SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 3127-3138
Open Access. Online & Print

DOI: 10.5281/zenodo.122.126241

SEGMENTATION OF TUBERCULOSIS CASES IN EAST JAVA USING SEM-PLS WARD CLUSTERING

Bambang Widjanarko Otok[1], Akbar Maulana Ibrahim[2*], Purhadi[2], Anak Agung Bagus Wirayuda[3], Muhammad Sjahid Akbar[2], Andi Alfian Zainuddin[4], Riry Sriningsih[5], Isnawati[6]

[1,2]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya
[3]Medicine Study Program, Institut Teknologi Sepuluh Nopember, Surabaya
[4]Department of Public Health and Community Medicine, Universitas Hasanuddin
[5]Faculty of Mathematics and Natural Sciences, Universitas Negeri Padang
[6]Environmental Sanitation Study Program, Poltekkes Kemenkes Banjarmasin, South Kalimantan

Received: 11/12/2025
Accepted: 02/02/2026

Corresponding Author: Akbar Maulana Ibrahim (akbarmlnibrahim@gmail.com)

ABSTRACT

Tuberculosis is an infectious disease that is still a health issue faced by most countries in the world. This disease is not only a public health problem, but also affects social and economic aspects in various countries. East Java is one of the provinces with the highest number of TB cases. The management of this disease is not only related to medical factors, but also social, economic, and environmental factors. This study aims to group regions in East Java with similar characteristics to overcome data heterogeneity. This grouping is performed using a ward clustering method based on the pattern of errors in the initial (global) model. Data were obtained from secondary sources 2023 East Java Health Office and Central Statistics Agency publications and measured five key factors: TB incidence, community behavior, environmental health, social determinants, and access to health services. Overall, social determinants were found to significantly influence community behavior and environmental health. However, TB incidence was directly linked only to community behavior and social determinants. Segmentation of the global residual model with the Ward method resulted in 2 segments. Analysis specific to each segment provided better insights, showing that access to health services was the only key factor linked to TB incidence. In this regional view, determinant social factors did not directly drive TB incidence but still influenced community behavior and environmental health.

KEYWORDS: Segmentation, Tuberculosis, Ward Clustering.

1. INTRODUCTION

Tuberculosis remains a major infectious disease and a health issue facing most countries worldwide. This disease is not only a public health problem but also impacts social and economic aspects in various countries. Among the global efforts implemented to address this issue is the End TB Strategy launched by the World Health Organization (WHO). The WHO's End TB Strategy targets an 80% reduction in tuberculosis incidence and a 90% reduction in TB deaths by 2030 compared to 2015 (WHO, 2015). At the national level, the Indonesian Ministry of Health responded to this global challenge by issuing a national tuberculosis control strategy document for 2020-2024. This strategy is designed to address various aspects affecting the spread and control of tuberculosis in Indonesia (Kemenkes RI, 2023).

However, even though the national strategy has been carefully formulated, challenges in implementation in the field have caused TB cases in Indonesia to continue to increase, even reaching a peak in 2023. Based on the Global TB Report 2024, Indonesia is ranked second with the highest number of TB cases in the world after India (World Health Organization, 2023). In Indonesia, East Java is one of the provinces with a fairly high number of TB cases. As a province with a large population and a fairly high level of urbanization, factors such as population density, access to health services, environmental conditions, and the level of public awareness about TB contribute to the high number of cases in this area. It was recorded that in 2023, the number of suspected people affected by TB reached 98,040 people (East Java Provincial Health Office, 2023).

Structural Equation Modeling (SEM) and its applications have advanced significantly in recent years. SEM is particularly valuable for investigating factors related to tuberculosis (TB), as it can simultaneously assess relationships between indicators and latent variables, as well as between latent variables themselves. While SEM typically requires assumptions such as multivariate normality and large sample sizes (Shela et al., 2023). Alternative methods like partial least squares (PLS) enable analysis with smaller samples. In studies applying SEM-PLS, researchers often assume sample homogeneity (Rožman et al., 2020). However, if the population is inherently heterogeneous and this assumption is overlooked, model estimates may lose accuracy. To account for unobserved heterogeneity, various techniques can be employed, including REBUS PLS, FIMIX PLS, PLS MFC, and residual-based segmentation using hierarchical clustering, among others (Otok et al., 2024).

Residual-Based Segmentation with Hierarchical Clustering is an approach similar to the REBUS PLS algorithm (Response-Based Unit Segmentation in PLS) but simpler, as it does not involve iterative processes. This method involves analyzing residuals from the initial PLS model to detect heterogeneity in the data, followed by clustering and then re-estimating each cluster using PLS. The clustering stage groups observations based on the similarity of their residual patterns. Residual-based segmentation in SEM-PLS offers a simple and flexible alternative for addressing unobserved heterogeneity in data. Based on the above discussion, this study employs a combination of Structural Equation Modeling (SEM) with Partial Least Squares (PLS) and Hierarchical Clustering to develop a structural model applied to tuberculosis (TB) cases in the districts/cities of East Java Province. This research aims to examine and analyze the influence of various factors including social determinants, community behavior, environmental health, and access to healthcare services and facilities on tuberculosis (TB) prevalence in East Java. Subsequently, the clusters formed are interpreted to understand the distinct profiles of each region. Furthermore, it seeks to develop a regional segmentation model for TB case prevalence in East Java Province by integrating Partial Least Squares Structural Equation Modeling (SEM-PLS) and Hierarchical Clustering approaches.

2. THEORETICAL FOUNDATION AND HYPOTHESES DEVELOPMENT

2.1. Structural Equation Modelling Partial Least Square (SEM PLS)

Structural Equation Modeling (SEM) comprises two primary approaches: covariance-based SEM (CB-SEM) and variance-based SEM (Ringle et al., 2020). CB-SEM relies on large sample sizes and requires the data to meet multivariate normality assumptions. As an alternative solution to CB-SEM, variance-based SEM, also known as Partial Least Squares (PLS), has been developed. Unlike CB-SEM, PLS-SEM does not necessitate large sample sizes. A commonly applied minimum sample size rule is the "10-times rule", which requires the sample to be at least 10 times the number of structural paths in the SEM model (Pereira et al., 2024). Although still widely used, some studies caution that it may be overly simplistic or insufficient in complex models (Demir & Usak, 2025). Previous research [1] demonstrates that samples remain valid for partial least squares structural equation modeling (PLS-SEM) when the measurement model achieves

satisfactory quality standards, specifically with factor loadings exceeding 0.7 (Hair et al., 2017). Moreover, arguments favoring PLS-SEM solely due to small sample capability should be treated cautiously, as inadequate sample sizes can adversely affect accuracy and statistical power (Sarstedt et al., 2022).

In general, PLS-SEM consists of two submodels: the measurement model and the structural model. The structural model represents the relationship between independent latent variables (exogenous) and dependent latent variables (endogenous), which can be expressed by the following equation (Schumacker & Lomax, 2004).

$$\boldsymbol{\eta}_{i\times 1} = \boldsymbol{B}_{i\times i}\boldsymbol{\eta}_{i\times 1} + \boldsymbol{\Gamma}_{i\times j}\,\boldsymbol{\xi}_{j\times 1} + \boldsymbol{\zeta}_{i\times 1} \quad (1)$$

The measurement model specifies the relationships between latent variables and their corresponding indicators, represented by loading factors (λ). These loading factors quantify the correlation strength between indicators and their associated latent variables. Higher loading values indicate stronger indicator-latent variable relationships. Measurement models are categorized into two types: reflective and formative models. The reflective measurement models for both endogenous and exogenous latent variables can be expressed by the following equations:

$$\boldsymbol{y}_{(g\times 1)} = \boldsymbol{\Lambda}_{\mathrm{y}_{(\mathrm{g}\times \mathrm{i})}}\boldsymbol{\eta}_{(i\times 1)} + \boldsymbol{\varepsilon}_{(g\times 1)} \quad (2)$$

$$\boldsymbol{x}_{(h\times 1)} = \boldsymbol{\Lambda}_{x_{(\mathrm{h}\times \mathrm{j})}}\boldsymbol{\xi}_{(j\times 1)} + \boldsymbol{\delta}_{(h\times 1)} \quad (3)$$

where Λ_y represents the loading factor values for endogenous latent variables and Λ_x denotes the loading factor values for exogenous latent variables.

2.2. PLS SEM Model Estimation

Parameter estimation in PLS-SEM is obtained through three iterative stages: weight estimation, path estimation, and mean/location parameter estimation. The first stage, weight estimation, calculates latent variable factor scores. For reflective models, the weight estimation for indicators of exogenous latent variables is obtained by minimizing the sum of squared errors of δ_jh as follows.

$$\lambda_{jh} = \frac{cov(x_{jh},\xi_j)}{var\left(\xi_j^2\right)} = cor(x_{jh},\xi_j \quad (4)$$

Using an analogous derivation approach, the weight estimation for endogenous latent variable indicators is obtained as follows.

$$\lambda_{ig} = cor(y_{ig},\eta_i \quad (5)$$

The structural equation model connects latent variables through path coefficients. There are two types of path coefficients: beta coefficients (β) and gamma coefficients (γ). Beta coefficients (β) serve as links between endogenous latent variables (η), while gamma coefficients (γ) connect exogenous latent variables (ξ) with endogenous latent variables (η). The estimation of structural model parameters is determined based on three schemes: the path scheme, centroid scheme, and factor scheme.

2.3. Bootstrap and Hypothesis Testing

The bootstrap method serves as a robust statistical technique designed to mitigate uncertainty associated with distributional assumptions, particularly when normality assumptions may not hold. In the context of PLS analysis, this resampling approach involves repeatedly drawing bootstrap samples to calculate standard errors, thereby enabling researchers to assess both the significance and stability of parameter estimates across both measurement and structural models (Sarstedt et al., Partial least squares structural equation modeling. In Handbook of Market Research, 2022). By generating multiple subsamples from the original dataset through random sampling with replacement, the bootstrap technique facilitates the computation of reliable standard error estimates without relying on strict distributional assumptions. This approach proves particularly valuable in PLS-SEM applications where traditional parametric tests may be inappropriate, as it provides an empirical approximation of the sampling distribution that can be used to construct confidence intervals and evaluate the precision of path coefficients, loadings, and other model parameters. The implementation of standard error bootstrap resampling enhances the rigor of PLS-SEM results by offering a non-parametric alternative for statistical inference, ultimately strengthening the validity of conclusions drawn from the structural equation modeling analysis (Chin, 1998).

The hypothesis testing framework in Partial Least Squares Path Modeling examines two distinct sets of hypotheses: those pertaining to the Outer Model (λ) and those concerning the Inner Model (β and γ). The Outer Model hypotheses specifically evaluate the following relationships.

$H_0: \lambda_h = 0$ (The Indicator does not significantly contribute to the measurement of its latent construct)

$H_1: \lambda_h \neq 0$ (The Indicator demonstrates a significant measurement contribution to its latent construct)Test statistic,

$$T = \frac{\hat{\lambda}_h}{se(\hat{\lambda}_h)} \quad (6)$$

The null hypothesis is rejected when $|T| > T_{\frac{\alpha}{2},df}$, indicating a statistically significant measurement contribution at the specified α-level (typically 0.05) with appropriate degrees of freedom.

For the inner model (β and γ) the hypothesis is as follows:

Beta Coefficients (β)*:*

$H_0: \beta_i = 0$ (The i-th endogenous latent variable has no significant effect on another endogenous latent variable)

$H_1: \beta_i \neq 0$ (The i-th endogenous latent variable has a significant effect on another endogenous latent variable)

Gamma Coefficients (γ)

$H_0: \gamma_j = 0$ (The j-th exogenous latent variable has no significant effect on the endogenous latent variable)

$H_1: \gamma_j \neq 0$ (The j-th exogenous latent variable has a significant effect on the endogenous latent variable)

Test statistic,

$$T = \frac{\hat{\eta}_i}{se(\hat{\eta}_i)} \text{ or } T = \frac{\hat{\xi}_j}{se(\hat{\xi}_j)} \quad (7)$$

The null hypothesis is rejected when $|T| > T_{\frac{\alpha}{2},df}$.

2.4. Model Evaluation

The assessment of measurement models in Partial Least Squares Structural Equation Modeling (PLS-SEM) depends on whether the constructs are specified as reflective or formative. Each type requires distinct evaluation criteria to ensure reliability and validity. For reflective constructs, the indicators are assumed to be manifestations of the underlying latent variable (Amora, 2023). The following criteria are used for evaluation:

2.4.1. Convergent Validity

Convergent validity assesses the degree to which a latent variable correlates with its indicators in a reflective measurement model. In evaluating convergent validity, an indicator is considered valid if it has a loading factor > 0.7 for confirmatory research or > 0.5 for exploratory or early-stage research (Hair et al., 2017).

2.4.2. Discriminant Validity

Discriminant validity in reflective measurement models is evaluated through indicator cross-loadings on their respective latent variables. Additionally, this validity can be assessed using the Average Variance Extracted (AVE) metric. The condition is satisfied when the square root of AVE exceeds the correlation coefficients between latent variables.

Composite Reliability

A latent variable demonstrates adequate reliability when its composite reliability exceeds 0.7. The composite reliability value is computed using the following equation.

$$\rho_c = \frac{\left(\sum_{i=1}^{I} \lambda_i^2\right)^2}{\left(\sum_{i=1}^{I} \lambda_i^2\right)^2 + \sum_{i=1}^{I} (1-\lambda_i^2)} \quad (8)$$

As for the structural model or inner model, some of the evaluation measures used are as follows.

R-Square (R^2)

The coefficient of determination (R^2) serves as a key metric for evaluating model predictive power in PLS-SEM. If the R^2 is 0.67, the model can be considered or shown to be in a good category. If the R^2 is 0.33, the model is categorized as moderate. Meanwhile, if the R^2 is 0.19, the model is considered to be in the weak category.

Q-Square Predictive Relevance (Q^2)

This test is used to validate the predictive ability of the model. Interpretation of the Q^2 results shows that if the value is greater than 0, then the exogenous latent variable is considered good (relevant) as an explanatory variable that is able to predict the latent variable. The following is the equation for calculating the value of Q^2.

$$Q^2 = 1 - \frac{\sum_{i=1}^{I} \left(Y_{i,n} - \hat{Y}_{i,n}\right)^2}{\sum_{i=1}^{I} \left(Y_{i,n} - \bar{Y}_{(i)}\right)^2} \quad (9)$$

2.5. Ward Method Clustering

Cluster analysis is divided into two main types: hierarchical clustering and non-hierarchical clustering. Hierarchical clustering forms a hierarchy or cluster levels in the form of a dendrogram, where objects are grouped in stages based on their similarities or distances between them. Among the hierarchical clustering methods is the Ward method. Ward's method in hierarchical clustering aims to minimize the information loss that occurs when two groups are merged. This approach is based on the principle of minimizing the increase in the Sum of Squared Errors (SSE), so that the two groups combined are those that produce the smallest increase in SSE. Thus, Ward's method produces more homogeneous clusters (Coombes et al., 2021).

2.6. Hypotheses Development

The Commission on Social Determinants of Health (CSDH) framework developed by WHO, as

shown in Figure 2.1, provides in-depth insight into how social determinants influence health inequalities. This framework groups determinants into structural determinants (social, economic, and political factors that shape an individual's social position) and intermediary determinants (material conditions, behaviors, and psychosocial factors that contribute to health risks). In the context of TB, social determinants such as poverty, low education, limited access to health services, and poor environmental conditions are significant risk factors. Through this approach, it can be understood that TB control depends not only on medical interventions but also on strategies to address the underlying social determinants.

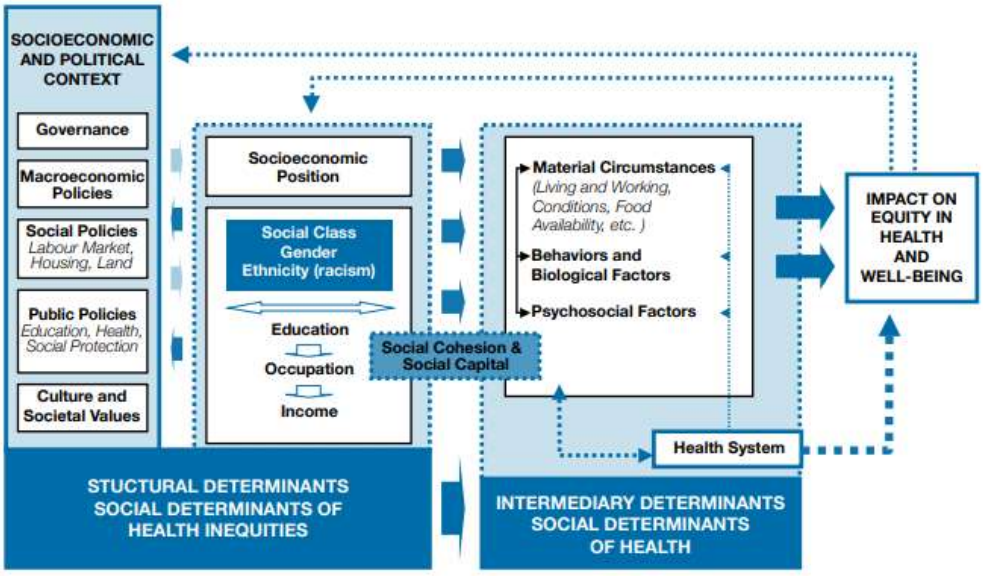

Figure 1: CSDH Conceptual Framework.

The following is an explanation of the risk factors for tuberculosis.

2.7. Social Determinants

Social inequities, including low education, low income, poverty, and precarious employment, play a significant role in increasing the risk of TB. Studies in the Philippines, Vietnam, Bangladesh, and Kenya have shown that individuals with low socioeconomic status have a TB prevalence 1.7–2 times higher than those with higher socioeconomic status (Van Leth et al., 2011). Then there's the educational aspect, which is related to a person's level of knowledge about health, their choices about maintaining their health, and their ability to manage their life. Furthermore, education is also closely related to an individual's income level and well-being. Higher educational attainment increases opportunities for higher earnings and income, which in turn are associated with healthier working conditions. (Braveman et al., 2011).

2.8. Access to Health Services

Access to health services, including the number of medical personnel and the number of health facilities, is a critical factor in the early diagnosis and treatment of TB. Studies show that better access to health services lowers the risk of TB, as this allows for early detection and timely treatment (Gu et al., 2009). Furthermore, the number of medical personnel and the availability of healthcare facilities directly impact the quality of services received by the public. Research in Bandar Lampung found that ease of access, as measured by distance and transportation availability to healthcare facilities, was significantly associated with TB incidence. The analysis showed that the greater the distance to healthcare facilities or the more difficult the transportation options, the greater the likelihood of TB cases occurring (Barker et al., 2002).

2.9. Community Behavior

Health behavior is a person's actions related to health, illness, the health care system, diet, and the environment. This behavior is influenced by an individual's knowledge and attitudes, as well as accessibility to available health facilities (Green, 1980). Among the health behaviors that are risk factors for TB is smoking. In their study, Slama et al. (2007) found that patients who stated they quit smoking made a positive contribution to TB recovery. A review of several previous studies also showed that the incidence of TB was higher in people who smoked (Wardani, 2014). Besides smoking, HIV

is also a risk factor for TB. HIV suppresses the immune system, allowing latent TB to progress to active TB. Furthermore, regarding diabetes mellitus, a review of WHO surveys in 46 countries showed that information related to diabetes mellitus (DM) in developing countries is still very limited. Meanwhile, in developed countries, DM is more common in individuals with low socioeconomic status. Furthermore, it was stated that in developed countries, DM increases the risk of TB with an Odds Ratio (OR) of 2.39; 95% CI 1.843-10 (Goldhaber-Fiebert, 2011).

2.10. Environmental Health

Several studies have highlighted the relationship between environmental health factors, such as the cleanliness of public facilities and food preparation facilities (TPP) that meet standards, and the incidence of tuberculosis (TB). Research by Amirus et al. (2022) shows that poor sanitation conditions, including substandard drinking water quality, can increase the risk of environmentally-related diseases, such as TB. Furthermore, hygiene and sanitation in food processing and preparation play a crucial role in preventing infectious diseases. Research by Fitri (2017) shows that the number of public places that meet health standards influences the incidence of TB. Although direct research linking TPP sanitation to TB is still limited, poor hygiene practices can contribute to a decline in an individual's health status, thereby increasing susceptibility to infections, including TB.

3. METHODS

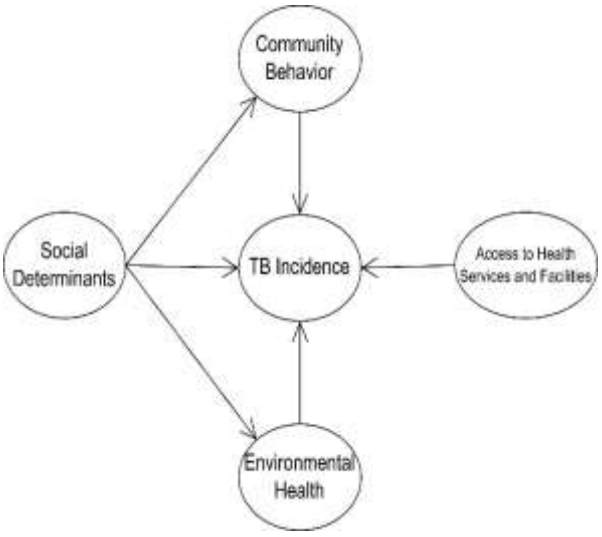

Figure 2: Research Conceptual Framework.

This study uses secondary data obtained from the Central Statistics Agency (BPS) of East Java Province and the East Java Provincial Health Office in 2023. The observations used are at the district/city level in East Java Province, which consists of 38 districts/cities. The research conceptual framework is presented in the following figure:

Based on the conceptual framework in Figure 1, tuberculosis incidence is caused by direct and indirect causal factors. Community behavior, environmental health, and access to health services and facilities are the direct causes, that is, factors thought to directly influence the number of tuberculosis cases, while social determinants are the indirect causes.

Table 1: Research Variables.

Latent Variables		Indicator
Community Behavior (η_1)	Y_{11}	Percentage of Population Aged >15 Years Who Smoke
	Y_{12}	Number of HIV Cases
Environmental Health (η_2)	Y_{21}	Percentage of Public Places and Facilities Monitored According to Standards
	Y_{22}	Percentage of Food Service Establishments Meeting Health Requirements
TB Incidence (η_3)	Y_{31}	Tuberculosis Prevalence
Social Determinants (ξ_1)	X_{11}	Percentage of Population Completing Higher Education
	X_{12}	Per Capita Income
	X_{13}	Percentage of Non-Poor Population
	X_{14}	Percentage of Working-Age Population
Access to Health Services and Facilities (ξ_2)	X_{21}	Number of Medical Personnel
	X_{22}	Number of Health Facilities

Based on Table 1, the variables used in the study consist of 5 latent variables, including 2 exogenous latent variables and 3 endogenous latent variables, comprising 11 indicator variables.

This study's analytical approach is bifurcated into two sequential stages. The initial stage employs Structural Equation Modeling with Partial Least Squares (SEM-PLS) to address the first research objective, resulting in a global model. The subsequent stage involves a clustering analysis founded upon the residuals generated from the first model, which is conducted to fulfill the second research objective. The model derived from this second stage is designated as the local model.

The first-stage analysis steps using the SEM-PLS method are as follows:

1. Constructing a measurement model (outer model) to describe the relationships between latent variables and their indicators.
2. Developing a structural model (inner model) to illustrate the causal relationships among latent variables.
3. Constructing a path diagram to visually represent the structural and measurement models.
4. Converting the path diagram into a system of equations for statistical analysis.
5. Calculate the estimated weights for latent variable factor scores
6. Estimating loading factor (λ), and path coefficients (β and γ).
7. Conduct hypothesis testing (bootstrap resampling).
8. Perform model evaluation (assessing validity and reliability for the measurement model, and R^2 and Q^2 for the structural model).

After obtaining the global model from the first stage analysis steps, residual extraction is carried out from the global model to then continue to the second stage analysis step. The second stage analysis steps are as follows:

1. Calculating the distance matrix
2. Performing hierarchical clustering with the Ward method
3. Label observations into clusters
4. Grouping data based on the segments or clusters formed
5. Perform SEM-PLS modeling for each cluster formed (local model)
6. Compare the results between the local and global models based on parameter significance criteria and the R-squared (R^2) criteria to determine the best model.
7. Interpret the best local model.
8. Draw conclusions.

4. RESULT AND DISCUSSION

4.1. Descriptive Analysis

As a preliminary step, data exploration was conducted, covering the indicator variables used in this study. The results of this exploration are expected to provide a general overview of the research data.

Table 2: Descriptive Statistics.

Indicator		Mean	Min	Max	St. Dev
Percentage of Population Completing Higher Education	X_{11}	9,90	2,44	23,06	5,42
Per Capita Income	X_{12}	12.287,05	9.363	18.977	2.233,28
Percentage of Non-Poor Population	X_{13}	89,71	78,24	96,69	4,26
Percentage of Working-Age Population	X_{14}	69,37	66,31	72,03	1,02
Number of Medical Personnel	X_{21}	2.658,45	796	18.382	2.876,19
Number of Health Facilities	X_{22}	88,05	22	528	85,60
Percentage of Population Aged >15 Years Who Smoke	Y_{11}	28,47	22,52	34,56	3,01
Number of HIV Cases	Y_{13}	280,82	32,00	1.260,00	249,51
Percentage of Public Places and Facilities Monitored According to Standards	Y_{21}	87,26	61,01	99,10	9,51
Percentage of Food Service Establishments Meeting Health Requirements	Y_{22}	71,07	37,11	99,44	12,95
Tuberculosis Prevalence	Y_{31}	264,71	80,00	800,00	159,69

The prevalence of TB (Y_{31}) varies considerably, from 80 to 800 per 100,000 population, with a standard deviation of 159.62, confirming that the disease burden is uneven and influenced by various factors.

Figure 3: Distribution of tuberculosis prevalence in East Java.

Figure 2 shows that the prevalence of tuberculosis tends to be higher in urban areas than in districts. Further analysis shows that there is inequality in access to health services as reflected in the

distribution of medical personnel (X_{21}) (796-18,382) and health facilities (X_{22}) (22-528 facilities) which are very uneven between districts/cities. Comorbid factors such as the number of HIV cases (Y_{12}) which also vary widely contribute to this TB distribution pattern. On the social determinant side, indicators such as education level (X_{11}), on average only 9.9% of the population has a higher education and per capita expenditure (X_{12}) of IDR 9.3-18.9 million, show socio-economic disparities which may be the root of this health disparity problem.

4.2. SEM PLS Analysis (Global Model)

The initial phase of the PLS-SEM analysis involves constructing a path diagram through the design of both a measurement model and a structural model, each aligned with the predefined conceptual framework. In this study, the structural model consists of two exogenous latent variables, namely social determinants (ξ_1) and access to health services and facilities (ξ_2), and three endogenous latent variables, namely community behavior(η_1), environmental health (η_2), and TB incidence (η_3). This study employs a reflective measurement model, wherein the latent variables influence the manifest variables. Additionally, a path diagram illustrates the relationships between the latent variables and their corresponding indicators, as well as the paths connecting the exogenous and endogenous latent variables.The path diagram can be seen in Figure 3 below.

In modeling using SEM PLS, it is necessary to identify valid and reliable indicator variables for the latent variables. Table 3 shows the results of the measurement model evaluation based on several evaluation values, including loading factor, AVE, and composite reliability.

As presented in Table 3, all indicators exhibit a loading factor greater than 0.5, confirming their validity. Furthermore, the Average Variance Extracted (AVE) values exceeding 0.5 demonstrate the fulfillment of convergent validity criteria.

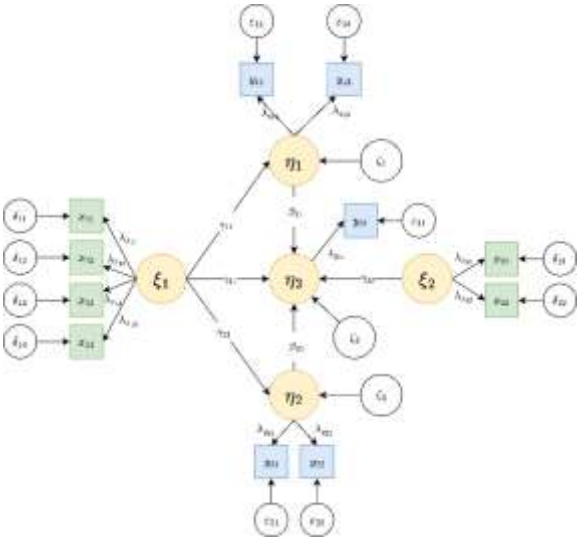

Figure 4: Path Diagram.

Table 1: Indicator Validity and Reliability Test on Latent Variables.

Latent Variables	Indicator	Loading Factor	AVE	Composite Reliability	P-Value
Social Determinants (ξ_1)	X_{11}	0,927	0,759	0,926	0,000*
	X_{12}	0,930			0,000*
	X_{13}	0,875			0,000*
	X_{14}	0,740			0,000*
Access to Health Services and Facilities (ξ_2)	X_{21}	0,996	0,987	0,994	0,000*
	X_{22}	0,991			0,000*
Community Behavior (η_1)	Y_{11}	0,970	0,609	0,741	0,000*
	Y_{12}	0,528			0,016*
Environmental Health (η_2)	Y_{21}	0,790	0,693	0,818	0,000*
	Y_{22}	0,873			0,000*
TB Incidence (η_3)	Y_{31}	1,000	1,000	1,000	

Meanwhile, all composite reliability values surpass 0.7, indicating that the reliability requirements are satisfied. Hypothesis testing for the measurement model was conducted using the bootstrap resampling method, with results confirming the statistical significance of all indicators with a significance level of 5%.

Table 4: Hypothesis Testing Result for Structural Models.

Paths	Path Coefficient	Bootstrap Stdev	T-stat	P-Value
$\xi_1 -> \eta_1$	-0,732	0,062	11,838	0,000*
$\xi_1 -> \eta_2$	0,651	0,101	6,455	0,000*
$\xi_1 -> \eta_3$	0,557	0,201	2,775	0,006*
$\xi_2 -> \eta_3$	-0,179	0,124	1,436	0,152
$\eta_1 -> \eta_3$	-0,377	0,189	1,991	0,047*
$\eta_2 -> \eta_3$	-0,205	0,159	1,291	0,197

Based on Table 4, it is evident that social determinants have a significant and negative influence on community behavior, indicating that better education and economic aspects in a region will reduce negative community behaviors (smoking and HIV). Social determinants also have a significant

and positive effect on environmental health, meaning that social health will improve as social determinants increase. Additionally, social determinants show a significant and positive influence on TB incidence, suggesting that improved social determinants actually increase TB cases. However, it is important to note that urban areas, which generally have better education and economic conditions, tend to have higher tuberculosis prevalence. The latent variable of community behavior has a significant negative effect on TB incidence, which contradicts theoretical expectations where lower smoking rates and fewer HIV cases should typically reduce TB occurrence. Based on Table 4, the structural equation model can be formulated as follows:

$$\eta_1 = -0{,}732\xi_1 + \zeta_1$$

$$\eta_2 = 0{,}651\xi_1 + \zeta_2$$

$$\eta_3 = -0{,}377\eta_1 - 0{,}205\eta_2 + 0{,}557\xi_1 - 0{,}179\xi_2 + \zeta_3$$

Next, the structural model evaluation is conducted by analyzing the R^2 and Q^2 values. The Community Behavior variable (η_1) has an R^2 value of 0.536, indicating that social determinants account for 53.6% of its variance. For the Environmental Health variable (η_2), the R^2 value of 0.424 falls into the moderate category, while the TB Incidence variable (η_3) achieves an R^2 of 0.486, demonstrating that the combination of factors in the model contributes significantly. Further assessment of the model's predictive capability is performed using the Q^2 (predictive relevance) metric. The Q^2 values are calculated via the blindfolding procedure (D=5) based on Equation (9), yielding Q^2 values of 0.257, 0.268, and 0.401. Since all values exceed 0, the exogenous latent variables in the model meet the predictive relevance criterion and can be considered valid predictors for the endogenous latent variables.

4.3. Segmentation Analysis with Ward Method Clustering

The initial SEM-PLS modeling yielded global model residuals, which were subsequently analyzed using cluster analysis to identify homogeneous segments for further modeling. Ward's hierarchical clustering method was employed, producing two distinct clusters that differentiated between high and low TB prevalence regions.

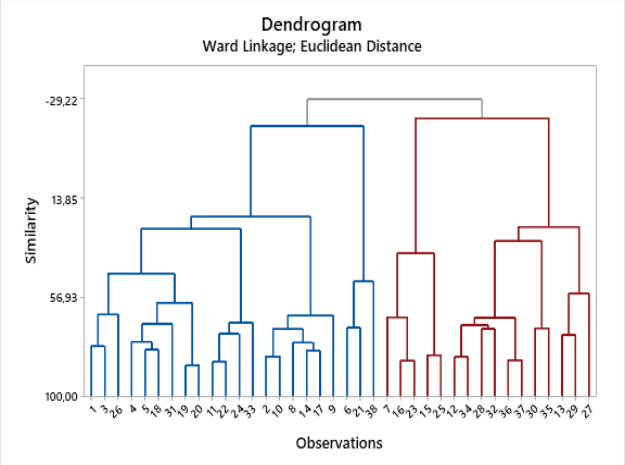

Figure 5: Dendogram Ward Method (2 Clusters).

Figure 6: Distribution of Cluster Ward.

This grouping resulted in segment 1, consisting of 22 members, representing 18 regencies and 4 cities, while segment 2 consisted of 16 members, representing 11 regencies and 5 cities. The distribution of each segment is shown in Figure 5 below. After clustering has been performed and two clusters have been obtained, each cluster is then modeled using SEM-PLS. The modeling for the first segment/cluster is called local model 1 (L1), while the modeling for the second segment is called local model 2 (L2). Local model 1 and local model 2 are then compared with the initial model in the first stage (global model).

Table 5: Comparison of Loading Factors L1, L2 and GM.

Indicator		GM	L1	L2
Percentage of Population Completing Higher Education	X_{11}	0,927	0,916	0,947
Per Capita Income	X_{12}	0,930	0,901	0,945

Percentage of Non-Poor Population	X_{13}	0,875	0,778	0,951
Percentage of Working-Age Population	X_{14}	0,740	0,693	0,750
Number of Medical Personnel	X_{21}	0,996	0,990	0,944
Number of Health Facilities	X_{22}	0,991	0,978	0,989
Percentage of Population Aged >15 Years Who Smoke	Y_{11}	0,970	0,759	0,939
Number of HIV Cases	Y_{13}	0,528	-0,553	0,832
Percentage of Public Places and Facilities Monitored According to Standards	Y_{21}	0,790	0,790	0,913
Percentage of Food Service Establishments Meeting Health Requirements	Y_{22}	0,873	0,935	0,777
Tuberculosis Prevalence	Y_{31}	1,000	1,000	1,000

Table 5 shows that each indicator in both the global and local models has a loading factor value > 0.5, thus all indicators are valid. However, in L1, one indicator has a negative loading factor. Next, the path coefficient and its significance between the local and global models are compared as a measure of the significance of the influence between the latent variables.

Table 2: Comparison Of Path Coefficients L1, L2 And GM.

Paths	GM (38)		L1 (22)		L2 (16)	
	Path Coef	P-Value	Path Coef	P-Value	Path Coef	P-Value
$\xi_1 -> \eta_1$	-0,73	0,00*	-0,57	0,34	-0,84	0,00*
$\xi_1 -> \eta_2$	0,65	0,00*	0,62	0,00*	0,88	0,00*
$\xi_1 -> \eta_3$	0,56	0,01*	0,48	0,18	-0,04	0,93
$\xi_2 -> \eta_3$	-0,18	0,15	0,14	0,53	-0,36	0,03*
$\eta_1 -> \eta_3$	-0,38	0,05*	0,42	0,38	-0,57	0,06
$\eta_2 -> \eta_3$	-0,20	0,20	0,27	0,40	0,43	0,18

As presented in Table 6, Local Model 1 (L1) exhibits only one significant path coefficient: the positive relationship between social determinants (ξ_1) and environmental health (η_2). This implies that improvements in a region's educational and economic conditions are associated with enhanced environmental health. However, no other variables in L1 significantly influence TB incidence. In contrast, Local Model 2 (L2) demonstrates three significant path coefficients: (1) the effect of social determinants on community behavior, (2) social determinants on environmental health, and (3) healthcare access on TB incidence. All three relationships align with theoretical expectations. A comparison of path coefficient significance between the global model (heterogeneous data) and local models (segmented data) reveals that SEM-PLS modeling on homogeneous data yields more robust and interpretable results than modeling on heterogeneous data.

Table 7: Comparison of R^2 values for L1, L2 and GM.

Latent Variable	GM	L1	L2
Community Behavior (η_1)	0,536	0,325	0,705
Environmental Health (η_2)	0,424	0,383	0,770
TB Incidence (η_3)	0,486	0,355	0,730

Table 7 analysis reveals that Local Model 2 demonstrates superior performance with a substantially higher R^2 value (>0.7) compared to the Global Model, while Local Model 1 shows a relatively low R^2 (approximately 0.3). These results suggest that segmentation modeling using Ward's method yields better performance in Local Model 2. However, it should be noted that this superiority is not absolute when compared to other models. This finding reinforces that data segmentation can enhance modeling quality, particularly for specific data clusters.

4.4. Implications

The findings of this study carry significant implications across theoretical, methodological, and practical domains for tuberculosis control and public health research.

4.4.1. Theoretical Implications

This study strengthens and sharpens the

theoretical framework of the Commission on Social Determinants of Health (CSDH) by illustrating its relevance in relation to tuberculosis in East Java. The worldwide model validated the important influence of social factors and community actions on the incidence of TB. Significantly, the segmentation analysis indicated that these relationships vary across different regions. The local models, especially Local Model 2, offer a more detailed insight by revealing that in specific segments (noted for their higher R^2 values), the direct impact of access to health services and facilities on TB incidence becomes crucial, while the direct influence of social determinants on TB incidence lessens. This indicates that the pathways and levels of influence outlined by the CSDH framework can differ greatly depending on regional traits, supporting a more context-driven use of the theory.

4.4.2. Methodological Implications

From a methodological perspective, this research highlights the vital significance of tackling unobserved heterogeneity in public health data analysis employing SEM-PLS. The comparison of the global model with the local models clearly shows that residual-based segmentation through Ward's clustering can reveal hidden data structures that a singular global model hides. Local Model 2, originating from a more uniform segment, demonstrated enhanced explanatory capability (greater R^2) and more theoretically consistent path coefficients in relation to the global model. This confirms the effectiveness of integrating SEM-PLS with hierarchical clustering as a strong analytical method for creating region-specific health intervention models for diverse populations.

4.4.3. Practical Implications

For health policymakers in East Java, these findings advocate for a segmented TB control strategy. In high-performing cluster areas (Local Model 2), resources should be strategically concentrated on directly improving healthcare access, such as equitable distribution of medical staff and facilities. For the other cluster (Local Model 1), a broader, multi-sectoral approach is necessary, simultaneously addressing social determinants, community behavior, and environmental health. This ensures efficient resource allocation and maximizes intervention impact by targeting the most influential factors in each specific regional context.

5. CONCLUSION

The PLS-SEM analysis confirmed that all measurement indicators adequately represented their corresponding latent variables. Structural model analysis demonstrated that tuberculosis incidence was significantly influenced by both social determinants and community behavior. Additionally, social determinants were found to directly affect community behavior and environmental health. Cluster analysis of model residuals identified two distinct subgroups within the data. Comparative evaluation revealed that Local Model 2 showed superior performance to the global model, with stronger parameter estimates and better explanatory power. In this local model, tuberculosis incidence was primarily associated with healthcare access and facilities, while social determinants maintained their influence on community behavior and environmental health factors. These findings suggest that population subgroups may require different intervention approaches for tuberculosis prevention and control.

REFERENCES

Amora, J. (2023). On the validity assessment of formative measurement models in PLS-SEM. Data Analysis Perspectives Journal, 4(2), 1-7.

Barker, R. D., Nthangeni, M. E., & Millard, F. C. (2002). Is The Distance A Patient Lives from Hospital A Risk Factor for Death from Tuberculosis in Rural South Africa? The International Journal of Tuberculosis and Lung Disease, 98-103.

Braveman , P. A., Egerter, S. A., & Mockenhaupt, R. E. (2011). Broadening the Focus The Need to Address the Social Determinants of Health. American Journal of Preventive Medicine, S4-S18.

Chin, W. W. (1998). The Partial Least Squares Approach for Structural Equation Modeling. Modern Method for Business Resarch, 295-362.

Coombes, C., Liu, X., Abrams, Z., Coombes, K., & Brock, G. (2021). Simulation-derived best practices for clustering clinical data. J Biomed Inform.

Demir, S., & Usak, M. (2025). Analyzing the Implementation of PLS-SEM in Educational Technology Research: A Review of the Past 10 Years. Sage Open, 15(2).

East Java Provincial Health Office. (2023). East Java Province Health Profile 2023. Surabaya: DINKES JATIM.

Gu, D., Zhang, Z., & Zeng, Y. (2009). Access to Healthcare Services Makes A Difference in Healthy Longevity

among Older Chinese Adults. Social Science & Medicine, 210-219.
Hair, J. F., Hult, G. M., Ringle, C. M., & Sarstedt, M. (2017). A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM) (2 ed.). Los Angeles: SAGE.
Kemenkes RI. (2023). 2022 Tuberculosis Control Program Report. Jakarta: Kementrian Kesehatan RI.
Otok, B. W., Purhadi, Sriningsih, R., & Dila, D. S. (2024). Segmentation of toddler nutritional status using REBUS and FIMIX partial least square in Southeast Sulawesi. MethodsX, 12.
Pereira, L., Rodriguez, V., & Freires, F. (2024). Use of Partial Least Squares Structural Equation Modeling (PLS-SEM) to Improve Plastic Waste Management. Applied Sciences, 14(2), 628.
Ringle, C., Sarstedt, M., Mitchell, R., & Gudergan, S. (2020). Partial least squares structural equation modeling in HRM research. The International Journal of Human Resource Management, 31(12), 1617–1643.
Rožman, M., Tominc, P., & Milfelner, B. (2020). A Comparative Study Using Two SEM Techniques on Different Samples Sizes for Determining Factors of Older Employee's Motivation and Satisfaction. Sustainability, 12, 2189.
Sarstedt, M., Hair, J., Liengaard, B., Radomir, L., & Ringle, C. (2022). Progress in partial least squares structural equation modeling use in marketing research in the last decade. Psychology & Marketing.
Sarstedt, M., Radomir, L., Moisescu, O., & Ringle, C. (2022). Latent class analysis in PLS-SEM: A review and recommendations for future applications. Journal of Business Research, 138(C), 398-407.
Sarstedt, M., Ringle, C., & Hair, J. (2022). Partial least squares structural equation modeling. In Handbook of Market Research. Springer.
Schumacker, R., & Lomax, R. (2004). A Beginner's Guide toStructural Equation Modeling (2nd ed.). New Jersey: Lawrence Erlbaum Associates.
Shela, V., Ramayah, T., Aravindan, K., Ahmad, N., & Alzahrani, A. (2023). A systematic review of PLS-SEM application in strategic management research among developing nations. Heliyon, 9(12).
Van Leth, F., Guilatco, R. S., Hossain, S., Van't Hoog, A. H., Hoa, N. B., Van der Werf, M. J., & Lönnroth, K. (2011). Measuring socio-economic data in tuberculosis prevalence surveys. The International Journal of Tuberculosis and Lung Disease, 558-563.
WHO. (2015). The END TB Strategy. Geneva: WHO.
World Health Organization. (2023). Global Tuberculosis Report. Geneva: World Health Organization.