Journal of Theoretical and Applied Information Technology
30th September 2021. Vol.99. No 18

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

COMPARISON OF CLUSTER VALIDITY INDEX USING INTEGRATED CLUSTER ANALYSIS WITH STRUCTURAL EQUATION MODELINGTHE WAR-PLS APPROACH

ADJI ACHMAD RINALDO FERNANDES[1], SOLIMUN[2], FARID UBAIDILLAH[3], AISYAH ARYANDANI[4], ABELA CHAIRUNISSA[5], AISYAH ALIFA[6], ENDANG KRISNAWATI[7], ERLINDA CITRA LUCKI EFENDI[8], NI MADE AYU ASTARI BADUNG[9], ALIFYA AL ROHIMI[10], EVA FADILAH RAMADHANI[11], FATHIYATUL LAILI NUR RASYIDAH[12]

[1,2]Lecturer, University of Brawijaya, Department of Statistics, East Java, Indonesia

[3,4,5,6,7,8,9,19,11,12]Student. University of Brawijaya, Department of Statistics, East Java, Indonesia

E-mail: [1]fernandes@staff.ub.ac.id, [2]solimun@ub.ac.id, [3]farid.ub@student.ub.ac.id, [4]aisyaharyandani@student.ub.ac.id, [5]abelacns@gmail.com, [6]Aisyha@student.ub.ac.id, [7]endangkrisna@student.ub.ac.id, [8]erctr13@student.ub.ac.id, [9]ayuastari@student.ub.ac.id, [10]alifyaalrohimi@student.ub.ac.id, [11]evafadilah_@student.ub.ac.id, [12]fathiyatullaili@student.ub.ac.id

ABSTRACT

This study wants to compare the Integrated Cluster Analysis and SEM model of the Warp-PLS approach with various cluster validity indices on data on Service Quality, Environment, Fashions, Willingness to Pay, and Compliant Paying Behavior of Bank X Customers. The data used in this study are primary data. The variables used in this study are service quality, environment, fashion, willingness to pay, and compliance with paying behavior at Bank X. 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 customer 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 SEM analysis of the Warp-PLS approach were carried out with the ward linkage method and the euclidean distance on various cluster validity indices, including: Sillhouette index, Krzanowski-Lai, Dunn, Gap, Davies-Bouldin, Index C, Global Sillhouette, Goodman-Kruskal in this study were used as analysis tools. This research uses R software. Integrated cluster with index C is better for modeling influence between variables than index Silhouette, Krzanowski-Lai, Dunn, GAP, Davies-Bouldin, Global Sillhouette, and Goodman-Kruskal. The novelty in this study is the application of Integrated Cluster Analysis and SEM of the Warp-PLS approach to compare 8 cluster validity indices, namely the Silhouette Index, Krzanowski-Lai, Dunn, Gap, Davies-Bouldin, C Index, Global Sillhouette, Goodman-Kruskal simultaneously

Keywords: *Cluster Analysis, Sem, Warp-Pls, Integration Model, Dummy Variable, Cluster Validity Index*

1. INTRODUCTION

Cluster analysis is one of the multiple variables (multivariate) analysis included in the interdependency method, namely the independent or explanatory variables are not differentiated from the dependent variable or response. 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). The method that is included in the non-hierarchical method is the K-Means method.

Hierarchical methods and non-hierarchical methods have differences in determining the number of clusters. The hierarchical method of determining the number of clusters has not been determined, while the non-hierarchical method of determining the number of clusters has been determined first. Hierarchical methods have advantages over non-hierarchical methods. The advantage of the hierarchical method is that it is

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

easier to study all clusters that are formed and more informative because, in the hierarchical method, the stages of grouping are presented in the form of a dendogram or tree diagram.

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 Structural Equation Modeling (SEM) 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 SEM) Structural Equation Modeling (SEM) is a complex multivariate analysis method used to determine the relationship between variables that cannot be measured directly (latent variables). SEM can be categorized into 2 models, namely structural models and measurement models.

Based on RI Law number 10 of 1998 article 1 paragraph 2 concerning banking, a bank is a business entity that collects public funds in the form of savings and simultaneously distributes funds to the public in the form of credit and/or other forms to improve people's lives. Credit is one of the functions of a bank that is very helpful for the community. One type of credit provided by a bank is a Home Ownership Credit (KPR). KPR is one of the financing products provided by banks for home buyers with a financing scheme of up to 90% of the house price. Debtors who have non-current credit are one of the credit problems that can harm the bank. Before a bank gives credit to a debtor, it is necessary to measure whether the debtor can carry out his obligations in credit or not. From these problems, it is necessary to have supervision in the provision of KPR. One of the statistical tools that can be used in this problem is cluster analysis which is integrated with SEM.

In this study, the researcher will compare the integrated cluster analysis model and SEM using the Euclidean distance measure with different cluster validity indices. Therefore, the cluster distance size will be compared with eight cluster validity indices, namely statistical gap, silhouette, krznawski-lai, goodman kruskal, dunn, global silhouette, index C, Davies-Bouldin. The cluster validity test is used to evaluate the results of the cluster analysis. quantitative 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 [1].

2. LITERATURE REVIEW

2.1 Cluster Analysis

According to [2], 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 [3]. 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 [3]. The agglomerative method has several algorithms used to form clusters, namely single linkage, complete linkage, and average linkage [4]. 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 [3].

According to [5], 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 SEM 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 [3].

2.2 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 [1]. The selection of the cluster validity index in this study was based on the most commonly used validity index. The cluster validity indices used include statistical gap, silhouette, krznawski-lai, goodman kruskal, dunn, global silhouette, index C, and Davies-Bouldin.

2.2.1 Gap statistics

Subsub 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 customer is considered very satisfied with the company's services. Conversely, if not, the gap is negative, then the customer 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 [6]. One way to estimate the optimal number of clusters is to use a statistical gap [7]. 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$X_{ij} C_1, C_2, \ldots, C_3$ with are observations in the r and th cluster $C_r n_r$ is the number of objects in the r-th cluster, so it can be defined as follows:

$$D_r = \sum_{(ik,jk)} d_{ik} \quad (1)$$

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}

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

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

2.2.2 Silhouette

According to Charrad et al., (2014), in 1987 Rousseuw introduced the Silhouette index with the following equation:

$$S = \frac{\sum_{i=1}^{n} S(i)}{n} \quad (3)$$

Where,

$$S(i) = \frac{b(i) - a(i)}{max[a(i), b(i)]} \quad (4)$$

$a(i) = \frac{\sum d_{ik}}{n_r - 1}$, is the average distance between the ith object and all the observed objects in the same cluster

$b(i) = min\left(\frac{\sum d_{ik}}{n_s}\right)$, is the average distance of the ith object with all the observed objects contained in the other clusters.

The maximum value of the index is used to determine the number cluster optimal.

2.2.3 Krzanowski-lai

According to [8], an index that is based on a decrease in the value of the number of squares in a cluster can be defined as follows:

$$DIFF(k) = \left[(k-1)^{1/v} W(k-1) - k^{2/v} W(k)\right] \quad (5)$$

then choose one that makes the value below the maximum$k$

$$KL(k) = \frac{DIFF(k)}{DIFF(k+1)} \quad (6)$$

Suppose that it is defined as the number of optimal clusters. If there is an increase in the number of clusters that form up to, the value will decrease drastically for. $ggW(k)k < gDIFF(k)kk = gKL(k)k$is expected to be of little value for all values except. This will create the maximum value for optimal.

2.2.4 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. 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:

$$GK = \frac{S_c - S_d}{S_c + S_d} \quad (7)$$

where S_c= number of concordant pairs
S_d = number of discordant pairs
Large GK values indicate the optimum group (Bolshakova, 2003).

2.2.5 Dunn

The Dunn validation index denoted by D is calculated by the following formula:

$$D = \min_{1\le i\le n}\left\{\min_{1\le j\le n, i\ne j}\left\{\frac{d(c_i,c_j)}{\max_{1\le k\le n}(d`(c_k))}\right\}\right\} \quad (8)$$

Where d (ci, cj) = distance between ci and cj groups
d '(ck) = distance in group ck
The greatest value of D is taken as the optimum number of groups.

2.2.6 Silhoutte global index

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

$$S(i) = \frac{(b(i) - a(i))}{max\{a(i), b(i)\}} \quad (9)$$

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

2.2.7 Index C

This index can be explained as follows:

$$C = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (10)$$

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 [9].

2.2.8 Davies-bouldin

Davies-Bouldin Index is one of the methods used to measure the validity of clusters in a clustering method, cohesion is defined as the sum of data closeness to the cluster center point of the cluster being followed. While the separation is based on the distance between the cluster center points to the cluster. This Davies-Bouldin Index measurement maximizes the inter-cluster distance between clusters Ci and Cj and at the same time tries to minimize the distance between points in a cluster. If the inter-cluster distance is maximum, it means that the similarities in the characteristics between each cluster are slight so that the differences between the clusters are more pronounced. If the intra-cluster distance is minimal, it means that each object in the cluster has a high level of characteristic similarity [10].

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i\ne j}(R_{i,j}) \quad (11)$$

Where,

$$SSW_i = \frac{1}{m_i}\sum_{j=i}^{m_i} d(x_j, c_j) \quad (12)$$

$$SSB_{i,j} = d(c_i, c_j) \quad (13)$$

$$R_{i,j} = \frac{SSW_I + SSW_J}{SSB_{i,j}} \quad (14)$$

K is the number of clusters used. Based on the above calculations, it can be observed that the smaller the SSW value, the better the clustering results will be obtained. Essentially, the DBI wants a value as small (non-negative ≥ 0) as possible to assess the goodness of the cluster obtained. The index is obtained from the average of all cluster indexes, and the value obtained can be used as decision support to assess the number of clusters that are most suitable for use. DBI is also widely used to assist k-means in determining the right number of clusters to use because usually, k-means cannot yet determine the number of clusters used for data clustering. [11].

2.3 Structural Equation Modeling (SEM)

SEM is a type of multivariate analysis, which has the ability and superiority to analyze multivariate and multi-relational data at the same time, which are relatively complex. SEM is usually used to study the causal relationship between latent variables. SEM has a high degree of flexibility in combining theory and empirical knowledge by modeling observation errors, combining theory and empirical analysis, confirming theory and data

(hypothesis testing), and developing theory and data [12]. SEM is a technique used to describe the simultaneous linear relationship between observed variables, which also involves latent variables that cannot be measured directly. SEM analysis combines a system of simultaneous equations, path analysis, regression analysis, and factor analysis [13].

There are two methods in SEM analysis, namely covariance-based and component-based or variant-based methods. Several assumptions that must be fulfilled when using variant-based SEM are multivariate normal distribution, large sample size, and reflective indicators when forming latent variables. SEM analysis based on variance is guided by the fact that theory plays a very important role in constructing the causal relationship of the structural model, and its aim is only to confirm whether the theory-based model differs from the empirical model. Unlike covariance-based SEM, variant-based SEMs such as Partial Least Squares (PLS) and Generalized Structured Component Analysis (GSCA) do not require assumptions. If the data cannot be analyzed by covariance-based SEM, then variant-based SEM can be used [14].

Statistical modeling that involves the simultaneous relationship between variables and indicator models is called structural equation modeling (SEM). SEM analysis as a representation of the system under study should be able to explain the behavior of the system close to real conditions. Initially, SEM analysis combines a system of simultaneous equations, or path analysis, or regression analysis with factor analysis. In this case, factor analysis is used as a method to obtain latent variable data. The process of estimating parameters and testing is based on the concept of the variance-covariance matrix, so it is often called covariance-based SEM.

The explanation regarding structural equation modeling will be easier to understand if it is given an illustration such as Figure 1. Figure 1 is a structural model, with X1 being the exogenous variable, Y1 being the endogenous mediating variable, and Y2 being the endogenous dependent variable.

2.4 SEM with the WarpPLS Approach

The WarpPLS analysis is an extension of the partial least squares (PLS) analysis. PLS is a combination of path and factor analysis and component analysis. PLS is usually referred to as variant-based SEM. If there is a problem with a weak theoretical basis, PLS is the more appropriate method because it can be used for prediction.

PLS was developed as an alternative to research with a weak theoretical basis or indicators that do not meet the reflective measurement model. In PLS it is possible to carry out structural modeling using reflective and formative indicators. PLS can be applied to all data scales, does not require many assumptions, and can be used on small sample sizes so it is a powerful analysis [15]. PLS is usually used as theory confirmation (hypothesis testing) but can also be used for proposition testing.

The focus of analysis in PLS shifts from only parameter estimation to validity and accuracy of prediction because it is based on a shift in analysis from estimating model parameters to estimating relevant parameters. There are two characteristics of indicators in PLS, namely reflective indicators, and formative indicators. If the structural model to be analyzed is non-recursive and the latent variables have indicators that are formative, reflective, or mixed, then one of the appropriate methods to be applied is WarpPLS [16]. WarpPLS is a method and package application software program developed by Ned Kock to analyze SEM models based on variants or PLS. WarpPLS software is also equipped with moderating variable analysis with the interaction variable approach.

2.5 Quality of Service

Service quality is a model that describes the condition of customers in forming expectations for service from past experiences, word of mouth promotion, and advertisements by comparing the services they expect with what they receive/feel [17]. Meanwhile, according to [18], service quality is all forms of activities carried out by companies to meet consumer expectations. Service, in this case, is defined as a service or service delivered by the service owner in the form of ease, speed, relationship, ability, and hospitality addressed through attitudes and characteristics in providing services for customer satisfaction. Service quality can be identified by comparing consumers' perceptions of services that are actually received or obtained with services that are actually expected or desired for the service attributes of a company. Five indicators can measure service quality, namely: 1) Reliability, 2) Responsiveness, 3) Assurance, 4) Empathy, and 5) Tangibles.

2.6 Environment

The environment is all objects and conditions, including humans and their actions, which are contained in the space where humans live and affect the life and welfare of humans and other living bodies. In another definition, the environment is defined as a spatial unit with all objects and conditions of living things including humans, and their behavior and other living things [19]. Meanwhile, according to [20], the environment is institutions or outside forces that have the potential to affect organizational performance, the environment is formulated into two, namely the general environment and the special environment. The general environment is anything outside the organization that has the potential to influence the organization. This environment is in the form of social and technological conditions, while the special environment is the part of the environment that is directly related to the achievement of an organization's goals. Two indicators can measure the environment, namely: 1) Physical Environment, and 2) Non-Physical Environment.

2.7 Fashion

Currently, fashion is related to clothing or clothes, when in fact what is said to be fashion is everything that is trending in society. This includes clothing, appetite, entertainment, consumer goods, and so on. So actually fashion can include anything that is followed by many people and becomes a trend. Fashion is also related to novelty or novelty elements, therefore fashion tends to be short-lived and not eternal. And because what tends to move and change every time is clothing, fashion is often associated with clothing, whereas as long as there is something new about an artifact that involves the fun of many people, it can become a fashion [21]. Two indicators can measure the environment, namely: 1) Activities, 2) Interests, and 3) Opinions.

2.8 Willingness to Pay

Willingness to paymeans the willingness of the credit applied to pay the payment burden according to a predetermined amount of credit. Willingness to pay or willingness to pay is closely related to variables that affect the ability of lenders to force payments through legal compensation, this is an important part of the overall effectiveness analysis process and creditor friendliness [22]. Willingness to pay is a value where someone is willing to pay, sacrifice or exchange something to obtain goods or services. According to Permadi et al. (2013), five indicators can measure willingness to pay, namely: 1) Consultation before making payments; 2) Documents required to pay; 3) Information regarding the method and place of payment; 4) Information regarding the payment deadline; 5) Allocate payment funds.

2.9 Compliant Paying Conduct

According to [23] behavior is a human reaction or action caused by an impulse that is seen from values, habits, driving forces, motives, and strength of detention as endogenous or someone's reaction that appears. This is due to the experience of the stimulation process and learning from the environment. According to [24], obedience is an attitude of being willing to do whatever is based on self-awareness or coercion which causes behavior according to what is not expected. Obedient behavior is the interaction of individual, organizational, and group behavior [25]. Obedient paying behavior is defined as someone's action caused by self-awareness or compulsion to pay their obligations. According to [26], three indicators can measure customer compliance, namely: 1) Timeliness; 2) Accuracy of data; 3) Sanctions.

3. RESEARCH METHODS

Figures This study uses primary data, the variables used are service quality, environment, fashions, willingness to pay, and compliance with paying behavior at Bank X. The data consists of three exogenous variables, namely service quality, environment, and fashion, and two endogenous variables, namely willingness. to pay and compliant pay behavior. 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 customers.

Data analysis was carried out quantitatively, to explain each of the variables studied, a descriptive analysis was carried out first, then carried out by Structural Equation Modeling (SEM) analysis based on Partial Least Square (PLS) in this study used as an analysis tool. According to [16] using the WarpPLS program will obtain a PLS (Partial Least Square) analysis model for the following reasons: (1) The analysis model is tiered and the structural equation model meets the recursive model. (2) Measurement of latent

variables, namely any variables that cannot be measured directly.

Figure 2 is a picture of the research hypothesis model obtained based on the results of previous research.

Based on the research hypothesis model in Figure 1, the research hypothesis can be formulated as follows:

H1: Service Quality (X1) has a significant effect on willingness To Pay
H2: Environment (X2) has a significant effect on Willingness To Pay
H3: Fashions (X3) has a significant effect on Willingness To Pay
H4: Quality of Service (X1) has a significant effect on Compliant Paying Conduct
H5: Environment (X2) has a significant effect on Compliant Paying Conduct
H6: Fashions (X3) has a significant effect on Compliant Paying Conduct
H7: Willingness To Pay (Y1) has a significant effect on Compliant Paying Conduct

4. RESULTS AND DISCUSSION

4.1 Cluster Analysis

This study uses 8 cluster validity indices. The results of this study indicate that the number of members of cluster 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.

It can be seen from Table 1., the cluster mean for the index Silhouette, Krzanowski-Lai, Dunn, Gap, Davies-Bouldin, Global Sillhouette, and Goodman-kruskal have the same mean results, and for index C they have different means. This means that the Silhouette, Krzanowski-Lai, Dunn, Gap, Davies-Bouldin, Global Sillhouette, and Goodman-kruskal indexes have the same cluster members, as well as the C index. Judging from table 1, the best cluster indexes are the Silhouette and Davies index. Bouldin. Thus, in conducting SEM analysis of the Warp-PLS approach, the researcher uses the Silhouette index which will represent the Krzanowski-Lai, Dunn, Gap, Davies-Bouldin, Global Sillhouette, and Goodman-kruskal indices because they have the same members of each cluster and use the C index.

4.2 Model Integrated Cluster Index Silhouette

Based on the results of cluster analysis with the Silhouette index, it was found that the number of clusters was 2 clusters, with cluster 1 as many as 42 customers and cluster 2 as many as 58 customers. 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 customers who are in cluster 1 as dummy 1 and customers 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 (inner model) related to its assumptions. In this study, the criteria in determining the goodness/feasibility of the model for an integrated cluster with the SEM Warp-PLS approach can be seen in Table 2.

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, all the criteria have reached the expected value limit or have met the recommended Goodness of fit indices critical limit, so that the results of this modeling can be accepted or worthy of analysis. However, several criteria were rejected, including Average block VIF, Average full collinearity VIF, and-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 good results or following expectations, meaning that the empirical data (field data) has supported the theoretical model developed.

4.3 Model Integrated Cluster Index C

Based on the results of cluster analysis with index C, it was found that the number of clusters was 2 clusters, with cluster 1 as many as 42 customers and cluster 2 as many as 58 customers. 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 customers who are in cluster 1 as dummy 1 and customers 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 (inner model) related to its assumptions. In this study, the criteria for determining the goodness/feasibility of the model for an integrated cluster with the SEM Warp-PLS approach can be seen in Table 3.

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, all the criteria have reached the expected value limit or have met the recommended Goodness of fit indices critical limit, so that the results of this modeling can be accepted or worthy of analysis. It can be stated that this test results in good confirmation of the variables as well as the causality relationship between variables. So, the overall model test shows good results or following expectations, meaning that the empirical data (field data) has supported the theoretical model developed.

4.4 Comparison of R^2 SEM Value with Warp-PLS approach and Integrated Cluster with Various Indices

In this study, the criteria for determining the best model for an integrated cluster with SEM with the Warp-PLS approach can be seen in Table 4.

Table 4. Comparison of R2 Value

Index	R2 value
Silhouette	0.264
Krzanowski-Lai	0.264
Dunn	0.264
GAP	0.264
Davies-Bouldin	0.264
C-index	0.990
Global Sillhouette	0.264
Goodman-Kruskal	0.264

Based on table 4, it can be seen that the integrated cluster model of the SEM Warp-PLS approach with the Silhouette, Krzanowski-Lai, Dunn, GAP, Davies-Bouldin, Global Sillhouette, and Goodman-Kruskal index has an R2 value of 0.264 which means the variable service quality, environment, fashions, and willingness to pay simultaneously affect obedient pay behavior by 26.4%, while the remaining 73.6% is influenced by other variables. The integrated cluster model of the SEM Warp-PLS approach with index C has an R2 value of 0.990 which means that the variables of service quality, environment, fashions, and willingness to pay simultaneously affect compliance pay behavior by 99.0%, while the remaining 1.0% is influenced by other variables. Based on the value of R^2,

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

$$Y_1 = 0,131 D_1 + 0,049 X_1 + 0,541 X_2 + 0,651 X_3 + 0,029 D_2 X_1 + 0,155 D_3 X_2 - 0,026 D_4 X_3$$

Cluster 1 (D = 1):

$$Y_1 = 0,131 + 0,078 X_1 + 0,696 X_2 + 0,625 X_3 \quad (15)$$

Cluster 2 (D = 0):

$$Y_1 = 0,049 X_1 + 0,541 X_2 + 0,651 X_3 \quad (16)$$

Based on equations 15 and 16, it can be concluded that service quality (X1) and environment (X2) in cluster 1 have a greater influence than cluster 2. While fashions (X3) in cluster 2 have a greater influence than cluster 1. 1, every increase of one environmental quality unit (X1) will increase the customer's willingness to pay (Y1) by 0.078 units. Every increase of one environmental unit (X2) for the customer in cluster 1, it will increase the customer's willingness to pay (Y1) by 0.696 units. Also, each increase of one unit of customer fashions (X3) in cluster 1 will increase the customer's willingness to pay (Y1) by 0.625 units.

In cluster 2, each increase of one environmental quality unit (X1) will increase the customer's willingness to pay (Y1) by 0.049 units. Every increase of one environmental unit (X2) for a customer in cluster 2, it will increase the customer's willingness to pay (Y1) by 0.541 units. Also, each increase of one unit of customer fashions (X3) in cluster 2 will increase the customer's willingness to pay (Y1) by 0.651 units.

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

$$Y_2 = 0,256 D_1 + 0,389 X_1 + 0,556 X_2 - 0,011 X_3 + 0,045 Y_1 + 0,076 D_2 X_1 + 0,158 D_3 X_2 + 0,282 D_4 X_3 + 0,287 D_5 Y_1$$

Cluster 1 (D = 1):

$$Y_2 = 0,256 + 0,465 X_1 + 0,714 X_2 + 0,271 X_3 + 0,332 Y_1 \quad (17)$$

Cluster 2 (D = 0):

$$Y_2 = 0,049 X_1 + 0,541 X_2 + 0,651 X_3 + 0,045 Y_1 \quad (18)$$

Based on equations 17 and 18, it can be concluded that service quality (X1), environment (X2), and willingness to pay (Y1) in cluster 1 have a greater influence than cluster 2. While fashions (X3) in cluster 2 have a greater influence. is bigger than cluster 1. In cluster 1, every increase of one environmental quality unit (X1) will increase customers' compliance to pay behavior (Y2) by 0.465 units. Every increase of one environmental unit (X2) of customers in cluster 1, it will increase the customer compliance behavior (Y2) of 0.714 units. Every increase of one unit of customer

fashions (X3) in cluster 1, it will increase customers' compliance to pay behavior (Y2) by 0.271 units. Also, for each increase of one unit of willingness to pay (Y1) of customers in cluster 1,

In cluster 2, each increase of one environmental quality unit (X1) will increase the customer's obedient pay behavior (Y2) by 0.049 units. Every increase of one environmental unit (X2) for the customer in cluster 2, it will increase the customer compliance behavior (Y2) by 0.541 units. Every increase of one unit of customer fashions (X3) in cluster 2, it will increase customers' compliance to pay behavior (Y2) by 0.651 units. Also, every increase of one unit of willingness to pay (Y1) of customers in cluster 1, it will increase the compliance behavior of customers (Y2) by 0.045 units.

5. CONCLUSIONS AND SUGGESTIONS

The limitation in this study is that it only uses eight cluster validity indices, namely statistical gap, silhouette, krznawski-lai, goodman kruskal, dunn, global silhouette, index C, and Davies-Bouldin. In addition, this study also uses data from one of the state-owned banks in Indonesia. This study also limits using only Euclidean distance and average linkage. The conclusion that can be given based on the results of the analysis is the application of an integrated cluster in SEM with the Warp-PLS approach with various cluster validity index methods resulting in many clusters and the same cluster members causing the same dummy variables. The value of R^2 in the integrated cluster with the C index is better than the Silhouette, Krzanowski-Lai, Dunn, GAP, Davies-Bouldin, Global Sillhouette, and Goodman-kruskal indexes. Service quality variables (X1), environment (X2), and willingness to pay (Y1) in cluster 1 have a greater influence than cluster 2. While fashions (X3) in cluster 2 have a greater influence than cluster 1. These results are in line with the research conducted by [31], [32], and [33].

Suggestions that can be given are based on the results of the integrated cluster on SEM with the Warp-PLS approach, namely for further research to compare the effect of using linkage, as well as the distance on the discriminant integrated cluster analysis which results in a high R2 value.

REFRENCES:

[1] Jain, AK, & Ambassador, RC. "Algorithms for data clustering". Prentice-Hall, Inc. 1988.

[2] Siswadi and B. Suharjo. "Multiple Variable Data Exploration Analysis". Final Project Not Published. Bogor: Department of Mathematics, Faculty of Mathematics and Natural Sciences IPB, Bogor. 1998.

[3] Johnson, RA and Wichern, DW. "Applied Multivariate Analysis", Third Edition, New Jersey: Prentice Hall Inc. 1992.

[4] Supranto. "Multivariate Analysis of Meaning and Interpretation", Jakarta: PT. Rineka Cipta. 2004.

[5] Hair, JF, Anderson, RE, Tatham, RL, and Black, WC. "Multivariate Data Analysis. Fifth edition". Jakarta: Gramedia. Main Library. 2006.

[6] Irawan, B. "Stabilization of Upland Agriculture Under El Nino-Induced Climatic Risk: Impact Assessment and Mitigation Measures in Indonesia". No. 1438-2016-118920. 2002.

[7] 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.

[8] Krazanowski, WJ, & Lai, YT. "A criterion for determining the number of groups in a data set using sum of squares clustering". Biometrics, Vol. 44. 1988. pp. 23–34.

[9] Bolshakova, N., & Azuaje, F. "Cluster validation techniques for genome expression data". Signal processing, Vol. 83 No. 4. 2003. pp. 825-833.

[10] Wani, MA, & Riyaz, R. "A novel point density based validity index for clustering gene expression datasets". International Journal of Data Mining and Bioinformatics, Vol. 17 No. 1, 2017. pp. 66-84.

[11] Bates, A., & Kalita, J. "Counting clusters in twitter posts. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies". 2016, March. pp. 1-9.

[12] Fornell, C., & Larcker, DF. "Structural equation models with unobservable variables and measurement error: Algebra and statistics". Division of Research Journal. Vol. 266. 1981.

[13] Solimun. "Structural Equation Modeling (SEM) Lisrel and Amos". Malang: Faculty of Mathematics and Natural Sciences, Brawijaya University. 2002.

[14] Ghozali, I. "Structural equation modeling: An alternative method with partial least squares (pls)". Diponegoro University Publishing Agency. 2008.

[15] Solimun. "Multivariate Analysis of Structural Modeling". Malang: CV. Image of Malang. 2010.
[16] Solimun, Fernandes, AAR, & Nurjannah. "Multivariate Statistical Methods of Structural Equation Modeling (SEM) WarpPLS Approach". Malang: UB Press. 2017.
[17] Kotler, Philip. "Marketing Management Volume 1 (11th ed.)". PT. Index. Jakarta. 2005.
[18] Putro, SW. "The Effect of Service Quality and Product Quality on Customer Satisfaction and Consumer Loyalty in Happy Garden Restaurants". Journal of Marketing Strategy, Vol. 2 No. 1. 2014. pp. 1-9.
[19] Soeriaatmadja, RE. "Environmental Science". Bandung: ITB Publisher. 1997.
[20] Robbinson and Stephen. "Organizational Behavior Controversy Concept, Application", Jakarta Prehalinda. 2002.
[21] Thio, Alex. "Sociology (An Introduction)". New York: Westview. 1987.
[22] Golin, J., & Delhaise, P. "The bank credit analysis handbook: a guide for analysts, bankers, and investors". John Wiley & Sons. 2013.
[23] Budiono, D. "The behavior of Corporate Taxpayers in Fulfilling Tax Obligations: Humanistic Theory Perspective". 2016.
[24] Sulistiyono, A., & Ayuvisda, ADINCHA. "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, Vol. 1 No. 1. 2012.
[25] Siat, CC, & Toly, AA. "Factors Affecting Taxpayer Compliance in Fulfilling Tax Obligations in Surabaya". Tax & Accounting Review, Vol. 1 No. 1. 2013. 41.
[26] Bambang & Widi. "The Influence of Attitudes, Subjective Norms, Perceived Behavior Control, and Sunset Policy on Taxpayer Compliance with Intention as an Intervening Variable". 2010
[27] Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, MM. "Package 'nbclust'". Journal of statistical software. Vol. 61 No. (6). 2014. pp. 1-36.
[28] Hox, JJ, & Bechger, TM. "An introduction to structural equation modeling". Family Science Review. Vol. 11. 1998. pp. 354-373.
[29] Munadjat D. "Environmental law (book V: Sectoral): Indonesian environmental law (in the National & International system)". Bandung: Binacipta. 1984
[30] Permadi, T., Nasir, A., & Anisma, Y. "Study of willingness to pay taxes on individual taxpayers who do independent work (the case at KPP Pratama Tampan Pekanbaru)". Journal of Economics, Vol. 21 No. 02. 2013.
[31] Shim, Y., Chung, J., & Choi, I. C. "A comparison study of cluster validity indices using a nonhierarchical clustering algorithm". International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). Vol. 1, 2005, November. pp. 199-204). IEEE.
[32] Bandyopadhyay, S., & Maulik, U. "Nonparametric genetic clustering: comparison of validity indices". IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). Vol 31 No. 1. 2001, 120-125.
[33] Legany, C., Juhasz, S., & Babos, A. "Cluster validity measurement techniques". Proceedings of the 5th WSEAS international conference on artificial intelligence, knowledge engineering and data bases (pp. 388-393). World Scientific and Engineering Academy and Society (WSEAS) Stevens Point, Wisconsin, USA. 2006, February.

Journal of Theoretical and Applied Information Technology
30th September 2021. Vol.99. No 18

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

APPENDIX

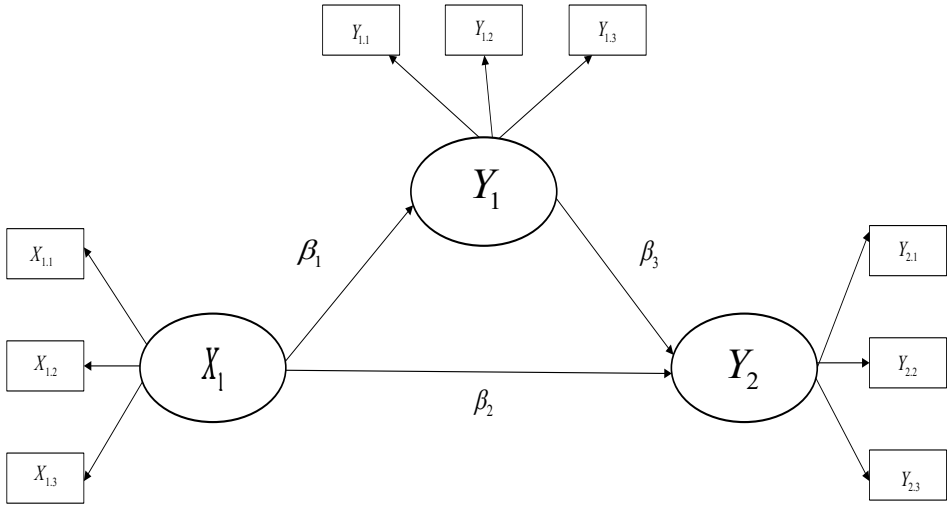

Figure 1: Structural Model Illustration

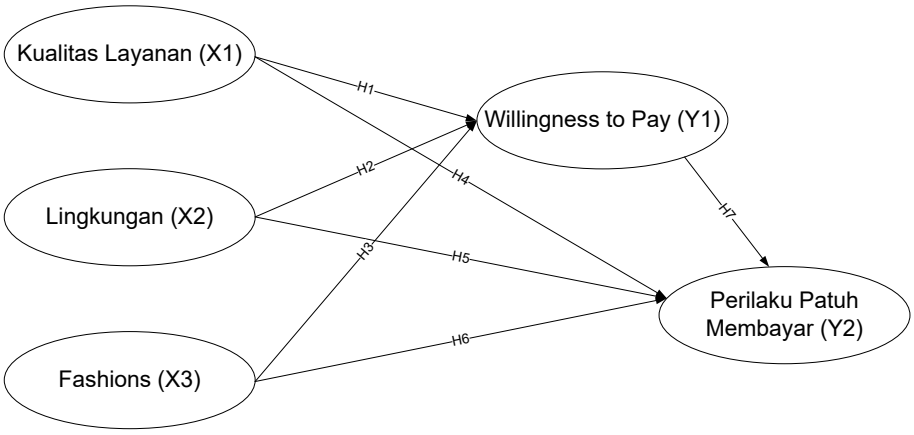

Figure 2. Research Hypothesis Model

Table 1. Average Cluster Members for Each Index

Index	Average									
	X1		X2		X3		Y1		Y2	
	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
1. Silhouette	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.11
2. Krzanowski-Lai	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.11
3. Dunn	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
4. GAP	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
5. Davies-Bouldin	3.94	3.34	4.06	3.30	3.92	3.39	3.98	3.30	3.97	3.32
6. C-index	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.11
7. Global Sillhouette	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12
8. Goodman-Kruskal	3.94	3.17	4.05	3.10	3.93	3.23	3.97	3.13	4.00	3.12

Source: Primary Data Processed (2021)

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

Table 2. Model Feasibility Test Results for Integrated Cluster with SEM Approach Warp-PLS Silhouette Index

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 = 123.7	acceptable if AVIF ≤ 5 ideal if AVIF ≤ 3,3	Rejected
5	Average full collinearity VIF	AFVIF = 129.4	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 = 0.725	acceptable if the SPR ≥ 0.7 ideal if SPR = 1	Acceptable
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 3. Model Feasibility Test Results for Integrated Cluster with SEM Approach Warp-PLS Index C

No.	Model Fit / Quality Index	Score	Criteria	Information
1	Average path coefficient	APC = 0.228P = 0.004	P <0.05	Significant
2	Average R-squared	ARS = 0.988 P <0.001	P <0.05	Significant
3	Average adjusted R-squared	AARS = 0.987P <0.001	P <0.05	Significant
4	Average block VIF	AVIF = 3,605	acceptable if AVIF ≤ 5 ideal if AVIF ≤ 3,3	Acceptable
5	Average full collinearity VIF	AFVIF = 65.77	acceptable if AFVIF ≤ 5 ideal if AFVIF ≤ 3,3	Rejected
6	Tenenhaus GoF	GoF = 0.880	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 = 0.990	acceptable if RSCR ≥ 0.9 ideal RSCR = 1	Acceptable
9	Statistical suppression ratio	SSR = 0.938	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)