrix}; X_{Z_{XX_DY}} = \begin{bmatrix} Z_{X_1} & D_1 & Z_{Y_1} \\ Z_{X_2} & D_2 & Z_{Y_2} \\ Z_{X_3} & D_3 & Z_{Y_3} \\ \vdots & \vdots & \vdots \\ Z_{X_n} & D_n & Z_{Y_n} \end{bmatrix}$$

2.5 Ordinary Least Square

Theorem 1 OLS

In the linear model in parameters, the Ordinary Least Square (OLS) method can be used by minimizing the number of residual squares to estimate the path coefficient of the general form of the matrix operation, $\underset{\sim}{Y} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$, where $\underset{\sim}{\varepsilon} = \underset{\sim}{Y} - \mathbf{X}\underset{\sim}{\beta}$.

The OLS method minimizes the following functions:

$$\min\{Q\} = \min\left\{\underset{\sim}{\varepsilon}^{\mathrm{T}}\underset{\sim}{\varepsilon}\right\} = \min\left\{(\underset{\sim}{Y} - \mathbf{X}\underset{\sim}{\beta})'(\underset{\sim}{Y} - \mathbf{X}\underset{\sim}{\beta})\right\}$$

Parameter estimation with OLS approach by minimizing Q the following:

$$Q = \left(\underset{\sim}{\varepsilon}^{\mathrm{T}}\underset{\sim}{\varepsilon}\right) = \left(\underset{\sim}{Y} - \mathbf{X}\underset{\sim}{\beta}\right)'\left(\underset{\sim}{Y} - \mathbf{X}\underset{\sim}{\beta}\right)$$

$$= \left(\underset{\sim}{Y}' - \mathbf{X}'\underset{\sim}{\beta}'\right)\left(\underset{\sim}{Y} - \mathbf{X}\underset{\sim}{\beta}\right)$$

$$= \left(\underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{Y}'\mathbf{X}\underset{\sim}{\beta} - \underset{\sim}{\beta}'\mathbf{X}'\underset{\sim}{Y} + \underset{\sim}{\beta}'\mathbf{X}^{\mathrm{T}}\mathbf{X}\underset{\sim}{\beta}\right)$$

$$= \left(\underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}'\mathbf{X}'\underset{\sim}{Y} + \underset{\sim}{\beta}'\mathbf{X}'\mathbf{X}\underset{\sim}{\beta}\right)$$

The solution to the optimization of equation (2.18) is to do the derivative Q against $\underset{\sim}{\beta}$ and equal to 0.

$$\frac{\partial(Q)}{\partial(\underset{\sim}{\theta})} = \underset{\sim}{0}$$

$$-2\mathbf{X}'\underset{\sim}{Y} + 2\mathbf{X}'\mathbf{X}\underset{\sim}{\hat{\beta}} = \underset{\sim}{0}$$

$$-\mathbf{X}'\underset{\sim}{Y} + \mathbf{X}'\mathbf{X}\underset{\sim}{\hat{\beta}} = \underset{\sim}{0}$$

$$\mathbf{X}'\mathbf{X}\underset{\sim}{\hat{\beta}} = \mathbf{X}'\underset{\sim}{Y}$$

2.6 Hypothesis of Linear Parameter Function (HLPF)

a) Formulating a Linear Parameter Function Hypothesis.
Searle (1971) explains that there are procedures that can be used to write general and specific hypotheses. Hypothesis testing of linear parameter functions is used to simultaneously test the equations formed in the study. This hypothesis can be written in the form of a matrix as follows:

$$H : \mathbf{K'\beta = m}$$

b) Hypothesis Testing for Linear Function Parameters

$$\frac{Q}{s\sigma^2} = \frac{\left(\mathbf{K'\hat{\beta} - m}\right)'\left[\mathbf{K'(X'X)^{-1}K}\right]^{-1}\left(\mathbf{K'\hat{\beta} - m}\right)}{s\sigma^2} \sim F_{db_1, db_2}$$

2.7 Variables

2.7.1. Character

The character of the creditor's willingness to pay off his obligations as agreed in the credit agreement. Character creditors are divided into two, namely:

1) Creditors who pay credit at the beginning of the month (days 1 to 15), are called early payments.
2) Creditors who pay credit at the end of the month (16-30 days), are called late payments.

2.7.2. Capacity

Capacity is the ability of the creditor to run his business in order to make a profit so that he can repay the loan/financing from the profit generated. Capacity in this study is represented through the work of creditors which are divided into 4 categories, namely:

1) Creditor Income
2) Ability to Pay Installments
3) Ability to Settle Loans on Time

2.7.3. Capacity

Own capital will also be considered by the bank, as evidence of the seriousness and responsibility of creditors in running their business, because they share the risk of business failure.
Ways taken by banks to find out the capital owned by creditors, among others need to be considered:

1) Fixed source of income
2) Have another line of business as a source of income
3) Have savings or savings in the bank

2.7.4. Collateral

Collaterals are goods that are delivered as collateral for the financing they receive.Collateral must be assessed by the bank to determine the extent of the risk of the creditor's financial obligations to the bank. The assessment of this collateral includes the type, location, proof of ownership, and legal status. The indicators used by the bank in providing collateral include:

1) The selling value of collateralized goods is comparable to/exceeds the credit limit
2) Guarantees are physical or non-physical
3) Ownership of collateral and document authenticity

2.7.5. Condition

Condition is assessing credit by assessing current economic, social and political conditions and predictions for the future. Conditions in this research need to pay attention to several things, namely:

1) Business/business/investment development
2) Economic fluctuation
3) Socio-economic conditions/family problems

2.7.6. Intention to Pay Obedience

Intention to comply with the intention of the creditor in complying with his obligations to pay debts at the bank. The indicators used in this study which replicates from Bambang & Widi (2010) are:

1) Trend
2) Decision

2.7.7. Punctuality of Payment

Timely in paying can be defined as payment on time that has been determined. Indicators that measure on time pay are:

1) Desire to always pay on time
2) Monthly payments are always on time

3. RESEARCH METHOD

This study uses primary data, the variables used are character, capacity, capital, collateral, condition, intention to pay obedience, and punctuality of payment Bank X Creditors. The data consists of three exogenous variables, namely character, capacity, capital, collateral, and condition, and two endogenous variables, namely intention to pay obedience and punctuality of payment. 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. Purposive sampling is a sampling technique based on certain characteristics or conditions that are the same as the characteristics of the population. The sample used is 100 Bank X creditors. Data analysis was carried out quantitatively, to explain each of the variables studied, a descriptive analysis was carried out first, then carried out by path analysis. The path diagram used in this study is as shown in Fig 1.

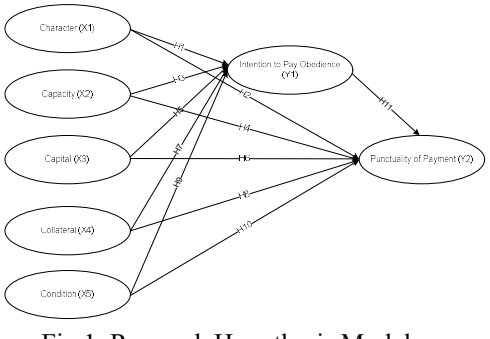

Fig 1. Research Hypothesis Model

4. RESULT AND DISCUSSION

4.1 Latent Variable Measurement Method

The results of the method of measuring demographic variables use an average score for each variable. The results of the average score can be seen in Table 4.1

Table 4.1 Quantification Results

No	X1	X2	X3	X4	X5	Y1	Y2
1	2,7	3,3	3,8	3,0	2,8	2,8	2,8
2	2,8	2,8	3,0	3,3	3,2	3,2	3,8
3	3,5	3,5	2,5	1,5	1,7	2,6	2,5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
98	3,7	2,2	3,3	3,3	2,8	2,2	2,5
99	2,7	3,0	4,2	2,5	2,8	3,3	3,8
100	1,5	3,8	4,0	4,0	2,2	3,2	3,5

4.2 Cluster Analysis

This study uses 8 cluster validity indices. The results of this study indicate that the number of members of clusters 1 and 2 for all indexes has the same number, namely, for cluster 1 there are 42 members and cluster 2 has 58 members. cluster validity index. The average results obtained can be seen in Table 1.

Table 4.2 Average Cluster Members for Index and Linkage respectively

Method	Average													
	X1		X2		X3		X4		X5		Y1		Y2	
	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
Gap Index														
Euclidean	3,9443	3,1772	4,0483	3,1033	3,9293	3,2333	3,9667	3,1266	4,0044	3,1155	3,9443	3,1772	4,0483	3,1033
Manhattan	3,8266	2,9588	3,7743	3,0533	3,8411	3,0009	3,8033	2,9950	3,8115	2,9956	3,8266	2,9588	3,7743	3,0533
Minkowski	3,9443	3,1772	4,0483	3,1033	3,9293	3,2333	3,9667	3,1266	4,0044	3,1155	3,9443	3,1772	4,0483	3,1033
Indeks C														
Euclidean	3,9443	3,1772	4,0483	3,1033	3,9293	3,2333	3,9667	3,1266	4,0044	3,9667	3,1266	4,0044	3,1155	3,9667
Manhattan	3,8266	2,9588	3,7743	3,0533	3,8411	3,0009	3,8033	2,9950	3,8115	3,8033	2,9950	3,8115	2,9566	3,8033
Minkowski	3,9443	3,1772	4,0483	3,1033	3,9293	3,2333	3,9667	3,1266	4,0044	3,9667	3,1266	4,0044	3,1155	3,9667
Global Sillhouette														
Euclidean	3,9443	3,1772	4,0483	3,1033	3,9293	3,2333	3,9667	3,1266	4,0044	3,1155	3,9443	3,1772	4,0483	3,1033

Manhattan	3,826	2,9588	3,7774	3,0533	3,8441	3,0099	3,8033	2,9500	3,8115	2,9556	3,8266	2,9588	3,774		3,0533
Minkowski	3,9433	3,1722	4,0488	3,1033	3,9289	3,2333	3,9667	3,1266	4,0044	3,1155	3,9433	3,1722	4,048		3,1033
Goodman- Kruskal															
Euclidean	3,9433	3,1722	4,0488	3,1033	3,9289	3,2333	3,9667	3,1266	4,0044	3,1155	3,1266	4,0065	3,115		3,1266
Manhattan	3,8266	2,9588	3,7774	3,0533	3,8441	3,0099	3,8033	2,9500	3,8115	2,9556	2,9500	3,8115	2,956		2,9500
Minkowski	3,9433	3,1722	4,0488	3,1033	3,9289	3,2333	3,9667	3,1266	4,0044	3,1155	3,1266	4,0044	3,115		3,1266

It can be seen from Table 1., the cluster means for the index Gap, C Index, Global Sillhouette, and Goodman-Kruskalhave the same average result for each linkage. That is, the Gap index, C Index, Global Sillhouette, and Goodman-Kruskal have the same cluster members for euclidean and minkowski distances and differ for the Gap index, C Index, Global Sillhouette, and Goodman-Kruskal distance in Manhattan. Thus, in conducting path analysis, the researcher uses the Gap index with euclidean and manhattan which will represent the C Index, Global Sillhouette, and Goodman-Kruskal index for each distance measure, because they have the same members of each cluster.

4.3 Integrated Cluster Index Model Gap and the Euclidean Distance

Based on the results of cluster analysis with the Gap index and euclidean distance, it was found that the number of clusters was 2 clusters, with cluster 1 as many as 42 creditors and cluster 2 as many as 58 creditors. Next, a dummy will be formed from the resulting clusters. The number of clusters formed is 2 clusters, so there is 1 dummy. The researcher determines creditors in cluster 1 as dummy 1 and creditors in cluster 2 as dummy 0.

Model feasibility test or Goodness of Fit testing the fit / suitability of the model with the research data held. The goodness of fit in question is an index or measure of the goodness of the relationship between latent variables related to its assumptions. In this study, the criteria in determining the goodness/feasibility of the model for an integrated cluster with the Path analysis can be seen in Table 2.

Table 2. Model Feasibility Test Results for Integrated Clusters with Path Analysis Gap Index and Euclidean Distance

No.	Model Fit / Quality Index	Score	Criteria	Information
1	Average path coefficient	APC = 0.607P <0.001	P <0.05	Significant
2	Average R-squared	ARS = 0.627 P <0.001	P <0.05	Significant
3	Average adjusted R-squared	AARS = 0.885P <0.001	P <0.05	Significant
4	Average block VIF	AVIF = 123,659	*acceptable* if AVIF ≤ 5 ideal if AVIF ≤ 3,3	Rejected
5	Average full collinearity VIF	AFVIF = 129,380	*acceptable if AFVIF ≤ 5 ideal if AFVIF ≤ 3,3*	Rejected
6	Tenenhaus GoF	GoF = 0.133	*small if GoF ≥ 0.1 medium if GoF ≥ 0.25 large if GoF ≥ 0.36*	Small
7	Sympson's paradox ratio	SPR = 0.625	*acceptable if the SPR ≥ 0.7 ideal if SPR = 1*	Rejected
8	R-squared contribution ratio	RSCR = 0.264	*acceptable if RSCR ≥ 0.9 ideal RSCR = 1*	Rejected
9	Statistical suppression ratio	SSR = 0.813	*acceptable if SSR ≥ 0.7*	Acceptable
10	Nonlinear bivariate causality direction ratio	NLBCDR = 1,000	*acceptable if NLBCDR ≥ 0.7*	Acceptable

Source: Primary Data Processed (2021)

Table 2 is a summary of the results obtained in the analysis and the recommended values for measuring the feasibility of the model. Based on the results of the feasibility test of the model as a whole, not all of the criteria reached the expected value limit or did not meet the recommended goodness of fit indices critical limit, so the results of this modeling could not be accepted or worthy of analysis. There are several criteria rejected, including Average block VIF, Average full collinearity VIF, Sympson's paradox ratio, and R-squared contribution ratio. It can be stated that this test results in a poor conformation of the variables as well as the causal relationship between variables. Thus, the overall model test shows unfavorable results or following expectations, meaning that the empirical data (field data) does not support the theoretical model developed.

4.4 GAP Index and Manhattan Distance Cluster Integrated Model

Based on the results of cluster analysis with the Gap index and Manhattan distance, it was found that the number of clusters was 2 clusters, with cluster 1 as many as 42 creditors and cluster 2 as many as 58 creditors. Next, a dummy will be formed from the resulting clusters. The number of clusters formed is 2 clusters, so there is 1 dummy. The researcher determines creditors who are in cluster 1 as dummy 1 and creditors in cluster 2 as dummy 0.

Model feasibility test or Goodness of Fit testing the fit / suitability of the model with the research data held. The goodness of fit in question is an index or measure of the goodness of the relationship between latent variables related to its assumptions. In this study, the criteria for determining the goodness/feasibility of the model for an integrated cluster with the Path Analysis can be seen in Table 3.

Table 3. Model Feasibility Test Results for Integrated Cluster with Path Analysis Manhattan Gap Index and Distance

No.	Model Fit / Quality Index	Score	Criteria	Information
1	Average path coefficient	APC = 0.451P <0.001	P <0.05	Significant
2	Average R-squared	ARS = 0.956 P <0.001	P <0.05	Significant
3	Average adjusted R-squared	AARS = 0.977P <0.001	P <0.05	Significant
4	Average block VIF	AVIF = 42,681	acceptable if AVIF ≤ 5 ideal if AVIF ≤ 3,3	Rejected
5	Average full collinearity VIF	AFVIF = 105,081	acceptable if AFVIF ≤ 5 ideal if AFVIF ≤ 3,3	Rejected
6	Tenenhaus GoF	GoF = 0.991	small if GoF ≥ 0.1 medium if GoF ≥ 0.25 large if GoF ≥ 0.36	Big
7	Sympson's paradox ratio	SPR = 0.875	acceptable if the SPR ≥ 0.7 ideal if SPR = 1	Acceptable
8	R-squared contribution ratio	RSCR = 0.747	acceptable if RSCR ≥ 0.9 ideal RSCR = 1	Rejected
9	Statistical suppression ratio	SSR = 0.875	acceptable if SSR ≥ 0.7	Acceptable
10	Nonlinear bivariate causality direction ratio	NLBCDR = 1,000	acceptable if NLBCDR ≥ 0.7	Acceptable

Source: Primary Data Processed (2021)

Table 3 is a summary of the results obtained in the analysis and the recommended values for measuring the feasibility of the model. Based on the results of the feasibility test of the model as a whole, not all of the criteria reached the expected value limit or did not meet the recommended Goodness of fit indices critical limit, so that the modeling results could be accepted or worthy of analysis. There are several criteria rejected, including Average block VIF, Average full collinearity VIF, and R-squared contribution ratio. It can be stated that this test resulted in a fairly good confirmation of the variables as well as the causal relationship between variables. So, the overall model test shows fairly good results or following expectations, meaning that the empirical data (field data) is sufficient to support the theoretical model being developed.

4.5 Comparison of R2 Path Analysis on an Integrated Cluster with Various Indices and Distances

In this study, the criteria for determining the best model for an integrated cluster with Path Analysis at various indices and distances can be seen in Table 5.

Table 5. Comparison of R2 Value

Index	R2 value
Silhouette	
Euclidean	0.885
Manhattan	0.977
Minkowski	0.885
Krzanowski-Lai	
Euclidean	0.885
Manhattan	0.977
Minkowski	0.885
Dunn	
Euclidean	0.885
Manhattan	0.977
Minkowski	0.885
Davies-Bouldin	
Euclidean	0.885
Manhattan	0.977
Minkowski	0.885

Based on table 5, it can be seen that the integrated cluster model of the Path Analysis approach using the Gap index, Index C, Global Sillhouette, and Goodman-Kruskal with the Manhattan distance has an R2 value of 0.977 which means the variable character, capacity, capital, collateral, condition, intention to pay obedience simultaneously affects punctuality of payment by 97.7%, while the remaining 6.3% is influenced by other variables. The integrated cluster model of Path model using the Gap index, Index C, Global Sillhouette, and Goodman-Kruskal with Euclidean and Minkowski distances has an R2 value of 0.885 which means that the variables of character,

capacity, capital, collateral, condition, and intention to pay obedience simultaneously affect punctuality of payment is 85.5%, while the remaining 14.5% is influenced by other variables.

$$Y_1 = \beta_1 D_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4 + \beta_6 D_2 X_1 + \beta_7 D_3 X_2 + \beta_8 D_4 X_3 + \beta_9 D_5 X_4$$

$$Y_1 = -0,643 D_1 + 0,430 X_1 + 0,318 X_2 + 0,328 X_3 + 0,345 X_4 + 0,790 D_2 X_1 + 0,199 D_3 X_2 + 0,420 D_4 X_3 + 0,423 D_5 X_4$$

Cluster 1 (D = 1):

$$Y_1 = -0,643 + 1,220 X_1 + 0,517 X_2 + 0,748 X_3 + 0,612 X_4 \quad (4.1)$$

Cluster 2 (D = 0):

$$Y_1 = 0,439 X_1 + 0,318 X_2 + 0,328 X_3 + 0,256 X_4 \quad (4.2)$$

Based on equations 4.1 and 4.2, it can be concluded that capacity (X1), capital (X2), collateral (X3), and condition (X4) in cluster 1 have a greater influence than cluster 2. In cluster 1, each increase is one unit of capacity (X1) it will increase the intention to pay obedience (Y1) by 1,220 units. Every increase of one capital (X2) for a creditor in cluster 1, it will increase the creditor's intention to pay obedience (Y1) by 0.517 units. Also, each increase of one unit of collateral (X3) in cluster 1 will increase the creditor's intention to pay obedience (Y1) by 0.748 units and each increase of one unit of condition (X4) in cluster 1 will increase the creditor's intention to pay obedience (Y1) by 0.653 units.

In cluster 2, each increase of one environmental quality unit (X1) will increase the creditor's willingness to pay (Y1) by 0.439 units. Every increase of one environmental unit (X2) for a creditor in cluster 2, it will increase the creditor's willingness to pay (Y1) by 0.318 units. Also, each increase of one unit of creditor fashions (X3) in cluster 2 will increase the creditor's willingness to pay (Y1) by 0.328 units.

$$Y_2 = \beta_1 D_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 Y_1 + \beta_6 D_2 X_1 + \beta_7 D_3 X_2 + \beta_8 D_4 X_3 + \beta_9 D_5 Y_1$$

$$Y_2 = -1,221 D_1 + 0,028 X_1 + 0,259 X_2 + 0,112 X_3 + 0,316 Y_1 + 0,937 D_2 X_1 + 0,463 D_3 X_2 + 0,462 D_4 X_3 + 0,294 D_5 Y_1$$

Cluster 1 (D = 1):

$$Y_2 = -1,221 + 0,491 X_1 + 0,721 X_2 + 1,049 X_3 + 0,610 Y_1 \quad (4.3)$$

Cluster 2 (D = 0):

$$Y_2 = 0,028 X_1 + 0,259 X_2 + 0,112 X_3 + 0,316 Y_1 \quad (4.4)$$

Based on equations 4.3 and 4.4, it can be concluded that capital (X1), capacity (X2), collateral (X3), condition (X4), and intention to pay obedience (Y1) in cluster 1 have a greater influence than cluster 2. In cluster 1, each increase one capital (X1) will increase creditors' punctuality of payment (Y2) by 0.491 units. Every increase of one environmental unit (X2) of creditors in cluster 1, it will increase the creditor punctuality of payment (Y2) of 0.721 units. Every increase of one unit of creditor collateral (X3) in cluster 1, it will increase creditors' punctuality of payment (Y2) by 1,049 units. Also, every increase of one unit of intention to pay obedience (Y1) of creditors in cluster 1, it will increase the punctuality of payment (Y2) by 0.610 units.

In cluster 2, each increase of one capital unit (X1) will increase the creditor's punctuality of payment (Y2) by 0.028 units. Every increase of one capacity unit (X2) for creditors in cluster 2, it will increase the creditor punctuality of payment (Y2) by 0.259 units. Every increase of one unit of creditor collateral (X3) in cluster 2, it will increase creditors' punctuality of payment (Y2) by 0.112 units. Also, every increase of one unit of intention to pay obedience (Y1) of creditors in cluster 1, it will increase the punctuality of payment (Y2) by 0.316 units.

5. CONCLUSION AND SUGGESTIONS

The conclusion that can be given based on the results of the analysis is the application of an integrated cluster in Path Analysis with various cluster validity index methods resulting in many clusters and the same cluster members causing the same dummy variables. The value of R2 on the integrated cluster with the Gap index, C Index, Global Sillhouette, and Goodman-Kruskal with Manhattan Distance is better than Gap, Index C, Global Sillhouette, and Goodman-Kruskal with Euclidean and Minkowski Distance. Variable Capacity (X1), Capital (X2), Collateral (X3), Condition (X4) and Intention to pay obedience (Y1) in cluster 1 have a greater influence than cluster 2.

Suggestions that can be given are based on the results of the integrated cluster on Path Analysis, namely for further research to compare the effect of using $_1$ linkage, as well as the distance on $_4$ the discriminant integrated cluster analysis which results in a high R2 value.

REFERENCES:

[1] Bambang & Widi. 2010. The Influence of Attitudes, Subjective Norms, Perceived Behavior Control, and Sunset Policy on Taxpayer Compliance with Intention as an Intervening Variable.

[2] Bates, A., & Kalita, J. (2016, March). Counting clusters in twitter posts. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (pp. 1-9).
[3] Bolshakova, N., & Azuaje, F. (2003). Cluster validation techniques for genome expression data. Signal processing, 83 (4), 825-833.
[4] Budiono, D. 2016. The behavior of Corporate Taxpayers in Fulfilling Tax Obligations: Humanistic Theory Perspective.
[5] Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, MM (2014). Package 'nbclust'. Journal of statistical software, 61 (6), 1-36.
[6] Fornell, C., & Larcker, DF (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. Division of Research Journal, 266.
[7] Ghozali, I. (2008). Structural equation modeling: An alternative method with partial least squares (pls). Diponegoro University Publishing Agency.
[8] Golin, J., & Delhaise, P. 2013. The bank credit analysis handbook: a guide for analysts, bankers, and investors. John Wiley & Sons.
[9] Hair, JF, Anderson, RE, Tatham, RL, and Black, WC 2006. Multivariate Data Analysis. Fifth edition. Jakarta: Gramedia. Main Library.
[10] Hox, JJ, & Bechger, TM (1998). An introduction to structural equation modeling. Family Science Review, 11, 354-373.
[11] Irawan, B. (2002). Stabilization of Upland Agriculture Under El Nino-induced Climatic Risk: Impact Assessment and Mitigation Measures in Indonesia (No. 1438-2016-118920).
[12] Jain, AK, & Ambassador, RC (1988). Algorithms for data clustering. Prentice-Hall, Inc.
[13] Johnson, RA and Wichern, DW 1992. Applied Multivariate. Analysis, Third Edition, New Jersey: Prentice Hall Inc.
[14] Krazanowski, WJ, & Lai, YT (1988). A criterion for determining the number of groups in a data set using the sum of squares clustering. Biometrics, 44, 23–34.
[15] Kotler, Philip. 2005. Marketing Management Volume 1 (11th ed.). PT. Index. Jakarta.
[16] Munadjat D. (1984). Environmental law (book V: Sectoral): Indonesian environmental law (in the National & International system). Bandung: Binacipta.
[17] Permadi, T., Nasir, A., & Anisma, Y. 2013. Study of willingness to pay taxes on individual taxpayers who do independent work (the case at KPP Pratama Tampan Pekanbaru). Journal of Economics, 21 (02).
[18] Putro, SW (2014). The Effect of Service Quality and Product Quality on Creditor Satisfaction and Consumer Loyalty in Happy Garden Restaurants. Journal of Marketing Strategy, 2 (1), 1-9.
[19] Robbinson and Stephen. (2002). Organizational Behavior Controversy Concept, Application, Jakarta Prehalinda.
[20] Siat, CC, & Toly, AA 2013. Factors Affecting Taxpayer Compliance in Fulfilling Tax Obligations in Surabaya. Tax & Accounting Review, 1 (1), 41.
[21] Siswadi and B. Suharjo. 1998. Multiple Variable Data Exploration Analysis. Final Project Not Published. Bogor: Department of Mathematics, Faculty of Mathematics and Natural Sciences IPB, Bogor.
[22] Soeriaatmadja, RE (1997). Environmental Science. Bandung: ITB Publisher.
[23] Supranto, 2004. Multivariate Analysis of Meaning and Interpretation, Jakarta: PT. Rineka Cipta.
[24] Sulistiyono, A., & Ayuvisda, ADINCHA 2012. The Influence of Motivation on Taxpayer Compliance in Paying Personal Income Taxes of Entrepreneurs (Study at the Bead Production Center, Plumbongambang Village, Gudo District, Jombang Regency, East Java Province). Journal of Accounting Unesa, 1 (1).
[25] Thio, Alex. 1987. Sociology (An Introduction). New York: Westview.
[26] Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63 (2), 411-423.
[27] Wani, MA, & Riyaz, R. (2017). A novel point density based validity index for clustering gene expression datasets. International Journal of Data Mining and Bioinformatics, 17 (1), 66-84.