SIMULATION STUDY FOR UNDERSTANDING THE PERFORMANCE OF PARTIAL LEAST SQUARES–MODIFIED FUZZY CLUSTERING (PLSMFC) IN FINDING GROUPS UNDER STRUCTURAL EQUATION MODEL

Moch. Abdul Mukid[1], Bambang Widjanarko Otok[2], Suparti[1]
[1] Department of Statistics, Universitas Diponegoro, Indonesia
[2] Department of Statistics, Institut Teknologi Sepuluh Nopember, Indonesia

e-mail: *mamukid@live.undip.ac.id*

Abstract: In structural equation modeling (SEM), it is usually assumed that all observations follow only one model. This becomes irrelevant if the observations contain natural groups, each of which has a different SEM model. Mukid et al (2002) have proposed the partial least squares-modified fuzzy clustering method (PLSMFC) as a way to find groups of observations and at the same time estimate the parameters of the SEM model. This research aims to understand the performance of the PLSMFC method in finding groups of observations characterized by different forms of structural equation models. The goal was achieved by conducting a simulation study involving factors such as SEM model specification and number of clusters. The procedure used is to force the generated data into a different number of segments. The segment validity measures used are the fuzziness performance index (FPI) and normalized classification entropy (NCE). The correct number of segments is indicated by the smallest FPI and NCE values. Based on simulation studies, it is known that the PLSMFC method can detect segments accurately, especially if the size of the segments used to reallocate observations is larger than the number of segments used to generate the data.

1. INTRODUCTION

Structural Equation Modeling (SEM) has become a statistical tool that is widely used to investigate the relationship between latent variables, especially for social research (Aktepe et al., 2015; Gokarna et al., 2022; Chuah et al, 2021; Hidayat et al., 2018; Otok et al., 2018). The application of SEM usually assumes that a set of data only follows a model. In certain cases, this assumption may not be met. The existence of "natural" groups in the data structure is the cause of failed efforts to "fit" a model. If the existence of groups in this population is ignored, universal solutions may not be appropriate and will provide inaccurate and/or inadequate results (Lamberti et al., 2016). Unfortunately, conventional clustering techniques for manifest variables or latent variable scores do not take into account the relationship between latent variables at all. On the other hand, the local model (each group) obtained by cluster analysis of the latent variable scores will give rise to group average differences in the latent variables but not necessarily different SEM models. The same method performed on manifest variables will most likely not produce different and well-separated models, both in

terms of model parameters and latent variable mean scores. In addition, the clustering procedure may present some theoretical problems as traditional cluster analysis assumes independence between variables, whereas SEM is based on the assumption that latent or manifest variables are correlated (Vinzi et al., 2008).

Several researchers have proposed various segmentation techniques in the context of SEM to overcome the heterogeneity in this model, for example, finite mixture SEM (Jedidi et al., 1997), finite mixture PLS (Hahn et al., 2002), response based unit segmentation PLS (REBUS PLS) (Vinzi et al., 2008), hierarchical bayesian SEM (Ansari et al., 2014), PLS Iterative Reweighted Regression Segmentation (PLS-IRRS) (Schlittgen et al., 2016) and PLS-SEM-KM (Fordellone & Vichi, 2020). Furthermore, a segmentation method based on the PLS genetic algorithm (PLS-GAS), which uses guided random search to find an optimal solution in a complex search space, was proposed by Ringle et al. (2013).

A fuzzy approach for segment discovery according to the PLS SEM model has also been studied (Mukid et al., 2022). There are several reasons to adopt a fuzzy clustering approach. Firstly, the fuzzy clustering algorithm is interesting in the context that the proposed method is compatible with the optimization procedure of distribution-free PLS methods. Second, due to the difficulty of identifying clear boundaries between segments in real-world problems, partial classification from fuzzy clustering appears more attractive than deterministic classification such as K-means. Third, fuzzy clustering algorithms are computationally more efficient because of the dramatic changes in segment membership values (McBratney & Moore, 1985). Fourth, fuzzy clustering is less affected by local optimality problems (Heiser & Groenen, 1997). Reviews of the performance of several segmentation methods in the context of PLS-SEM have also been carried out with the conclusion that there is no universal method that can be used for all cases (Sarstedt, 2008; Sarstedt et al., 2022).

This research examines the partial least squares-modified fuzzy clustering (PLSMFC) method proposed by Mukid et al (2022) through a simulation study. The main goal of this simulation study is to understand how well the method is at finding the right number of segments.

2. LITERATURE REVIEW
2.1. Partial Least Squares – Modified Fuzzy Clustering for SEM

The segmentation method studied in this research is a combination of the PLS method and modified fuzzy clustering. Each method has a different role. The PLS method is used to estimate SEM parameters and the fuzzy c-means method is used to find segments of the observations. The PLSMFC method (Mukid et al., 2022) is an alternative to finding data segments where each segment is characterized by different SEM models. To achieve this goal, the objective function used is the total weighted sum of squared errors which is as follows:

$$F = \sum_{j^*=1}^{J^*} \sum_{c=1}^{C} \sum_{n=1}^{N} u_{cn}^2 \hat{\zeta}_{j^*cn}^2 + \sum_{j=1}^{J_R} \sum_{k_j=1}^{K_j} \sum_{c=1}^{C} \sum_{n=1}^{N} u_{cn}^2 \hat{\varepsilon}_{jck_jn}^2 \qquad (1)$$

$$+ \sum_{j=1}^{J_F} \sum_{c=1}^{C} \sum_{n=1}^{N} u_{cn}^2 \hat{\delta}_{jcn}^2$$

where J^* is the number of endogenous latent variables, J_R is the number of latent variables measured using the reflexive model, J_F is the number of latent variables measured using the formative model, and $\sum_{c=1}^{C} u_{cn} = 1$, for each $n = 1, 2, ..., N$. $\hat{\zeta}_{j^*cn}^2$ is the residual of the inner model corresponding to the nth observation, the j^*-th endogenous latent variable in the c-th

cluster. $\hat{\varepsilon}_{jk_jcn}^2$ is the residual of the outer model for the reflexive measurement model at the nth observation in the kth indicator, the j-th latent variable in cluster c. $\hat{\delta}_{jcn}^2$ is the residual of the outer model for the formative measurement model at the nth observation in the c-th cluster associated with the j-th latent variable. Equation (1) is a modified version of the objective function of fuzzy clusterwise regression (Jajuga, 1986; Wedel & Steenkamp, 1989). We use $m = 2$ as a fuzzifier parameter. Equation (1) can be expressed as vector and matrix notation,

$$F = \sum_{j^*=1}^{J^*} \sum_{c=1}^{C} \left(\mathbf{U}_c \hat{\boldsymbol{\zeta}}_{cj^*}\right)^T \mathbf{U}_c \hat{\boldsymbol{\zeta}}_{cj^*} + \sum_{j=1}^{J_R} \sum_{k_j=1}^{K_j} \sum_{c=1}^{C} \left(\mathbf{U}_c \hat{\boldsymbol{\varepsilon}}_{cjk_j}\right)^T \mathbf{U}_c \hat{\boldsymbol{\varepsilon}}_{cjk_j} \tag{2}$$

$$+ \sum_{j=1}^{J_F} \sum_{c=1}^{C} \left(\mathbf{U}_c \hat{\boldsymbol{\delta}}_{cj}\right)^T \mathbf{U}_c \hat{\boldsymbol{\delta}}_{cj}$$

where $\hat{\boldsymbol{\zeta}}_{cj^*} = \begin{bmatrix} \hat{\zeta}_{cj^*1} \\ \hat{\zeta}_{cj^*2} \\ \vdots \\ \hat{\zeta}_{cj^*N} \end{bmatrix}, \hat{\boldsymbol{\varepsilon}}_{cjk_j} = \begin{bmatrix} \hat{\varepsilon}_{cjk_j1} \\ \hat{\varepsilon}_{cjk_j2} \\ \vdots \\ \hat{\varepsilon}_{cjk_jN} \end{bmatrix}, \hat{\boldsymbol{\delta}}_{cj} = \begin{bmatrix} \hat{\delta}_{cj1} \\ \hat{\delta}_{cj2} \\ \vdots \\ \hat{\delta}_{cjN} \end{bmatrix}, \mathbf{U}_c = \begin{bmatrix} u_{c1} & 0 & 0 & 0 \\ 0 & u_{c2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & u_{cN} \end{bmatrix}$

The Lagrange function containing a constraint $\sum_{c=1}^{C} u_{cn} - 1 = 0$ can be expressed as

$$F^* = \sum_{j^*=1}^{J^*} \sum_{c=1}^{C} \left[\left(\boldsymbol{\eta}_{j^*} - \boldsymbol{\eta}_{\to j^*}\hat{\boldsymbol{\beta}}_{j^*c}\right)^T \mathbf{U}_c^T\right] \left[\mathbf{U}_c\left(\boldsymbol{\eta}_{j^*} - \boldsymbol{\eta}_{\to j^*}\hat{\boldsymbol{\beta}}_{j^*c}\right)\right] \tag{3}$$

$$+ \sum_{j=1}^{J_R} \sum_{k_j=1}^{K_j} \sum_{c=1}^{C} \left[\left(\mathbf{x}_{jk_j} - \boldsymbol{\eta}_j\hat{\lambda}_{jk_jc}\right)^T \mathbf{U}_c^T\right] \left[\mathbf{U}_c\left(\mathbf{x}_{jk_j} - \boldsymbol{\eta}_j\hat{\lambda}_{jk_jc}\right)\right]$$

$$+ \sum_{j=1}^{J_F} \sum_{c=1}^{C} \left[\left(\boldsymbol{\eta}_j - \mathbf{X}_j\hat{\boldsymbol{\Lambda}}_{jc}\right)^T \mathbf{U}_c^T\right] \left[\mathbf{U}_c\left(\boldsymbol{\eta}_j - \mathbf{X}_j\hat{\boldsymbol{\Lambda}}_{jc}\right)\right] + \lambda\left(\sum_{c=1}^{C} u_{cn} - 1\right)$$

Estimation of parameters in the inner model is obtained through the first derivative of F^* at $\hat{\boldsymbol{\beta}}_{j^*c}$. $\frac{\partial F^*}{\partial \hat{\boldsymbol{\beta}}_{j^*c}} = 0$; for certain j^* and c.

$$\hat{\boldsymbol{\beta}}_B = \left[\mathbf{Y}_{\to j^*}^T \mathbf{V}_c \mathbf{Y}_{\to j^*}\right]^{-1} \left[\mathbf{Y}_{\to j^*}^T \mathbf{V}_c \mathbf{y}_{j^*}\right] \tag{4}$$

where $\mathbf{V}_c = \begin{bmatrix} u_{c1}^2 & 0 & 0 & 0 \\ 0 & u_{c2}^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & u_{cN}^2 \end{bmatrix}$

Estimation of parameters in the outer model for the reflexive measurement model is obtained through $\frac{\partial F^*}{\partial \hat{\lambda}_{jk_jc}} = 0$; for certain j, k, and c.

$$\hat{\lambda}_{jk_jc} = \left[\mathbf{y}_j^T \mathbf{V}_c \mathbf{y}_j\right]^{-1} \left[\mathbf{y}_j^T \mathbf{V}_c \mathbf{x}_{jk_j}\right]; k_j = 1,2,\dots,K_j \tag{5}$$

Estimation of parameters in a formative measurement model is obtained through $\frac{\partial F^*}{\partial \hat{\boldsymbol{\Lambda}}_{jc}} = 0$; for certain j and c.

$$\hat{\boldsymbol{\Lambda}}_{jc} = \left[\mathbf{X}_j^T \mathbf{V}_c \mathbf{X}_j\right]^{-1} \left[\mathbf{X}_j^T \mathbf{V}_c \mathbf{y}_j\right] \tag{6}$$

The fuzzy membership value is updated by using a formula in equation (7). The formula is obtained by first deriving F^* at u_{cn} and then equalizing it to zero, $\frac{\partial F^*}{\partial u_{cn}} = 0$, for certain n and c.

$$u_{cn} = \left[\sum_{c^*=1}^{C} \left[\frac{\left[\sum_{j^*=1}^{J^*} \hat{\zeta}_{nj^*c}^2 + \sum_{j=1}^{J_R} \sum_{k_j=1}^{K_j} \hat{\varepsilon}_{njk_jc}^2 + \sum_{j=1}^{J_F} \hat{\delta}_{njc}^2 \right]}{\left[\sum_{j^*=1}^{J^*} \hat{\zeta}_{nj^*c^*}^2 + \sum_{j=1}^{J} \sum_{k_j=1}^{K_j} \hat{\varepsilon}_{njk_jc^*}^2 + \sum_{j=1}^{J} \hat{\delta}_{njc^*}^2 \right]} \right] \right]^{-1} \tag{7}$$

Minimization of the objective function in equation (1) to find SEM model in several segments is carried out using the following algorithm:

1. Set the number of segment C, and the initial membership value u_{cn}
2. Estimate the weights and scores of latent variables using the PLS algorithm
3. Estimate the path and loading coefficients in the c-th segment using equations (4), (5), and (6).
4. Calculate the residual of the inner model and outer model in the c-th segment
5. Update the fuzzy membership value for the n-th observation in the c-th cluster using equation (7)
6. Calculate the objective function in the equation (1)
7. Calculate FPI and NCE using equations (8) and (9)
8. Repeat steps 1 to 7 for a different number of segments
9. Determine the optimal number of segments based on FPI and NCE values

2.2. Segment Validity Measures

In fuzzy clustering, a measure called segment validity (Bezdek, Ehrlich, & Full, 1984) is used to examine the status of separate segments. Based on synthetic data analysis, it can be concluded that the Fuzziness Performance Index (FPI) and Normalized Classification Entropy (NCE) are the most useful segment validity measures for fuzzy clustering. The FPI formula is as follows:

$$FPI = 1 - \frac{C \times PC - 1}{C - 1} \tag{8}$$

where PC is the Partition Coefficient which is defined by

$$PC = \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} u_{cn}^2$$

where C is the number of clusters. The formula for NCE is as follows:

$$NCE = \frac{PE}{\log C} \tag{9}$$

where PE is Partition Entropy which is defined by

$$PE = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} u_{cn} \log u_{cn}$$

The smaller the FPI or NCE value, the better the cluster formed at separating objects from each other.

3. MATERIAL AND METHOD

This section discusses the simulation design and the results obtained from the simulation studies carried out. A simulation study was conducted to evaluate the performance of the PLSMFC method. Specifically, this simulation aims to find out how good this method is at finding the right number of segments.

3.1. Simulation Design

The simulation design considers two factors including the number of segments and model specifications. Table 1 below is a list of factors and levels examined in this simulation study. The SEM model specifications considered consist of two forms, namely model 1 and model 2. Model 1 refers to the SEM model where the measurement model for latent variables is reflective while model 2 refers to the SEM model where the measurement model for latent variables is reflective and formative. The path diagram of the SEM model used in this simulation looks like in Figure 1 and Figure 2. The loading coefficients for the number of segments 2 and 3 are set to be the same, both in model 1 and model 2. The loading and path coefficients can be seen in Table 2. In general, this study involved 2×2 = 4 combinations. Each combination will be replicated 100 times.

Table 1. Factor and Levels in Simulation Study

No	Factor	Level
1	SEM Specification	1. Model 1
		2. Model 2
2	Number of Segment	1. Number of Segment 2
		2. Number of Segment 3

Table 2. Loading and Path Coefficient in Each Segment

True Parameter	Segment 1	Segment 2	Segment 3
λ_{11}	0.50	0.65	0.80
λ_{12}	0.55	0.70	0.85
λ_{13}	0.60	0.75	0.90
λ_{21}	0.50	0.65	0.80
λ_{22}	0.55	0.70	0.85
λ_{23}	0.60	0.75	0.90
λ_{31}	0.50	0.65	0.80
λ_{32}	0.55	0.70	0.85
λ_{33}	0.60	0.75	0.90
λ_{41}	0.50	0.65	0.80
λ_{42}	0.55	0.70	0.85
λ_{43}	0.60	0.75	0.90
λ_{51}	0.50	0.65	0.80
λ_{52}	0.55	0.70	0.85
λ_{53}	0.60	0.75	0.90
β_{21}	0.50	0.70	0.90
β_{31}	0.50	0.70	0.90
β_{41}	0.50	0.70	0.90
β_{51}	0.50	0.70	0.90

In this research, parameters in the SEM model in each segment are set as in Table 2. If the number of segments is two, then the SEM model parameters in each segment are as in columns 1 and 2, whereas if the number of segments is three, then the model parameters SEM in each segment are as in columns 1, 2, and 3.

3.2. Data Generation Proccess

This section explains the data generation process for model 1 and model 2. The specifications for model 1 and model 2 have different data generation methods.

3.2.1. Data Generation for Model 1

Synthetic data for model 1 were generated by following the following procedure:

1) Generate score of exogenous latent variable for η_2, η_3, η_4, and η_5 from N(0, 1).
2) Generate ζ_1 from Normal$(0, 0.05)$.
3) Compute score of endogenous latent variable η_1 based on the *inner model*
$$\eta_1 = \beta_{21}\eta_2 + \beta_{31}\eta_3 + \beta_{41}\eta_4 + \beta_{51}\eta_5 + \zeta_1$$
4) Generate $\varepsilon_{11}, \varepsilon_{12}$, and ε_{13} from distribution Normal$(0, 0.05)$
5) Compute score of indicators η_1 using formula
$$x_{11} = \lambda_{11}\eta_1 + \varepsilon_{11}$$
$$x_{12} = \lambda_{12}\eta_1 + \varepsilon_{12}$$
$$x_{13} = \lambda_{13}\eta_1 + \varepsilon_{13}$$
6) Generate $\varepsilon_{21}, \varepsilon_{22}$, and ε_{23} from distribution Normal$(0, 0.05)$
7) Compute score of indicators η_2 using formula
$$x_{21} = \lambda_{21}\eta_2 + \varepsilon_{21}$$
$$x_{22} = \lambda_{22}\eta_2 + \varepsilon_{22}$$
$$x_{23} = \lambda_{23}\eta_2 + \varepsilon_{23}$$
8) Generate $\varepsilon_{31}, \varepsilon_{32}$, and ε_{33} from distribution Normal$(0, 0.05)$
9) Compute score of indicators η_3 using formula
$$x_{31} = \lambda_{31}\eta_3 + \varepsilon_{31}$$
$$x_{32} = \lambda_{32}\eta_3 + \varepsilon_{32}$$
$$x_{33} = \lambda_{33}\eta_3 + \varepsilon_{33}$$
10) Generate $\varepsilon_{41}, \varepsilon_{42}$, and ε_{43} from distribution Normal $(0, 0.05)$
11) Compute score of indicators η_4 using formula
$$x_{41} = \lambda_{41}\eta_4 + \varepsilon_{41}$$
$$x_{42} = \lambda_{42}\eta_4 + \varepsilon_{42}$$
$$x_{43} = \lambda_{43}\eta_4 + \varepsilon_{43}$$
12) Generate $\varepsilon_{51}, \varepsilon_{52}$, and ε_{53} from distribution Normal$(0, 0.05)$
13) Compute score of indicators η_5 using formula
$$x_{51} = \lambda_{51}\eta_5 + \varepsilon_{51}$$
$$x_{52} = \lambda_{52}\eta_5 + \varepsilon_{52}$$
$$x_{53} = \lambda_{53}\eta_5 + \varepsilon_{53}$$

3.2.2. Data Generation for Model 2

The simulation data for model 2 has a different data generation procedure from model 1 because model 2 contains a mixed measurement model between reflective and formative. Simulation data from model 2 are obtained by the following procedure:

1) Generate score of indicators x_{21}, x_{22}, and x_{23} from distribution Normal$(0,1)$
2) Generate ζ_2 from distribution Normal$(0,0.05)$
3) Compute score of exogenous latent variable η_2 using formula
$$\eta_2 = \lambda_{21}x_{21} + \lambda_{22}x_{22} + \lambda_{23}x_{23} + \zeta_2$$
4) Generate score of indicators x_{31}, x_{32}, and x_{33} from distribution Normal$(0,1)$
5) Generate ζ_3 from distribution Normal$(0,0.05)$
6) Compute score of exogenous latent variable η_3 using formula
$$\eta_3 = \lambda_{31}x_{31} + \lambda_{32}x_{32} + \lambda_{33}x_{33} + \zeta_3$$

7) Generate score of exogenous latent variable for η_4 and η_5 from distribution Normal(0, 1).
8) Generate ζ_1 from distribution Normal(0,0.05).
9) Compute score of endogenous latent variable η_1 based on the inner model
$$\eta_1 = \beta_{21}\eta_2 + \beta_{31}\eta_3 + \beta_{41}\eta_4 + \beta_{51}\eta_5 + \zeta_1$$
10) Generate $\varepsilon_{11}, \varepsilon_{12},$ dan ε_{13} from distribution Normal(0; 0.05)
11) Compute score of indicators x_{11}, x_{12}, and x_{13} using formula
$$x_{11} = \lambda_{11}\eta_1 + \varepsilon_{11}$$
$$x_{12} = \lambda_{12}\eta_1 + \varepsilon_{12}$$
$$x_{13} = \lambda_{13}\eta_1 + \varepsilon_{13}$$
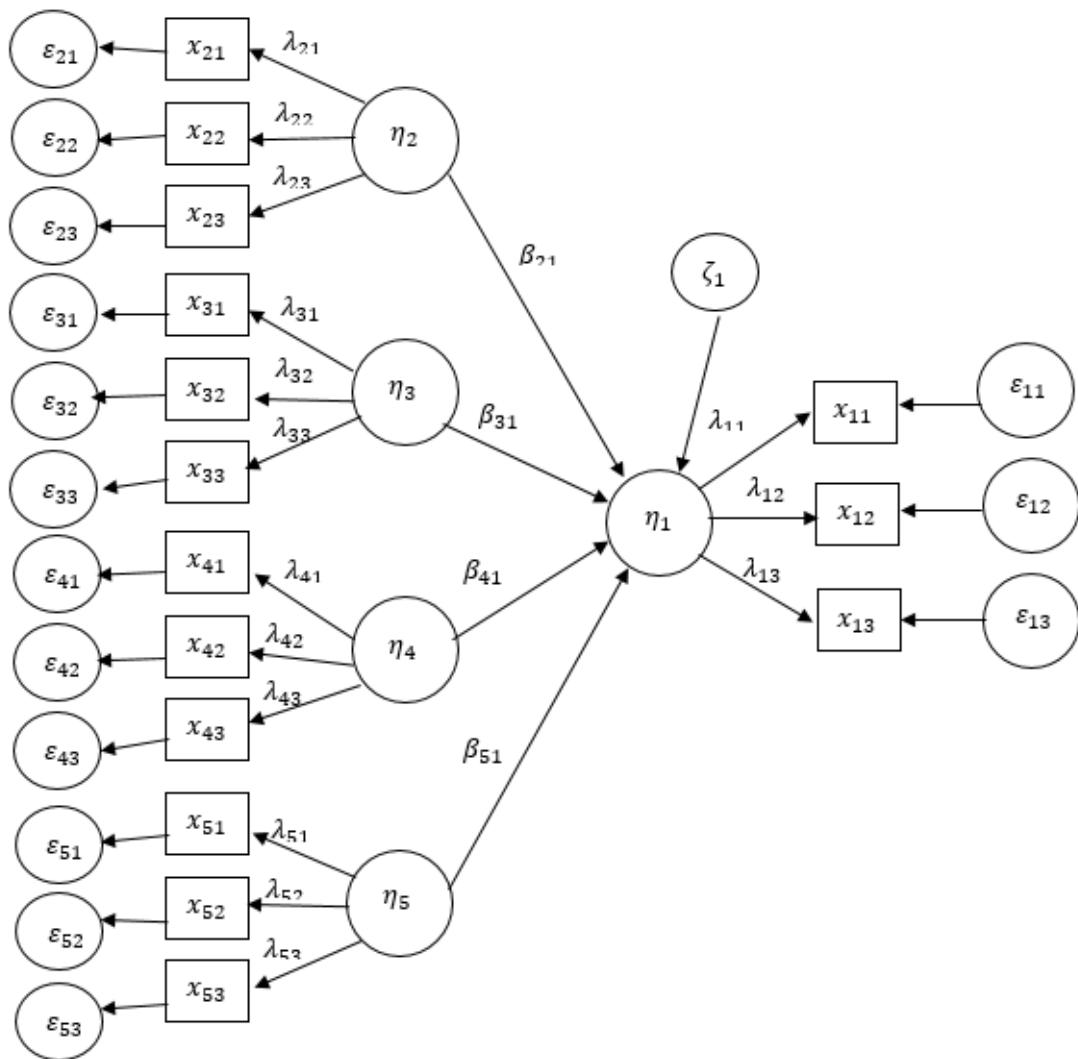
Figure 1. Path Diagram for Model 1
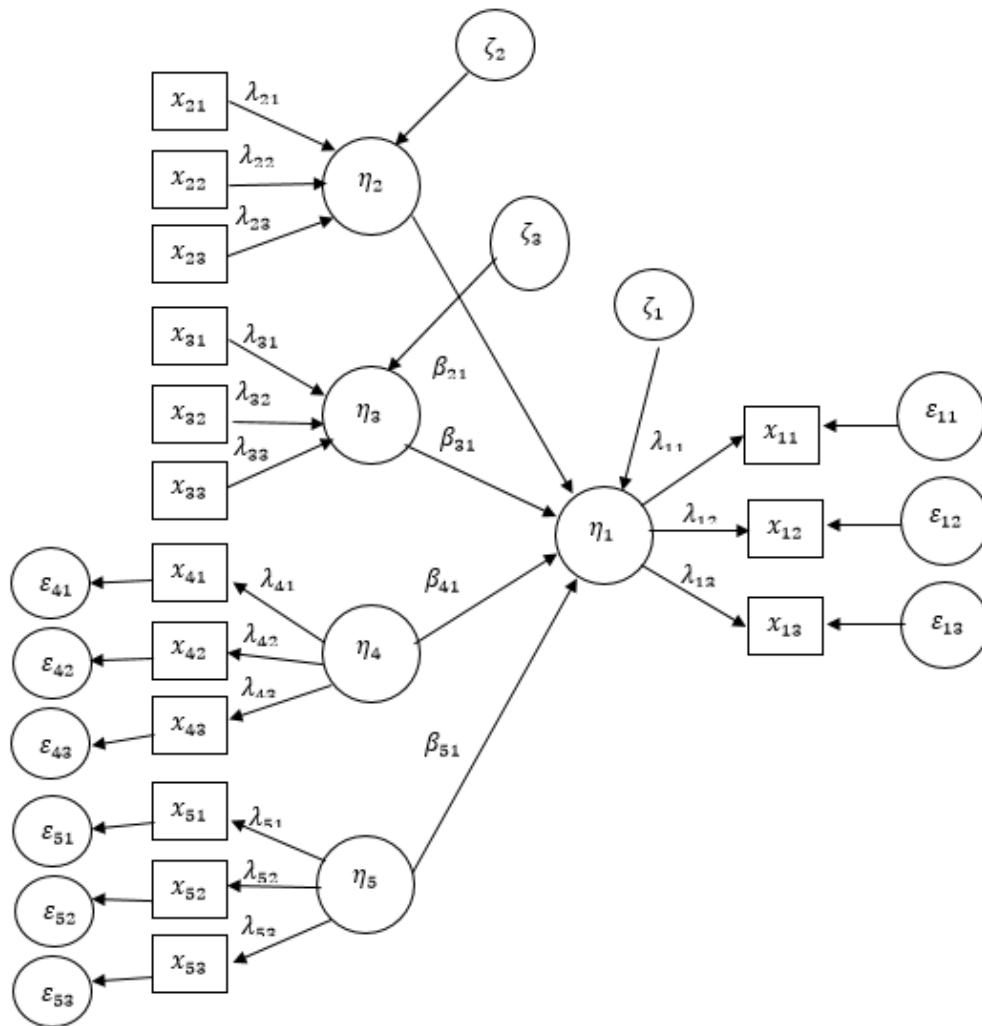
Figure 2. Path Diagram for Model 2

4. RESULTS AND DISCUSSION

In previous research, Mukid et al. (2022) carried out a simulation to see the performance of the PLSMFC method in its ability to reallocate observations to the right segments. In this simulation, the influence of five factors was examined for their influence on the ability to return the observation to its original segment. Unfortunately, the simulation study did not examine PLSMFC's ability to find the right number of segments. This section explains the performance of the PLSMFC method to recover the correct number of segments. The procedure used is to force the generated data into a different number of segments. In this research, the data generated uses scenarios of two and three segments. If initially the data is generated based on a scenario with several segments of two, it will be forced to enter several segments of three. If initially the data is generated based on a scenario with several segments of three, it will be forced to enter several segments of two or four.

The criteria for determining that the PLSMFC method can find the right number of segments is to look at the FPI and NCE values as in equations (8) and (9). The FPI or NCE value from the number of segments used to generate data will be compared with the FPI or NCE value obtained from the other number of segments. If the FPI or NCE value of the data generated from segment number two and re-entered into segment number two is smaller than the FPI or NCE value when entered into several segments of three, then the PLSMFC method

is said to be able to find the correct number of segments. Likewise, if the FPI or NCE value of data generated from three segments and re-entered into three segments is smaller than the FPI or NCE value when entered into two or four segments, the PLSMFC method is said to be able to find the correct number of segments.

Table 3. Average of FPI and NCE from Number of Segments Two and Three
(Data Generated from The Number of Segment Two)

Model	Allocated to The Number of Segment Two		Allocated to The Number of Segment Three	
	FPI	NCE	FPI	NCE
Model 1	0.6591*	0.7143*	0.8028	0.8127
Model 2	0.8605*	0.8875*	0.9153	0.9209

*) the smallest value when compared with the corresponding segment validity measures in different scenarios

Table 3 is the average FPI and NCE values obtained from the number of segments two and three where the data was generated from the number of segments two. In Model 1, the data generated from several segments of two and reinserted into the number of segments two has an FPI value of 0.6591 and an NCE value of 0.7143. Meanwhile, the data generated from segment number two and entered into segment number three has an FPI value of 0.8028 and an NCE value of 0.8127. The FPI and NCE values from the data scenario that is generated from the number of segments two and re-entered into the number of segments two are smaller than the FPI and NCE values from the data scenario that is generated from the number of segments two and re-entered into the number of segments three. Therefore, it can be concluded that in Model 1, the PLSMFC method can find the correct number of segments.

In Model 2, the data generated from several segments of two and by the PLSMFC method placed back into the number of segments two has an FPI value of 0.8605 and an NCE value of 0.8875. Meanwhile, the data generated from segment number two and placed into segment number three has an FPI value of 0.9153 and an NCE value of 0.9209. The FPI value from the data scenario that is generated from two segments and placed back into two segments is smaller than the FPI value from the data scenario that is generated from two segments and placed into three segments. The NCE value from the data scenario that is generated from two segments and placed back into two segments is smaller than the NCE value from the data scenario that is generated from two segments and placed into three segments. Therefore, based on the simulation study, it can be concluded that in model specification 2, the PLSMFC method can find the number of segments correctly.

Table 4. Average of FPI and NCE from Number of Segments Two and Three
(Data Generated from Number of Segments Three)

Model	Allocated to The Number of Segment Two		Allocated to The Number of Segment Three	
	FPI	NCE	FPI	NCE
Model 1	0.6590*	0.7151*	0.7328	0.7536
Model 2	0.8009*	0.8370*	0.8367	0.8513

*) the smallest value when compared with the corresponding validity measure in different scenarios

Table 4 is the average FPI and NCE values obtained from the number of segments two and three where the data was generated from the number of segments three. In Model 1, the data generated from the number of segments three and put back into the number of

segments three has an FPI value of 0.7328 and an NCE value of 0.7536. Meanwhile, the data generated from segment number three and entered into segment number two has an FPI value of 0.6590 and an NCE value of 0.7151. The FPI and NCE values from the data scenario that is generated from the number of segments three and re-entered into the number of segments three are greater than the FPI and NCE values from the data scenario that is generated from the number of segments three and re-entered into the number of segments two. Therefore, it can be concluded that based on the simulation study in model specification 1, the PLSMFC method is unable to find the correct number of segments.

In Model 2, the data generated from the number of segments three and put back into the number of segments three has an FPI value of 0.8367 and an NCE value of 0.8513. Meanwhile, the data generated from total segment three and entered into total segment two has an FPI value of 0.8009 and an NCE value of 0.8370. The FPI and NCE values from the data scenario that is generated from the number of segments three and re-entered into the number of segments three are greater than the FPI and NCE values from the data scenario that is generated from the number of segments three and re-entered into the number of segments two. Therefore, it can be concluded that in model 2, the PLSMFC method is unable to find the correct number of segments.

Table 5. Average of FPI and NCE from Number of Segments Three and Four
(Data Generated from Number of Segments Three)

Model	Allocated to The Number of Segment Three		Allocated to The Number of Segment Four	
	FPI	NCE	FPI	NCE
Model 1	0.7328[*]	0.7536[*]	0.7562	0.8035
Model 2	0.8367[*]	0.8513[*]	0.8771	0.8756

[*]) smallest value when compared with the corresponding validity measure in different scenarios

Table 5 shows the average FPI and NCE values obtained from three and four segments where the data was generated from three segments. In Model 1, the data generated from the number of segments three and put back into the number of segments three has an FPI value of 0.7328 and an NCE value of 0.7536. Meanwhile, the data generated from total segment three and entered into total segment four has an FPI value of 0.7562 and an NCE value of 0.8035. The FPI and NCE values from the data scenario that is generated from three segments and re-entered into three segments are smaller than the FPI and NCE values from the data scenario that is generated from three segments and re-entered into four segments. Therefore, it can be concluded that in Model 1, the PLSMFC method can find the correct number of segments.

In Model 2, the data generated from the number of segments three and put back into the number of segments three has an FPI value of 0.8367 and an NCE value of 0.8513. Meanwhile, the data generated from segment number three and entered into segment number four has an FPI value of 0.8771 and an NCE value of 0.8756. The FPI value from the data scenario that is generated from three segments and re-entered into three segments is smaller than the FPI value from the data scenario that is generated from three segments and re-entered into four segments. The NCE value from the data scenario that is generated from three segments and entered into three segments is also smaller than the NCE value from the data scenario that is generated from three segments and entered into four segments. Therefore, it can be concluded that based on the FPI and NCE values, the PLSMFC method can find the correct number of segments.

In general, it can be concluded that based on simulation studies, the PLSMFC method is capable of detecting segments correctly if the segment size used to reallocate observations is larger than the number of segments used to generate data.

5. CONCLUSION

This research examines a new method for estimating SEM parameters that contain heterogeneity in the populations. The method combines the Partial Least Squares (PLS) method and modified fuzzy clustering which allows an object to be in many segments even with different membership weights. PLSMFC method is a modification of the PLS SEM algorithm. PLS SEM algorithm is only intended for the process of estimating latent variable scores and estimating SEM parameters. This process is carried out partially on the parameters of the inner model and outer model. The PLSMFC method modifies the PLS SEM algorithm by adding one more process, namely segmentation (grouping) of observation objects. The segmentation process is carried out using a fuzzy approach where each object is given an initial weight randomly according to the specified number of segments. These weights are updated iteratively and sequentially by estimating the model parameters. The iteration process will stop when the difference in the objective function of two successive iterations is very small. Generally, it can be concluded that based on simulation studies, the PLSMFC method is capable of detecting segments correctly if the segment size used to reallocate observations is larger than the number of segments used to generate data.

REFERENCES

Aktepe, A., Ersöz, S., & Toklu, B. (2015). Computers & Industrial Engineering Customer Satisfaction and Loyalty Analysis with Classification Algorithms and Structural Equation Modeling. *Computers & Industrial Engineering*, *86*, 95–106.

Ansari, A., Jedidi, K., & Jagpal, S. (2014). A Hierarchical Bayesian Treating Equation Methodology for Structural Models in Heterogeneity. *Marketing Science*, *19*(4), 328–347.

Bai Gokarna, V., Mendon, S., Thonse Hawaldar, I., Spulbar, C., Birau, R., Nayak, S., & Manohar, M. (2022). Exploring the Antecedents of Institutional Effectiveness: A Case Study of Higher Education Universities in India. *Economic Research-Ekonomska Istrazivanja* , *35*(1), 1162–1182.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The Fuzzy c-Means Clustering Algorithm. *Computers & Geosciences*, *10*(2), 191–203.

Chuah, S. H. W., Tseng, M. L., Wu, K. J., & Cheng, C. F. (2021). Factors Influencing The Adoption of Sharing Economy in B2B Context in China: Findings from PLS-SEM and fsQCA. *Resources, Conservation and Recycling*, *175*(September), 105892.

Fordellone, M. & Vichi, M. (2020). Finding Groups in Structural Equation Modeling Through The Partial Least Squares Algorithm. *Computational Statistics and Data Analysis*, *147*, 106957.

Hahn, C.; Johnson, M.D.; Herrmann, A.; Huber, F. Capturing Customer Heterogeneity using a Finite Mixture PLS Approach. Schmalenbach Bus. Rev. 2002, 54, 243–269.

Heiser, W. J., & Groenen, P. J. F. (1997). Cluster Differences Scaling with A Within-Clusters Loss Component and A Fuzzy Successive Approximation Strategy to Avoid Local

Minima. In *Psychometrika* (Vol. 62, pp. 63–83).

Hidayat, R. N., Poernomo, E., Waluyo, M., & Otok, B. W. (2018). The Model of Risk of Travel Ticket Purchasing Decisions on Marketing Communication Mix in Online Site Using Structural Equation Modeling. *International Journal of Civil Engineering and Technology*, *9*(5), 847–856.

Hwang, H., DeSarbo, W. S., & Takane, Y. (2007). Fuzzy Clusterwise Generalized Structured Component Analysis. *Psychometrika*, *72*(2), 181–198.

Jajuga, K. (1986). Linear Fuzzy Regression. *Fuzzy Sets and Systems*, *20*(3), 343–353.

Jedidi, K., Jagpal, H. S., & Desarbo, W. S. (1997). Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity. *Marketing Science*, *16*(1), 39–59.

Lamberti, G., Aluja, T. B., & Sanchez, G. (2016). The Pathmox Approach for PLS Path Modeling Segmentation. *Applied Stochastic Models in Business and Industry*, *32*(4), 453–468.

Lohmöller, J. B. (2013). *Latent Variable Path Modeling with Partial Least Squares*. Berlin: Physica-Verlag HD.

McBratney, A. B., & Moore, A. W. (1985). Application of Fuzzy Sets to Climatic Classification. *Agricultural and Forest Meteorology*, *35*(1–4), 165–185.

Mukid, M. A., Otok, B. W., & Suharsono, A. (2022). Segmentation in Structural Equation Modeling Using a Combination of Partial Least Squares and Modified Fuzzy Clustering. *Symmetry*, *14*(11).

Otok, B. W., Sustrami, D., Hastuti, P., Purhadi, Sutikno, Purnami, S. W., & Suharsono, A. (2018). Structural Equation Modeling The Environment, Psychology, Social Relationships Against Physical Health in Determination Quality of Elderly Community Surabaya. *International Journal of Civil Engineering and Technology*, *9*(12), 926–938.

Sarstedt, M. (2008). A Review of Recent Approaches for Capturing Heterogeneity in Partial Least Squares Path Modelling. *Journal of Modelling in Management*, (July 2008). https://doi.org/10.1108/17465660810890126

Sarstedt, M., Radomir, L., Moisescu, O. I., & Ringle, C. M. (2022). Latent Class Analysis in PLS-SEM: A Review and Recommendations for Future Applications. *Journal of Business Research*, *138*(September 2021), 398–407.

Schlittgen, R., Ringle, C. M., Sarstedt, M., & Becker, J. (2016). Segmentation of PLS Path Models by Iterative Reweighted Regressions. *Journal of Business Research*, *69*(10), 4583–4592. https://doi.org/10.1016/j.jbusres.2016.04.009

Vinzi, V. E., Trinchera, L., Squillacciotti, S., & Tenenhaus, M. (2008). REBUS-PLS: A Response-Based Procedure for Detecting Unit Segments in PLS Path Modelling. *Applied Stochastic Models in Business and Industry*, *24*, 439–458. https://doi.org/10.1002/asmb

Wedel, M., & Steenkamp, J. B. E. M. (1989). A Fuzzy Clusterwise Regression Approach to Benefit Segmentation. *International Journal of Research in Marketing*, *6*(4), 241–258. https://doi.org/10.1016/0167-8116(89)90052-9