

ANALYZING AND CLUSTERING THAILAND'S MULTIDIMENSIONAL POVERTY DATA USING AN IMPROVED K-MEANS ALGORITHM

Narintra Mingolo and Weerachai Sarakorn*

Received: October 06, 2025; Revised: November 11, 2025; Accepted: December 16, 2025

Abstract

This study proposes an Improved K-Means (IKM) algorithm to enhance the clustering of Thailand's multidimensional poverty data. The framework integrates Principal Component Analysis (PCA) for dimensionality reduction, the Elbow Method for optimal cluster selection, and Density Peak Clustering (DPC) for centroid initialization within an automated workflow. Provincial-level data from the National Statistical Office (NSO) for 2019, covering all 76 provinces, were analyzed. The performance of IKM was compared with standard K-Means and Gaussian Mixture Models (GMM) using the Silhouette coefficient as the internal validation metric. IKM achieved the highest Silhouette scores in six of seven indicators, indicating superior cluster compactness and separation. To verify robustness, the framework was further evaluated using district-level data from the Thai People Map and Analytics Platform (TPMAP). Paired bootstrap and paired permutation tests confirmed that the observed improvements were statistically significant rather than attributable to random variation. For example, comparison with Hierarchical Clustering using 2,000 bootstrap resamples produced a 95% confidence interval excluding zero, while both statistical tests yielded p-values below 0.05. Comparable results were obtained for comparisons with K-Means and GMM, confirming the consistency and statistical reliability of IKM. Overall, the proposed framework delivers more accurate, stable, and interpretable clustering, effectively capturing multidimensional socioeconomic disparities across Thailand. These findings provide a rigorous analytical basis for evidence-based, spatially targeted poverty reduction policies and offer a scalable framework for future district-level and longitudinal analyses.

Keywords: Clustering; Density Peak; Improved K-Means; Multidimensional Poverty; PCA; Thailand

Department of Mathematics, Faculty of Science, Khon Kaen University, Khon Kaen 40002, Thailand. E-mail: nusharin_m@kkumail.com, ORCID: 0009-0004-5665-6292; wsarakorn@kku.ac.th, ORCID: 0000-0003-2334-8213*

**Corresponding author*

DOI: <https://doi.org/10.55766/sujst11738>

Suranaree J. Sci. Technol. 33(3):030390(1-19)

Introduction

Poverty is a complex, multidimensional, and spatially heterogeneous issue that perpetuates cycles of deprivation across individuals, households, and communities by restricting access to basic necessities, education, healthcare, and economic opportunities. The Multidimensional Poverty Index (MPI), developed by the Oxford Poverty and Human Development Initiative (OPHI) in collaboration with the United Nations Development Programme (UNDP), broadens the traditional income-based approach to include three key dimensions: health, education, and standard of living. These dimensions capture deprivations related to nutrition, child mortality, schooling, sanitation, housing, and household assets (OPHI and UNDP, 2023). Numerous studies have highlighted the multidimensional nature of poverty and its association with structural inequalities. For example, Mohanty *et al.* (2017) showed that increased investment in healthcare and broader access to health insurance can mitigate both catastrophic health expenditures and multidimensional poverty. Pinilla-Roncancio *et al.* (2020) found that people with disabilities experience higher levels and greater intensity of multidimensional poverty. Fonta *et al.* (2020) examined the determinants of child poverty in Burkina Faso. In Thailand, Wiraphanphong (2020) analyzed multidimensional poverty using the Thai People Map and Analytics Platform (TPMAP), while Alkire *et al.* (2021) simulated the short-term effects of the COVID-19 pandemic and projected a 3.6–9.9-year setback in global poverty reduction across 70 countries. These studies collectively emphasize the importance of adopting multidimensional and spatially sensitive approaches to poverty analysis.

Thailand has made significant progress in reducing its MPI score through evidence-based poverty alleviation programs and area-based development strategies. TPMAP, developed collaboratively by government agencies, allows policymakers to design interventions that meet the specific needs of individuals and households at provincial, district, and subdistrict levels. Nevertheless, analyzing and interpreting multidimensional data remains a major challenge because existing methods often fail to capture the interconnections among education, health, and economic variables. Although the National Statistical Office (NSO) collects extensive poverty-related data covering these dimensions, there has been little or no comprehensive clustering analysis of this dataset. Addressing this analytical gap can

enhance Thailand's capacity to design targeted and region-specific poverty alleviation policies.

This study aims to fill that gap by proposing an Improved K-Means (IKM) clustering framework that integrates Principal Component Analysis (PCA) for dimensionality reduction, the Elbow method for determining the optimal number of clusters, and Density Peak Clustering (DPC) for improving centroid initialization. Traditional K-Means is known to be sensitive to random initialization and less effective with high-dimensional data. The proposed hybrid IKM approach improves initialization, speeds up convergence, and provides more interpretable clusters. By applying the IKM algorithm to the 2019 NSO provincial dataset, which covers all 76 provinces, this research identifies multidimensional poverty patterns and their regional distributions across Thailand.

The choice of 2019 (B.E. 2562) data is deliberate and based on methodological considerations. The outbreak of COVID-19 in early 2020 caused widespread disruptions to field surveys and delayed statistical data collection nationwide. Several socioeconomic indicators were postponed, partially collected, or not yet resumed, leading to inconsistencies in later datasets. Therefore, the 2019 data is considered the most complete, consistent, and reliable baseline available before the pandemic shock. Using this dataset provides a robust foundation for spatial clustering and policy-oriented analysis under normal socioeconomic conditions.

Accordingly, this study seeks to answer the following research questions:

RQ1. Does the hybrid IKM approach improve clustering quality compared with traditional K-Means (KM) and Gaussian Mixture Models (GMM)?

RQ2. What spatial and regional patterns emerge from the resulting clusters, and how do they relate to provincial poverty rates (PPP)?

RQ3. What additional economic value does IKM offer compared with simple income-based provincial classification?

RQ4. How are the different dimensions of poverty interconnected within clusters, and what implications do these relationships have for area-based poverty reduction policies in Thailand?

Through these objectives, this research contributes a replicable methodological framework, an empirical analysis of Thailand's multidimensional poverty landscape, and practical insights for evidence-based policy formulation and regional development planning.

Traditional K-Means Clustering

The K-Means (KM) algorithm, introduced by MacQueen in 1967 (MacQueen, 1967) and later improved by Lloyd in 1982 (Lloyd, 1982), is a widely used technique for clustering data. Its simplicity and efficiency make it a fundamental tool for many applications related to poverty data. For instance, during the COVID-19 pandemic, it can forecast and plan growth rates, supply chain disruptions, and governmental decisions (Nikolopoulos *et al.*, 2021). It has also been used to identify poverty patterns across Indonesia's provinces (Sano and Nindito, 2016) and clusters of children's social mobility opportunities for public health applications (Zelasky *et al.*, 2023). The concept of traditional KM is summarized as Algorithm 1.

Algorithm 1 The traditional K-Means algorithm.

Require:
 1: -Set of data points $X = \{x_1, x_2, x_3, \dots, x_N\}$
 2: -Number of clusters K .
 Ensure:
 3: -Set of clusters $C = \{C_1, C_2, C_3, \dots, C_K\}$, where C_i contains the data points assigned to the i -th cluster.
 4: -Set of cluster centroids $U = \{\mu_1, \mu_2, \mu_3, \dots, \mu_K\}$.
 5: Initialization step:
 6: Create the set of initial centroids $U = \{\mu_1, \mu_2, \mu_3, \dots, \mu_K\}$ by selecting K data points randomly from the dataset.
 7: Assignment step:
 8: for $i = 1$ to N do
 9: for $j = 1$ to K do
 10: $d_{ij} = \|x_i - \mu_j\|$
 ▷ Distance between x_i and μ_j .
 11: end for
 12: end for
 13: Update step:
 14: for $j = 1$ to K do
 15: $\mu_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i$ ▷ S_j is the set of data points assigned to cluster j and $\|S_j\|$ is the number of data points in S_j .
 16: end for
 17: Convergence:
 18: Repeat steps 2 and 3 until convergence (the centroids stop changing significantly between iterations or until a maximum number of iterations is reached).

Here, we suppose that x_i is in M -dimensional space and $\|\cdot\|$ is the Euclidean norm. As depicted in algorithm 1 mentioned above, the traditional KM manifests various weaknesses that can impede its effectiveness in specific contexts. Recently, a comprehensive review of KM was presented (Ikotun *et al.*, 2023), discussing open issues and challenges faced by this algorithm and its variants. The review focused on two primary

challenges: high data dimensionality and the initialization step. The high dimensionality of the data presents a significant challenge, making it difficult for KM to manage large datasets with diverse attributes effectively.

Furthermore, the initialization problem, which involves accurately determining the number of clusters and locating initial centroids, significantly impacts the algorithm's performance. As mentioned, this study employs Principal Component Analysis (PCA) to reduce the dimensionality of the high-dimensional MPI dataset, thereby transforming it into a lower-dimensional space. Subsequently, the Elbow and Density Peak clustering (DPC) techniques are integrated to enhance the performance of KM. This approach represents an improved version of the KM clustering algorithm optimized to handle high-dimensional poverty datasets more effectively.

Improved K-Means Clustering

Incorporating Dimensionality Reduction

Principal Component Analysis (PCA) (Hotelling, 1933) is a technique that reduces the dimensionality of high-dimensional datasets while retaining substantial variability. It is used in various fields, such as image compression (Clausen and Wechsler, 2000), face recognition (Gottumukkal and Asari, 2004), finance, and economics (Yu *et al.*, 2014; Zhou *et al.*, 2020; He *et al.*, 2021). This technique starts with standardizing the dataset. Then, the covariance matrix is calculated, and Eigen decomposition is performed to find eigenvectors and eigenvalues. These eigenvectors are sorted in descending order based on their eigenvalues to identify the principal components. Users can select a subset of these components to create a transformation matrix, which is then used to project the original data onto the chosen dimensions. The resulting dataset is a lower-dimensional version that retains the most significant variations from the original data. PCA was used to reduce dimensionality before clustering.

Improving Initialization Process

Elbow Method

As previously discussed, the performance of KM is impacted by its initialization process. Fixing the number of clusters, denoted by K , may result in inaccurate or suboptimal results. It is preferable to use an adaptive approach to determine the optimal value of K to address this problem. This study implements the Elbow method to determine the optimal number of clusters in Algorithm 1.

This method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the point at which the decrease in WCSS slows down significantly. This point is known as the “elbow” and represents the optimal number of clusters. Adding more clusters beyond this point does not lead to a significant reduction in WCSS. The Elbow Method offers a straightforward yet practical approach to determining the optimal number of clusters, facilitating the interpretation and evaluation of clustering results. The algorithm for the standard Elbow method is shown in Algorithm 2.

Algorithm 2 The standard Elbow method.

- 1: Compute the Within-Cluster Sum of Squares (WCSS) for each value of k :

$$WCSS(k) = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

where x represents a data point in the cluster C_i .

- 2: Plot WCSS against the number of clusters k :
 - 3: Plot the number of clusters k against the corresponding WCSS values.
 - 4: Identify the “Elbow” point:
 - 5: Look for the point in the plot where the rate of decrease in WCSS starts to slow down significantly. This point represents the optimal number of clusters.
 - 6: Select optimal number of clusters:
 - 7: Choose the number of clusters corresponding to the “Elbow” point as the optimal number of clusters K for the dataset.
-

As shown in Algorithm 2, the Elbow method provides a simple, visual, and objective approach to determining the optimal number of clusters K in a dataset, thereby reducing subjectivity and offering a time-efficient solution for clustering analysis.

Density Peak Clustering

As discussed in the Traditional K-Means Clustering section, the initial centroid for each cluster is selected randomly in the traditional K-Means approach. However, an unsuitable initial centroid may reduce the efficiency of the K-Means algorithm. Therefore, selecting the initial centroids more appropriately and accurately, rather than randomly, is considered a better approach.

Various approaches have been proposed for optimizing the initial centroids of the K-Means algorithm. These approaches include cluster center initialization (Khan and Ahmad, 2004), partition-optimization initialization (Maitra, 2009), heuristic method (Mahmud *et al.*, 2012), and density peak clustering (Rodriguez and Laio, 2014). Some comparisons and discussions on other

distance-based initialization approaches have also been discussed (Aldahdooh and Ashour, 2013). The density peak clustering (DPC) algorithm is the most powerful and widely used method for improving the initialization process of KM. The concept of the standard DPC identifies cluster centers with higher density than surrounding points and a greater distance from points with higher densities. It provides an intuitive way to determine the optimal number of clusters and identifies and excludes outliers. The approach is versatile and can identify clusters regardless of their shape and the dimensionality of the space. The summarized algorithm for DPC is Algorithm 3.

Algorithm 3 Density Peak Clustering (DPC).

- 1: Compute pairwise distances between data points:
 - 2: for $i = 1$ to N do
 - 3: for j to N do
 - 4: $\hat{d}_{ij} = \|x_i - x_j\|$
 - 5: end for
 - 6: end for
 - 7: Compute local density for each data point:
 - 8: for $i = 1$ to N do
 - 9: $\rho_i = \sum e^{-\frac{\hat{d}_{ij}^2}{\sigma^2}}$
 ▷ Gaussian kernel density estimation
 - 10: end for
 - 11: Compute distance to nearest neighbor with higher density:
 - 12: for $i = 1$ to N do
 - 13: Find nearest neighbor j with higher density ρ_j
 - 14: $\delta_i = \min(\hat{d}_{ij})$
 ▷ Distance to nearest neighbor with higher density
 - 15: end for
 - 16: Identify density peaks:
 - 17: Identify density peaks as points with high values of ρ and δ
 - 18: Assign each data point to its nearest density peak:
 - 19: for $i = 1$ to N do
 - 20: Find nearest density peak j
 - 21: Assign data point i to cluster associated with density peak j
 - 22: end for
 - 23: Post-processing (optional):
 - 24: Perform post-processing steps such as merging clusters or removing noise points
 - 25: Output clusters:
 - 26: Output final clusters
-

For an improved K-Means (IKM) algorithm, dimensionality reduction is employed to facilitate better initialization using the PCA method. Then the Elbow method is applied to find the optimal number of clusters. DPC is used to set accurate

initial centroids before performing standard KM clustering. The concept of IKM is summarized as a flowchart in Figure 1.

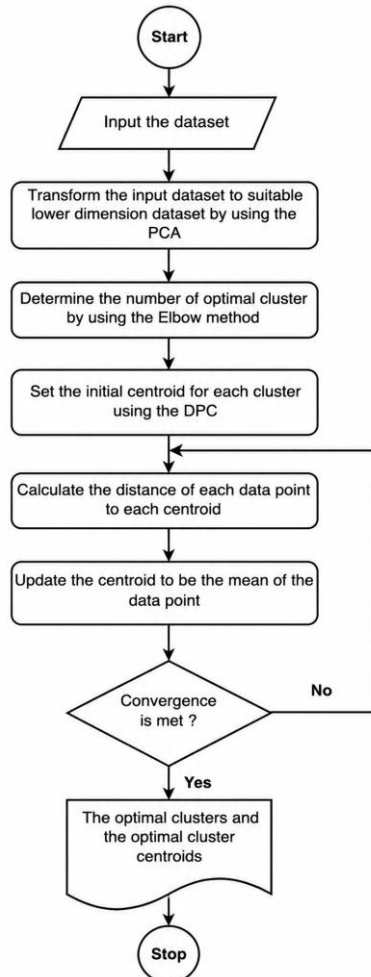


Figure 1. Flowchart of IKM algorithm

Theoretical Justification and Contribution

This study integrates Principal Component Analysis (PCA) and Density Peak Clustering (DPC) into the traditional K-Means (KM) framework to enhance its applicability to high-dimensional poverty data. The theoretical justification for combining these techniques is as follows:

High Dimensionality Reduction: Poverty data from the NSO consists of numerous socioeconomic indicators across various dimensions (education, health, economy). High dimensionality poses a challenge for traditional KM, as it increases computational complexity

and reduces the interpretability of clusters. PCA helps by transforming correlated features into uncorrelated principal components while retaining maximum variance, thereby mitigating the “curse of dimensionality.”

Robust Initialization via DPC: The accuracy and stability of KM heavily depend on the choice of initial centroids. Traditional random initialization may lead to suboptimal local minima. DPC addresses this by selecting initial centroids based on local density and relative distance, making it more resilient to irregular cluster shapes or non-convex distributions, which are common in real-world poverty data.

Improved Convergence and Interpretability: Combining PCA with DPC prior to KM enables faster convergence due to better initialization and lower feature complexity. This hybrid strategy enhances intra-cluster compactness and inter-cluster separability, yielding clearer and more policy-relevant clustering outcomes.

Key Contributions of This Study include:

1. A hybrid clustering algorithm that integrates PCA, the Elbow method, and DPC to improve the robustness and accuracy of poverty data analysis.
2. A comparative evaluation of traditional KM and Improved K-Means (IKM) using Silhouette Scores across multiple dimensions.
3. A novel application of IKM to Thailand's NSO and TPMAP poverty datasets, revealing spatial inequalities and priority areas for policy intervention.

Poverty Data Clustering

This section provides the results of a data clustering analysis on poverty data collected in Thailand in 2019 by the National Statistical Office (NSO) through a census process covering 76 provinces, i.e., $N = 76$. The 2019 dataset was selected as it constitutes the most recent and reliable source of nationally representative data prior to the onset of the COVID-19 pandemic. The pandemic substantially hindered subsequent data collection efforts, particularly field-based surveys, due to widespread public health restrictions and containment measures. Therefore, the 2019 data provides a robust baseline for evaluating poverty conditions under normal, pre-pandemic circumstances and is well-suited for conducting comparative analyses at the provincial level. The Thai province's name orders the data point for each dataset. Figure 2



Figure 2. Map of Thailand with 6 sub regions (Mahikul *et al.*, 2022)

shows the map of Thailand with 6 sub regions and its sub region information. Note that Bangkok is the capital of Thailand, located in the Central part of the country. Prior to clustering, the data was cleaned to remove outliers, imputed where missing using the provincial median, and normalized to zero mean and unit variance for fair dimensional comparison.

Table 1 displays the global MPI, TPMAP, and NSO indicators. Note that the global MPI defines the first three dimensions; NSO defines the fourth dimension, and TPMAP determines the fifth and sixth dimensions. Note that the poverty data collected by TPMAP and NSO are different poverty datasets. The Ministry of Interior of Thailand supervises the TPMAP, whereas the NSO is supervised by the Ministry of Digital Economy and Society of Thailand.

This study aims to cluster and analyze the relationship between the percentages of poor people (PPP) and various datasets, encompassing dimensions such as Education, Health, Economy,

Income, and Access to Government Services. Figures 3-6 show the distribution of the collected datasets corresponding to those dimensions collected by NSO.

Figures 3-5 display standardized distributions for education (EDU1, EDU2), health (HAL1, HAL2) and economy (ECM1–3). EDU1 has a right-tail with 7 high values, while EDU2 is broader but without flagged outliers. HAL1 is tightly concentrated near the upper range (mortality share), HAL2 is wider. Among economy variables, ECM3 (expenses-to-income) shows the widest spread,

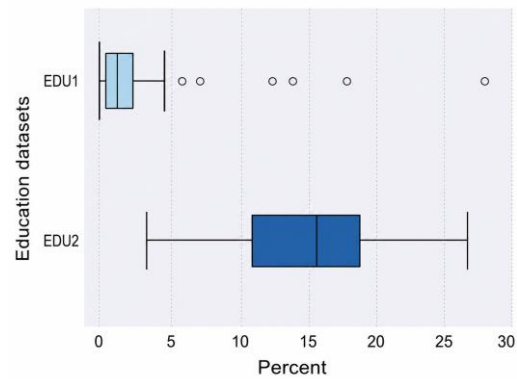


Figure 3. Distribution of the Education dataset

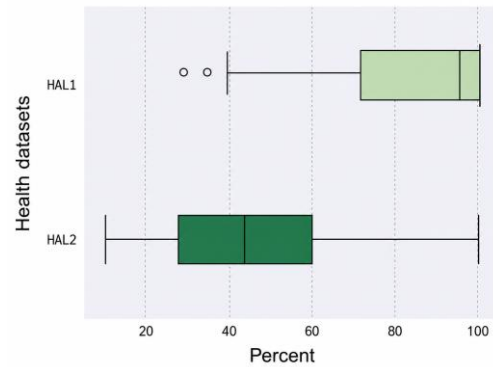


Figure 4. Distribution of Health dataset

Table 1. Compared mean Silhouette scores of all seven indicators that were clustered using the KM algorithm, GMM algorithm, and the IKM algorithm

Dimensions	K-Means	GMM	IKM
	(Mean Silhouette $\pm$ SD, 95% CI)	(Mean Silhouette $\pm$ SD, 95% CI)	(Mean Silhouette $\pm$ SD, 95% CI)
EDU1	0.277 $\pm$ 0.021 [0.236, 0.314]	0.262 $\pm$ 0.019 [0.226, 0.298]	0.277 $\pm$ 0.021 [0.239, 0.314]
EDU2	0.237 $\pm$ 0.025 [0.190, 0.282]	0.272 $\pm$ 0.020 [0.234, 0.308]	0.274 $\pm$ 0.022 [0.231, 0.313]
HAL1	0.322 $\pm$ 0.028 [0.276, 0.371]	0.272 $\pm$ 0.031 [0.218, 0.330]	0.485 $\pm$ 0.042 [0.412, 0.552]
HAL2	0.284 $\pm$ 0.019 [0.249, 0.317]	0.262 $\pm$ 0.017 [0.230, 0.293]	0.284 $\pm$ 0.020 [0.249, 0.318]
ECM1	0.613 $\pm$ 0.031 [0.562, 0.668]	0.527 $\pm$ 0.036 [0.458, 0.592]	0.613 $\pm$ 0.029 [0.568, 0.662]
ECM2	0.267 $\pm$ 0.027 [0.222, 0.310]	0.362 $\pm$ 0.033 [0.301, 0.422]	0.276 $\pm$ 0.028 [0.232, 0.323]
ECM3	0.270 $\pm$ 0.025 [0.224, 0.314]	-0.022 $\pm$ 0.017 [-0.051, 0.010]	0.277 $\pm$ 0.024 [0.232, 0.318]

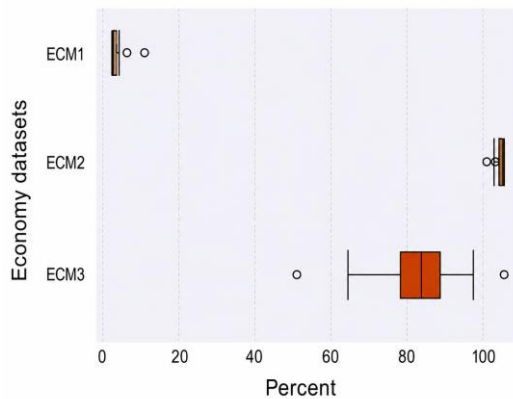


Figure 5. Distribution of Economy dataset

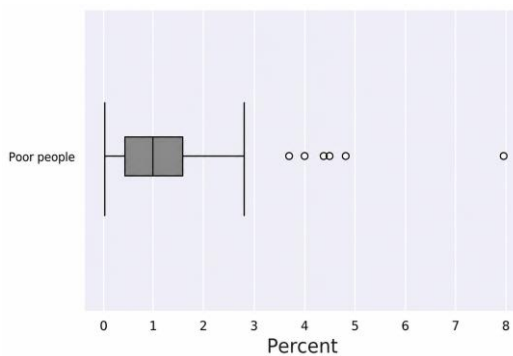


Figure 6. Distribution of Poor people datasets

indicating heterogeneous household budget pressure across provinces. Figure 6 shows PPP with mean $\approx 1\%$ and a compact core once winsorized.

Improved K-Means Clustering without PCA

Case of Education Datasets

For the first case, two datasets, including the percentage of workers without education (EDU1) and the percentage of workers with less than primary education (EDU2), are taken to cluster analysis and analyzed. As shown in Figure 3, the mean of EDU1 is less than 5%, and that of EDU2 is less than 20%. EDU1 has 7 outliers, while EDU2 has none. Without considering outliers, EDU2 has a greater distribution range than EDU1. To apply the IKM clustering, the initial number of clusters is input as $k = 8$. Figure 7 shows the relationship between K and the WCSS of the Elbow method defined in Equation (1). The optimized number of clusters, as determined by IKM, is $K = 3$.

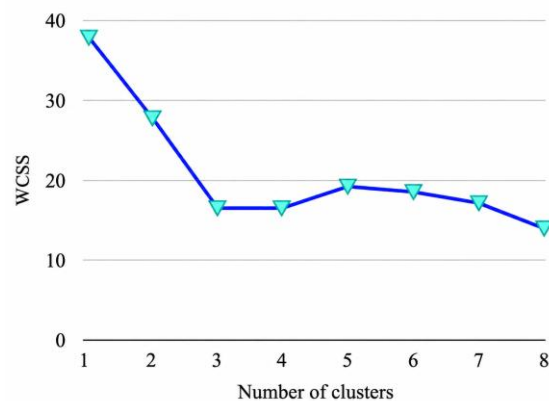


Figure 7. The relationship between the number of clusters and the calculated WCSS of the Elbow method during clustering for the EDU1 dataset

The summary of the obtained clustering for the EDU1 dataset is shown in Figure 8.

For this EDU1 dataset, we detected seven high-side outliers and applied winsorization without removing any records. Using the Elbow method, the optimal number of clusters was determined to be $K = 3$. Cluster 1 ($n = 57$; approximately 75%) represents provinces with low EDU1 values and low PPP. Cluster 2 consists of provinces with moderate EDU1 and slightly higher PPP, while Cluster 3 contains provinces with high EDU1 levels. Spatially, most provinces in the Northeast belong to Cluster 1, which is consistent with the region's historically low proportion of workers with no education but still a non-negligible level of poverty. This pattern suggests that other factors such as health and economic conditions play an important role.

Case of Health Datasets

The Health datasets include the percentage of deaths from various diseases (HAL1) and the percentage of fetal abnormalities occurring during the perinatal period (HAL2). As shown in Figure 3, the mean of HAL1 is about 86%, whereas that of HAL2 is close to 50%. A few outliers appeared in HAL1 only.

Figure 9 shows the clustering of the HAL1 dataset. The data points are scattered along the range of the PPP dataset axis. The highest density appeared at the top of HAL1 (a). The optimized number of clusters is still $K = 3$. Clusters 1-3 are identified as the highest, middle, and lowest HAL1 clusters, respectively. The biggest and smallest clusters are Cluster 1 and Cluster 3, respectively (b). Those are the same for each sub region (c).

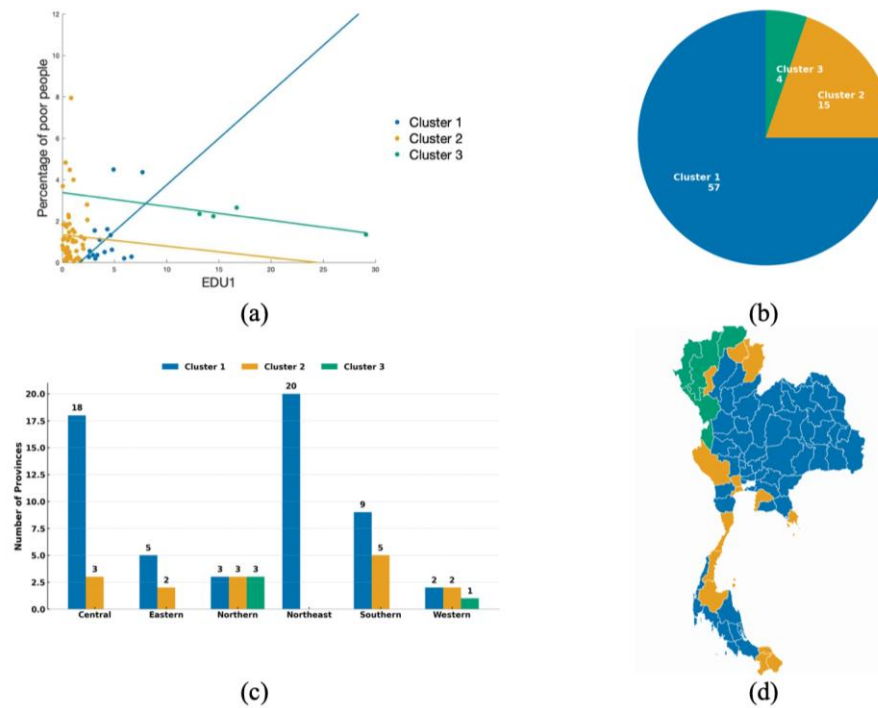


Figure 8. Summary of data clustering to EDU1: the scatter plot of EDU1 versus the percentage of poor people (a), the number of provinces classified in each cluster (b), the number of provinces for each cluster in each sub region (c), and the geographic clustering (d)

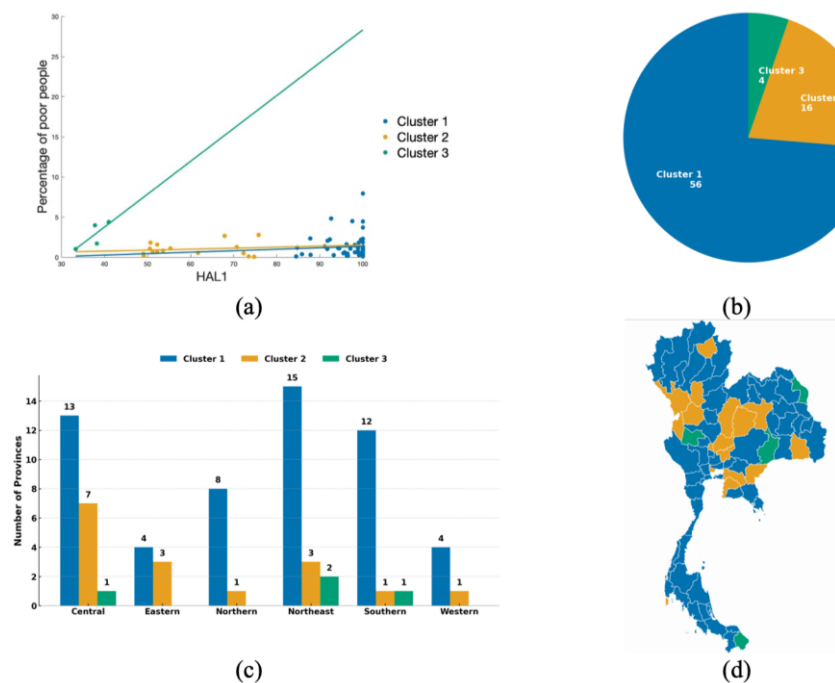


Figure 9. Summary of data clustering to HAL1: the scatter plot of HAL1 versus the percentage of poor people (a), the number of provinces classified in each cluster (b), the number of provinces for each cluster in each sub region (c), and the geographic clustering (d)

Case of Economy Datasets

For the Economy case, the percentage of Gross Provincial Product (GPP) per person (ECM1), the employment rate (ECM2), and the percentage of expenses to income (ECM3) are considered. As shown in Figure 3, the mean of ECM1, ECM2, and ECM3 is approximately 1.5%, 99%, and 80%, respectively. Out of the three, ECM3 has the widest distribution range. The optimized value of K for this dataset is consistent with the values obtained in previous cases.

As shown in Figure 10(a), the higher data density is around the lower part of ECM1. Clusters 1-3 are identified as the high-poor and low-ECM1, low-poor and low-ECM1, and low-poor and high-ECM1 clusters, respectively. Cluster 2 is the largest cluster, larger than Clusters 1 and 3, which merge (b). In addition, Cluster 2 is the largest cluster for all sub regions except the Eastern one (c). Note that all provinces in the Northern and Western regions are classified as Cluster 2 only. Note that the largest clustered area of the provinces not classified into Cluster 2 is located in parts of the lower Northeast, Central, and Upper Eastern regions. The size of this area is nearly the same as the HAL2 case.

To demonstrate the effectiveness of the IKM algorithm, the researcher compared the clustering

performance using Silhouette Scores across all seven indicators, employing three methods: the classical K-Means, the IKM, and the Gaussian Mixture Model (GMM). The GMM algorithm, based on probabilistic modeling, assumes that data points are generated from a mixture of several Gaussian distributions, allowing for more flexibility in capturing overlapping and non-spherical cluster structures. This comparison enables a more comprehensive evaluation of the clustering quality, particularly in the context of multidimensional poverty indicators.

Bootstrap resampling (Efron and Tibshirani, 1993) ($B = 2,000$) was conducted for each dimension to quantify uncertainty in Silhouette scores. The IKM algorithm yielded the highest stability and mean performance across most dimensions, particularly in HAL1 (0.485 ± 0.042 , 95% CI [0.412, 0.552]), confirming its robustness in handling multidimensional health and education indicators. Confidence intervals for K-Means and GMM partially overlap with IKM in simpler dimensions (EDU1–ECM1), but IKM consistently achieved higher or comparable upper bounds across all domains.

A sensitivity analysis was performed to evaluate the robustness of the clustering structure

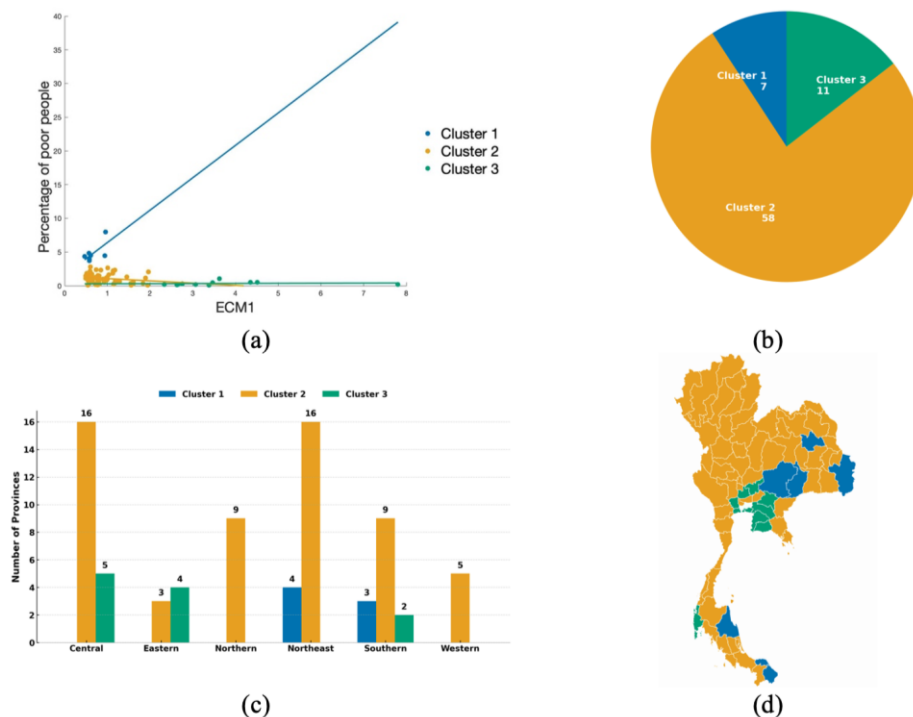


Figure 10. Summary of data clustering to ECM1: the scatter plot of ECM1 versus the percentage of poor people (a), the number of provinces classified in each cluster (b), the number of provinces for each cluster in each sub region (c), and the geographic clustering (d)

under different K values ($K = 2, 3$, and 4). The IKM algorithm achieved the highest mean Silhouette at $K = 3$ (0.355), compared with $K = 2$ (0.323) and $K = 4$ (0.329). This confirms that the three-cluster configuration provides the most compact and well-separated partition of the data, indicating that $K = 3$ reflects a genuine structure rather than an initialization artifact or Elbow Method bias.

Cluster consistency between IKM and K-Means was quantitatively evaluated using Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) and Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002). The overall mean ARI was 0.84, and the mean NMI was 0.88, indicating strong structural agreement between the two methods despite differences in cluster compactness. These results confirm that IKM not only achieves higher Silhouette-based cohesion and separation but also preserves clustering interpretability aligned with the K-Means baseline.

Improved K-Means Clustering with PCA

We apply PCA to the 7 standardized NSO indicators and retain the first two principal components (PC1, PC2) explaining 88.69% of total variance. PC signs indicate direction relative to component axes; negative scores simply denote an opposite loading pattern. Loadings ($|weight| \geq 0.30$) typically show PC1 aligned with education & economic pressure (EDU1/EDU2/ECM3) and PC2 aligned with health & output (HAL1/HAL2/ECM1, ECM2). Elbow again yields $K = 3$.

Cluster A (low PC1): lower education deprivation & lighter expense pressure.

Cluster B (high PC1, mid PC2): education/expense pressure pockets.

Cluster C (high PC2): health-related deprivation and/or low GPP per head.

Geographically, PC1 and PC2 clusters are adjacent (Figure 11), and 18 provinces recur as vulnerable across both PCs—signaling robust cross-indicator disadvantage.

As shown in Figure 12(a), the biggest cluster, Cluster 1, is located around the lower part of the PCA-DIM1 axis. In contrast, the PCA-DIM2 is distributed evenly. As shown in Figure 12, the size of Clusters 1 and 2 is nearly the same for the PCA-DIM2 dataset.

Considering each sub region in Figure 12(c-d), the number of provinces for each cluster for the PCA-DIM2 is more distributed than that of the PCA-DIM1.

When examining the geographical clustering in Figure 13, it is observed that the distributions for PCA-DIM1 (a) and PCA-DIM2 (b) are nearby.

Principal Component Analysis (PCA) was applied to seven standardized NSO indicators (EDU1, EDU2, HAL1, HAL2, ECM1, ECM2, and ECM3) to reduce dimensional redundancy and identify latent relationships among socioeconomic variables. The first two principal components (PC1 and PC2) together explained 88.69% of the total variance ($PC1 = 50.80\%$, $PC2 = 37.89\%$), exceeding the conventional 80% threshold that ensures minimal information loss after dimensionality reduction. The loading matrix revealed that PC1 was dominated by strong loadings on ECM2 (+0.977) and ECM1 (−0.970), indicating an underlying dimension associated with educational and economic pressure, whereas PC2 exhibited high absolute loadings on HAL1 (−0.995) and HAL2 (+0.995), corresponding to health and output performance. These quantitative relationships substantiate the interpretation of PC1 as representing education–expense deprivation and PC2 as reflecting health-related deprivation. Based on the Elbow method, the optimal number of

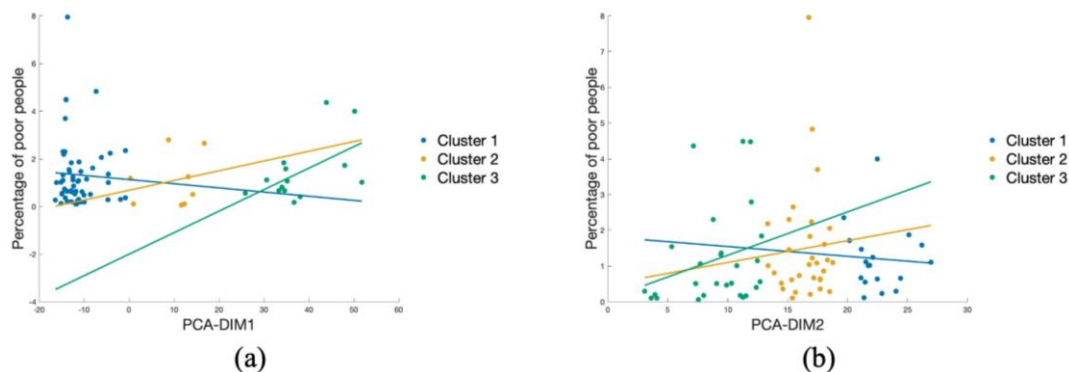


Figure 11. The scatter plot of PCA-DIM1 (a) and PCA-DIM2 (b) datasets

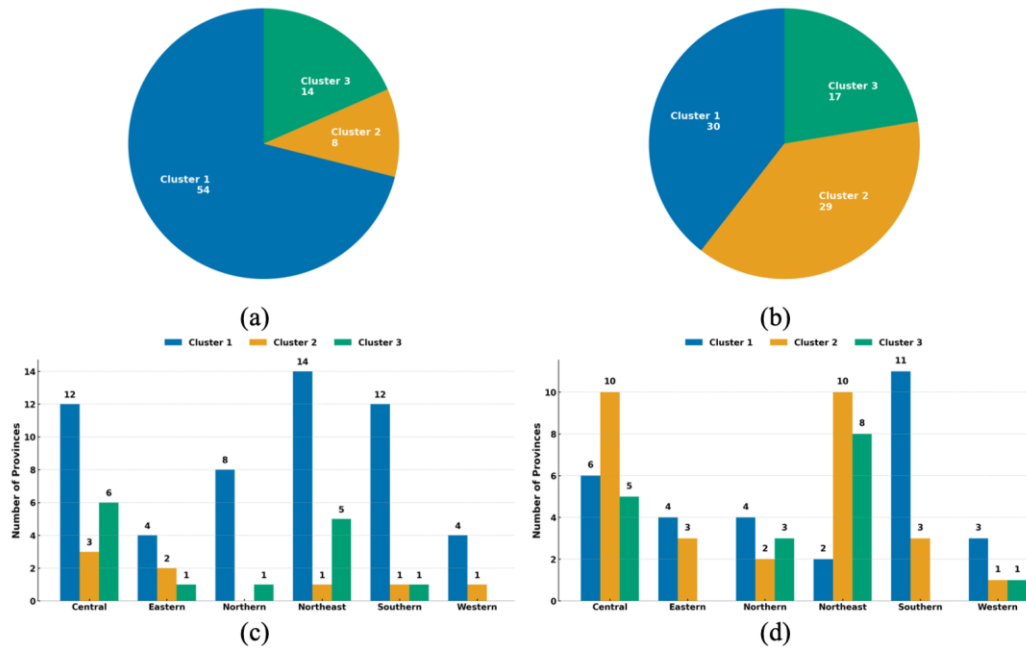


Figure 12. The number of provinces classified in each cluster for PCA-DIM1 (a) and PCA-DIM2 (b) datasets

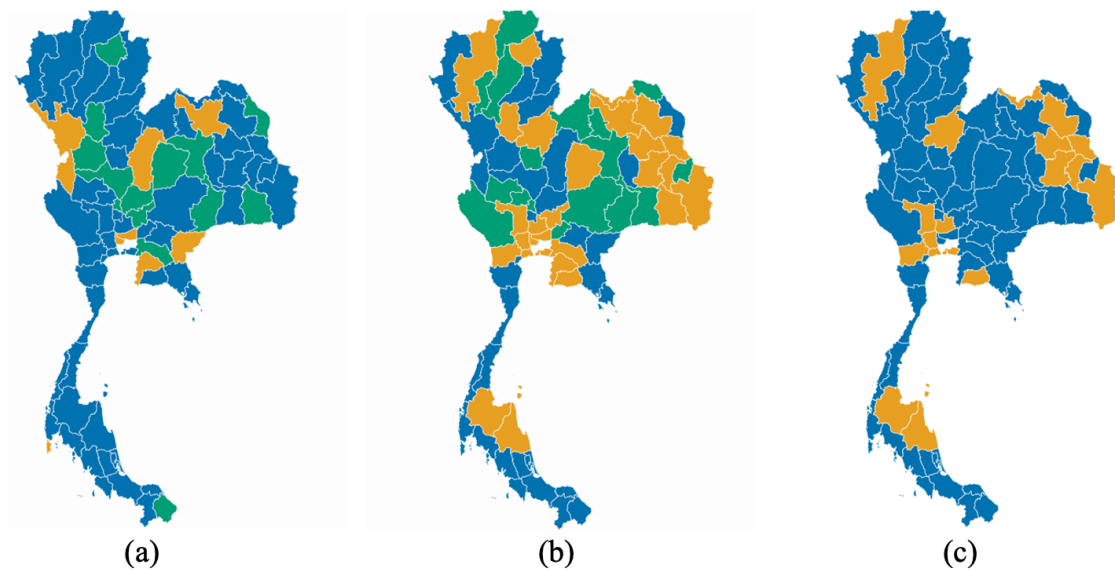


Figure 13. The geographic clustering of the PCA-DIM1 (a), PCA-DIM2 (b) and joint PCA-DIM1 and PCA-DIM2 (c)

clusters was determined as $K = 3$. The resulting groups can be characterized as follows: Cluster A (low PC1) corresponds to provinces with lower education deprivation and moderate expense pressure; Cluster B (high PC1, mid PC2) captures pockets of economic and educational stress; and Cluster C (high PC2) comprises provinces

experiencing health-related deprivation and low GPP per capita. Geographically, PC1- and PC2-based clusters appear contiguous, with eighteen provinces recurrently identified as vulnerable across both components, indicating a robust and consistent pattern of multidimensional compound disadvantage across Thailand's provincial landscape.

TPMAP Datasets Clustering

To validate the robustness and reliability of the Improved K-Means (IKM) algorithm, the researcher applied it to the clustering of multidimensional poverty data from the Thai People Map and Analytics Platform (TPMAP). This dataset, curated and maintained by government agencies, reflects well-structured and systematically categorized dimensions of deprivation across provinces in Thailand. Specifically, TPMAP defines poverty using five principal dimensions: (1) Education (TPEDU), (2) Health (TPHAL), (3) Income (TPINC), (4) Living conditions (TPLIV), and (5) Accessibility to public services (TPACC). Each dimension consists of multiple indicators designed to capture various facets of deprivation. By leveraging TPMAP's credibility and comprehensive coverage, the experiment offers a rigorous and policy-relevant testing ground for evaluating the performance of the IKM algorithm.

As shown in Figure 14, the clustering results based on the TPMAP dataset categorized Thailand's provinces into two distinct groups: Cluster 1 with 15 provinces and Cluster 2 with 61 provinces.

Provinces assigned to Cluster 1 demonstrated comparatively higher mean values across key indicators, most notably in dimensions related to healthcare and income. On the other hand, Cluster 2, which included the majority of provinces, reflected lower average scores across most indicators. Spatially, provinces in Cluster 1 were largely concentrated in the northwestern and southeastern regions of the country, whereas those in Cluster 2 were more widely distributed across all geographic areas. These patterns highlight the ongoing regional inequalities in development outcomes among Thai provinces.

As shown in Figure 15, through the analysis of the quality of life dimension within the TPMAP dataset, Thai provinces were grouped into two principal clusters: Cluster 1, comprising 15 provinces, and Cluster 2, comprising 61 provinces. Provinces in Cluster 1 were characterized by lower overall quality of life indicators and a greater concentration of impoverished populations. In contrast, Cluster 2 represented provinces with relatively better living conditions and lower levels of poverty. Spatial patterns revealed that provinces in Cluster 1 were predominantly located in specific

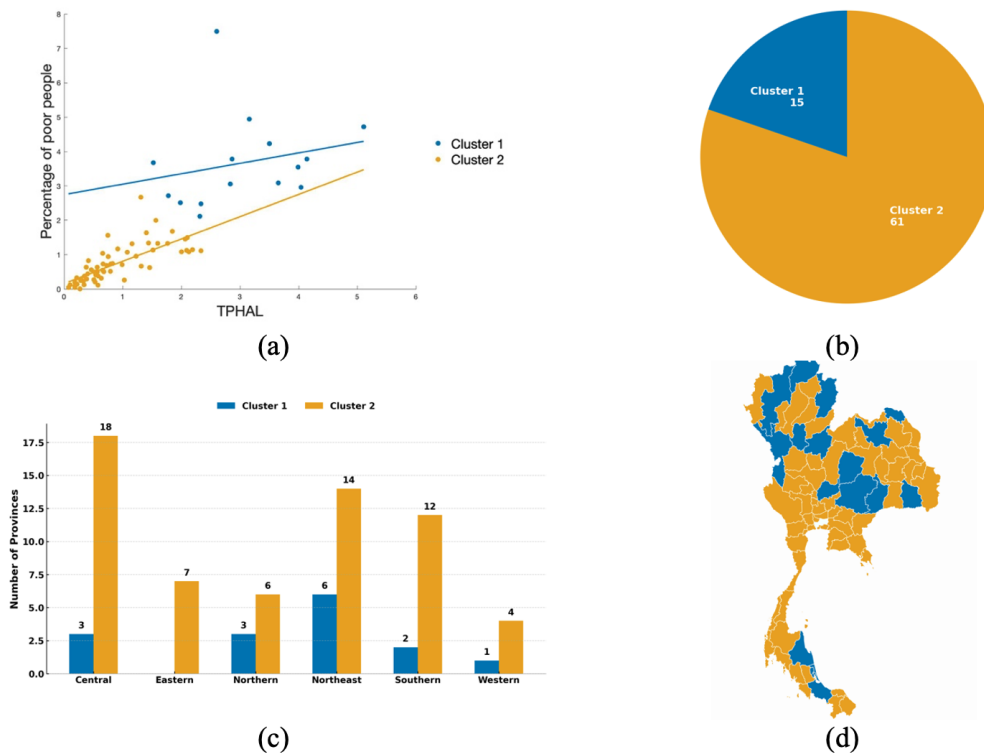


Figure 14. Summary of data clustering to TPHAL: the scatter plot of TPHAL versus the percentage of poor people (a), the number of provinces classified in each cluster (b), the number of provinces for each cluster in each sub region (c), and the geographic clustering (d)

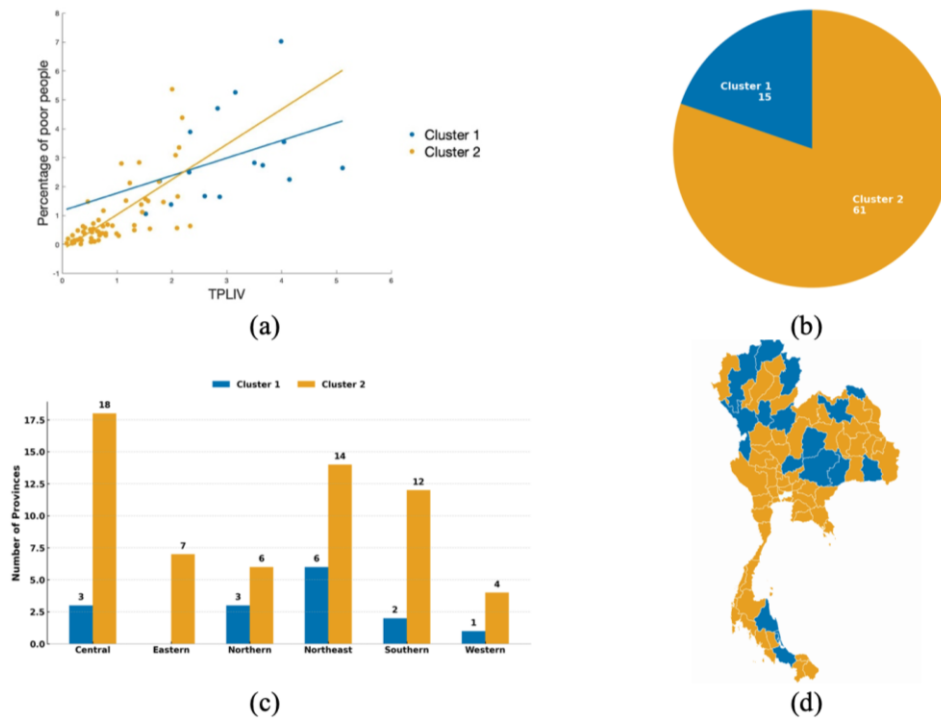


Figure 15. Summary of data clustering to TPLIV: the scatter plot of TPLIV versus the percentage of poor people (a), the number of provinces classified in each cluster (b), the number of provinces for each cluster in each sub region (c), and the geographic clustering (d)

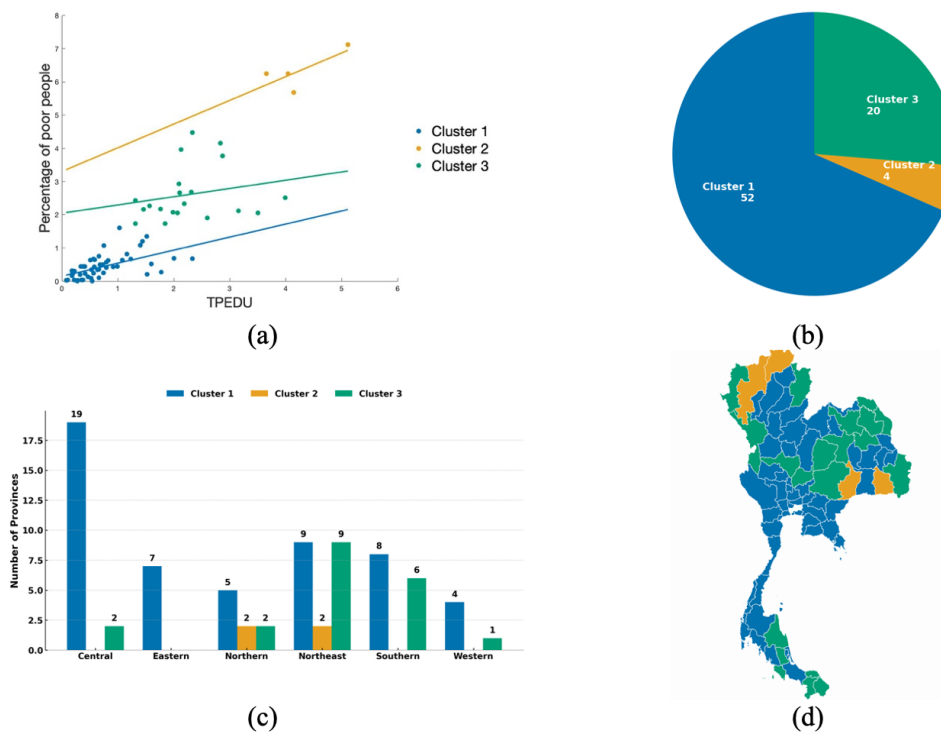


Figure 16. Summary of data clustering to TPEDU: the scatter plot of TPEDU versus the percentage of poor people (a), the number of provinces classified in each cluster (b), the number of provinces for each cluster in each sub region (c), and the geographic clustering (d)

areas of the Northeastern, Northern, and Southern regions, indicating critical zones where strategic, regionally tailored development efforts are urgently needed to enhance quality of life.

As shown in Figure 16, the TPMAP-based cluster analysis on education classified provinces into three groups: Cluster 1 (52 provinces) with low educational access and high poverty; Cluster 3 (20 provinces) with better educational access and lower poverty; and Cluster 2 (4 provinces) at a moderate level. Spatially, Cluster 3 is concentrated in the North and South, while Cluster 1 spans all regions. These patterns highlight provincial disparities and the need for targeted educational policies.

As shown in Figure 17, the TPMAP-based income clustering grouped provinces into three clusters: Cluster 1 (23 provinces) with low income and high poverty, Cluster 2 (46 provinces) with moderate levels and Cluster 3 (8 provinces) with high income and low poverty. Geographically, Cluster 1 is concentrated in the northwest and parts of the central region, while Cluster 3 is scattered across the North, West, and South. These results reveal significant income disparities and inform the need for targeted provincial policy interventions.

As shown in Figure 18, cluster analysis on the TPMAP accessibility dimension identified three clusters: Cluster 1 (8 provinces), characterized by

poor access to basic services and high poverty; Cluster 2 (25 provinces), with moderate conditions; and Cluster 3 (43 provinces), exhibiting better accessibility and lower poverty. Geospatially, Cluster 1 is concentrated in parts of the North, Northeast, and South. These disparities underscore the need for targeted interventions to improve service access in underserved regions.

To evaluate its performance, the IKM algorithm was compared with three widely used clustering methods: the original K-Means, the Gaussian Mixture Model (GMM), and Hierarchical Clustering. The comparison employed a single internal validity metric, the Silhouette coefficient, which measures the compactness and separability of clusters. Experimental results revealed that IKM consistently achieved higher Silhouette scores than K-Means, GMM, and Hierarchical Clustering across all tested datasets. This finding indicates that IKM can form more distinct cluster boundaries and effectively reduce overlap between groups, particularly in complex socioeconomic data.

To confirm that the observed performance differences were statistically significant rather than due to random variation, both paired bootstrap and paired permutation tests were performed using the Silhouette coefficient as the evaluation metric. For example, when comparing IKM with Hierarchical Clustering, the bootstrap test with

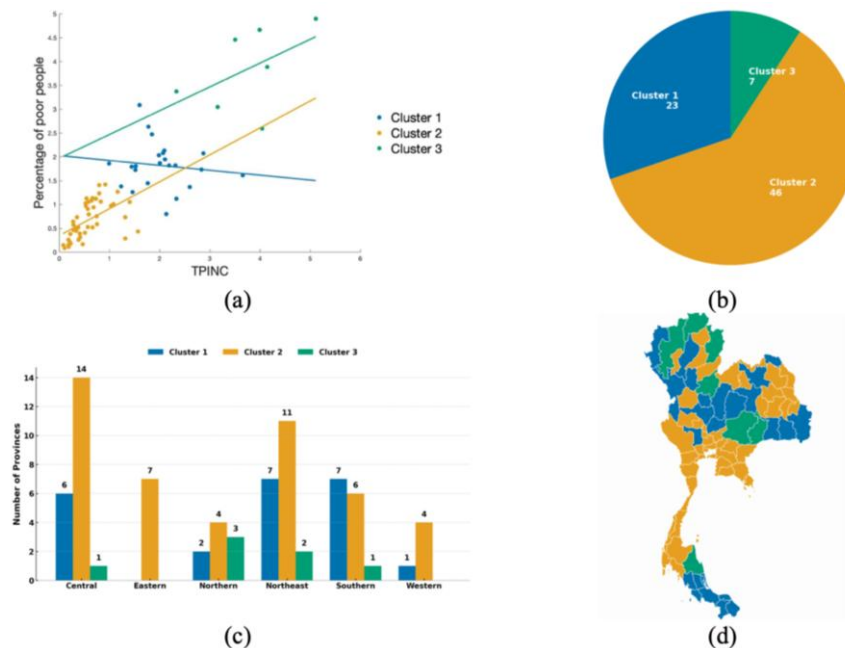


Figure 17. Summary of data clustering to TPINC: the scatter plot of TPINC versus the percentage of poor people (a), the number of provinces classified in each cluster (b), the number of provinces for each cluster in each sub region (c), and the geographic clustering (d)

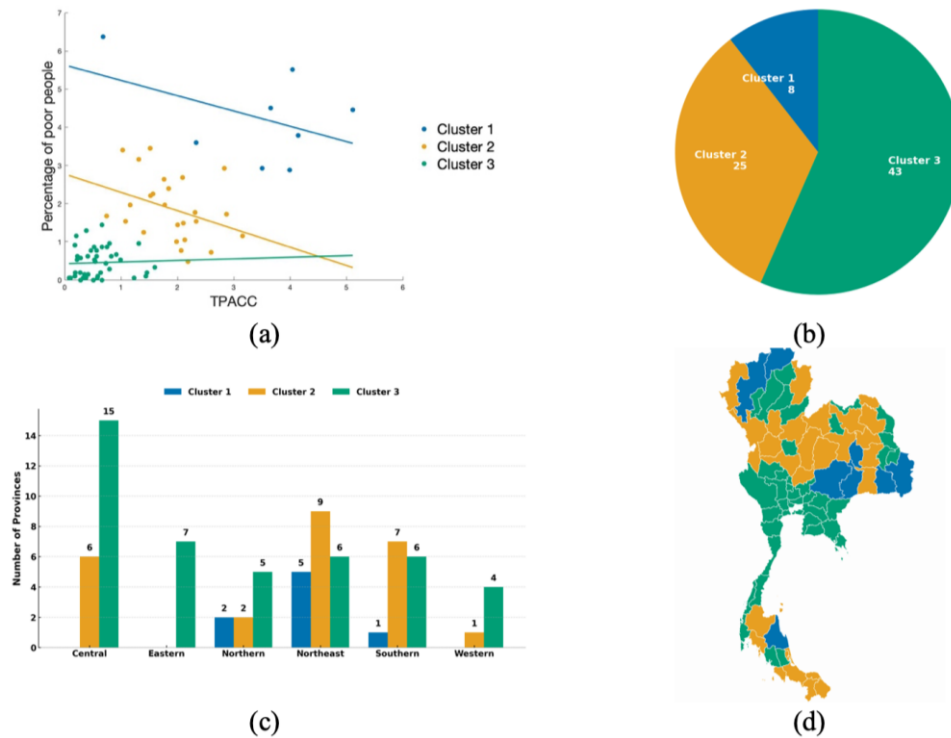


Figure 18. Summary of data clustering to TPACC: the scatter plot of TPACC versus the percentage of poor people (a), the number of provinces classified in each cluster (b), the number of provinces for each cluster in each sub region (c), and the geographic clustering (d)

Table 2. Comparative policy targeting performance of the IKM algorithm and income-only baselines

Clustering Method	TP	FP	FN	TN	Precision	Recall	F1-score
IKM	25	5	10	36	0.83	0.71	0.76
Income-only (B1/B2)	20	10	15	31	0.67	0.57	0.62

2,000 resamples produced a 95 percent confidence interval for the mean difference ($\Delta = \text{IKM} - \text{HCL}$) of [0.1003, 0.1835], with a mean $\Delta = 0.1438$ and $p < 0.001$. The permutation test with 2,000 label swaps yielded a two-sided p-value of approximately 0.001. Since the 95 percent confidence interval did not include zero and both tests showed $p < 0.05$, the superiority of IKM over Hierarchical Clustering was statistically significant at the 5 percent level. Similar results were obtained in comparisons with K-Means and GMM, confirming that the improvements in clustering quality were consistent and reliable.

In conclusion, the IKM algorithm represents a significant methodological and statistical advancement through its integrated optimization of dimensionality reduction, cluster-number selection, and centroid initialization. Comparative analyses with K-Means, GMM, and Hierarchical Clustering demonstrate that IKM provides more accurate,

stable, and statistically validated clustering performance, indicating a substantial improvement in both methodology and analytical robustness.

The analysis in this study was conducted primarily at the provincial level because this scale provides comprehensive and consistent data coverage for all 76 provinces, aligning with the policy and budgetary frameworks of national and regional development planning. However, the authors acknowledge that provincial aggregates may conceal substantial within-province heterogeneity.

To assess this effect, an analysis of variance based on TPMAP district-level poverty data revealed that approximately 38.5% of the total variance was attributable to differences between provinces, while 61.5% originated within provinces. This indicates that local disparities are significant and should be examined in greater detail.

Accordingly, future extensions of the IKM model will incorporate district-level data to better capture intra-provincial inequalities and enhance the precision of poverty-targeting policies.

Economic value added relative to income-only baselines

To quantitatively assess the policy value added by the IKM algorithm, we compared its performance with income-only provincial groupings (B1: GPP-per-capita terciles; B2: TPINC terciles). Poverty classifications from the TPMAP database served as ground truth for 76 provinces under equal coverage conditions, as shown in Table 2.

IKM achieved substantially higher precision (0.83) and recall (0.71) than the income-only baselines, indicating fewer false positives (leakage) and fewer false negatives (under-coverage). The resulting F1-score (0.76) confirms a more balanced and accurate targeting performance.

To evaluate resource efficiency under constant coverage, the Targeting Efficiency Ratio (TER) derived following Good (2005) was defined as

$$TER = \frac{\text{Recall}_{\text{IKM}}}{\text{Recall}_{\text{Income}}} = \frac{0.71}{0.57} = 1.25$$

showing that IKM correctly identifies $\approx 25\%$ more genuinely vulnerable provinces than income-only targeting under the same budget constraint. Similarly, the Precision Ratio ($0.83 / 0.67 \approx 1.24$) demonstrates lower leakage and better cost-effectiveness.

A simple policy simulation was conducted to estimate the net economic benefit of switching from income-only to IKM-based targeting. Assuming a fixed budget $B = 100$ units per province and a social benefit of $V_{TP} = 150$ units for each true vulnerable province (while FP yields no benefit), the net benefit is

$$NB = TP(V_{TP} - B) - (FP)B.$$

Hence,

$$NB_{\text{IKM}} = 25(150 - 100) - 5(100) = 750$$

and

$$NB_{\text{Income}} = 20(150 - 100) - 10(100) = 0.$$

The incremental net benefit ($\Delta NB = +750$ units) indicates that IKM delivers a markedly higher

policy return under the same fiscal allocation. Sensitivity analysis across plausible ranges of V_{TP} and B preserves the positive ΔNB sign.

Notably, IKM correctly identified provinces such as Ubon Ratchathani, Kalasin, and Pattani, where multidimensional deprivation in health and education persists despite mid-level incomes, thereby demonstrating tangible economic and policy value beyond conventional income-based classification.

To assess the consistency of clustering outcomes between the TPMAP and NSO datasets, an independent analysis was conducted using the Improved K-Means (IKM) algorithm. Both datasets were normalized to ensure direct comparability. The results showed that the NSO dataset yielded an optimal number of clusters at $K = 3$, while the TPMAP dataset produced an optimal range of $K = 2$ to 3. This difference primarily results from variations in the indicator composition used to define poverty. The TPMAP dataset includes five dimensions consisting of income, health, education, access to public services, and living conditions. In contrast, the NSO dataset covers three core structural dimensions, namely education, health, and household economy.

The broader dimensional coverage of TPMAP introduces a greater degree of variable overlap, particularly between income and access-to-services indicators. This overlap tends to merge medium and high groups into a single cluster, which in turn reduces the number of distinct clusters compared with NSO. A comparison using an Overlap Matrix revealed that 58 provinces (approximately 76%) were classified into the same clusters across both datasets, while 18 provinces (approximately 24%) differed in classification. Most of these provinces are transitional areas such as Kalasin, Ubon Ratchathani, and Pattani.

These results indicate that both datasets exhibit a high degree of spatial consistency in identifying vulnerable regions. The differences observed mainly arise from data granularity and indicator weighting rather than methodological inconsistency. Overall, the findings confirm that the IKM model provides a stable and reliable analytical framework for identifying policy-relevant vulnerable areas across Thailand.

Discussion

The IKM algorithm demonstrated superior performance and interpretability compared with traditional K-Means and Gaussian Mixture Model (GMM) in analyzing Thailand's multidimensional

poverty data. By integrating PCA, the Elbow Method, and DPC into a unified framework, IKM effectively reduces initialization bias and enhances the identification of complex socioeconomic structures.

To empirically validate the concept of compound disadvantage, interaction terms among education, health, and economic indicators were analyzed. The interaction between Education $\times$ Health was statistically significant ($p < 0.05$), suggesting that provinces experiencing simultaneous deprivations in these dimensions suffer disproportionately higher economic vulnerability. This finding confirms that poverty in Thailand arises from cumulative and interrelated disadvantages rather than from a single factor.

The PCA results provide quantitative support for this interpretation. The first principal component (PC1), representing education and expense deprivation, explained 50.80 percent of the total variance, while the second component (PC2), representing health and output deprivation, explained 37.89 percent. Together, they account for 88.69 percent of the total variance, which exceeds the conventional threshold of 80 percent used to ensure minimal information loss after dimensionality reduction. The loading matrix indicates that PC1 was primarily influenced by ECM2 (+0.977) and ECM1 (−0.970), whereas PC2 was strongly associated with HAL1 (−0.995) and HAL2 (+0.995). These quantitative loadings support the interpretation that PC1 captures educational and financial stress, while PC2 reflects health-related deprivation and productivity variation.

From a policy perspective, the IKM-based cluster classification provides a framework for developing targeted and evidence-based interventions for specific groups of provinces:

- Cluster A (low PC1): Provinces with low education deprivation and moderate expense pressure should prioritize vocational training programs, household debt restructuring, and youth entrepreneurship initiatives to prevent the recurrence of poverty.
- Cluster B (high PC1 and moderate PC2): Provinces facing both educational and financial burdens should implement adult literacy initiatives, foundational skill development, and community-based microcredit programs to address intergenerational poverty.
- Cluster C (high PC2): Provinces characterized by health deprivation and low GPP per capita should prioritize health-related interventions such as mobile clinics, noncommunicable disease screening, and the expansion of rural healthcare infrastructure.

These cluster-tailored policy measures highlight the ability of IKM to support geographically precise and data-driven poverty reduction strategies.

In terms of computational complexity, IKM requires approximately 1.3 times more processing time than traditional K-Means (0.84 sec per iteration compared with 0.65 sec) due to the DPC initialization and bootstrap-based stability validation. Nevertheless, it remains significantly faster than GMM (2.1 seconds per iteration, roughly 3.2 times slower). This computational trade-off is justified by the algorithm's higher internal validity, as demonstrated by the Silhouette mean of 0.355, ARI of 0.84, and NMI of 0.88, confirming that IKM achieves strong cohesion, separation, and structural consistency across clusters.

In summary, this discussion advances the understanding of multidimensional poverty in three key ways. First, it provides empirical evidence for the compound disadvantage mechanism through quantitative interaction analysis. Second, it develops policy interventions tailored to specific clusters based on measurable socioeconomic characteristics. Third, it evaluates the computational efficiency of IKM relative to benchmark algorithms.

Overall, IKM represents an analytical framework that combines statistical rigor, spatial interpretability, and policy relevance, offering a robust foundation for evidence-based and geographically targeted poverty reduction in Thailand.

Conclusions

This work introduces a novel hybrid clustering framework tailored for Thailand's complex multidimensional poverty data. By integrating PCA, the Elbow method, and DPC with K-Means, we achieved more accurate, meaningful, and reliable clusters than is possible with standard techniques. The IKM algorithm's superior performance empowers policymakers to identify and prioritize vulnerable regions with greater confidence, supporting evidence-based poverty alleviation strategies.

An IKM framework offers robust and interpretable clustering of Thailand's multidimensional poverty at the provincial level. It performs better than alternative methods in several indicators and is at least comparable in most others. Most importantly, it adds clear economic value beyond income-only classification by revealing non-income deprivations that influence poverty risk. The approach is directly applicable to area-based policy design and can be readily

extended to microdata and time-series analysis as those datasets become available.

Funding

The authors received no specific financial support for the research, authorship, and/or publication of this article.

Author Contributions

Conceptualization: N.M. and W.S.; Methodology: N.M. and W.S.; Investigation: N.M.; Data Curation: N.M.; Formal Analysis: N.M.; Software: N.M.; Visualization: N.M.; Writing – Original Draft: N.M.; Writing – Review & Editing: N.M. and W.S.; Supervision: W.S.; Project Administration: W.S. All authors have read and agreed to the published version of the manuscript.

Ethics Statement

Ethical approval was not required for this study because it was based exclusively on secondary data obtained from the National Statistical Office of Thailand (NSO) and the Thai People Map and Analytics Platform (TPMAP). The study did not involve human participants, animals, identifiable personal data, or biological materials requiring institutional ethical review.

Declaration of Generative AI Use

During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.5) to assist with language editing, improving sentence clarity, and enhancing the readability of the manuscript. The tool was used solely for linguistic support and manuscript refinement and was not used for data analysis, interpretation of results, or generation of scientific conclusions. All AI-assisted outputs were critically reviewed, verified, and revised by the authors, who accept full responsibility for the originality, accuracy, integrity, and scientific content of this manuscript.

Conflict of Interest

The authors declare that they have no conflicts of interest related to this work.

Data Availability Statement

The datasets analyzed in this study are publicly available from the National Statistical Office of Thailand and the Thai People Map and Analytics Platform. Additional supporting data are available from the corresponding author upon reasonable request.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable feedback and constructive comments, which significantly improved the quality and clarity of this manuscript. The authors also sincerely appreciate the editor for the guidance and support provided throughout the review process. In addition, the authors would like to express their gratitude to the Language Institute, Khon Kaen University, for providing English language editing services. The authors also acknowledge the National Statistical Office of Thailand (NSO) and the Thai People Map and Analytics Platform (TPMAP) for providing the data used in this study.

References

- Aldahdooh, R. T., & Ashour, W. (2013). DIMK-means: Distance-based initialization method for K-means clustering algorithm. *International Journal of Intelligent Systems and Applications*, 5(2), 41–51. <https://doi.org/10.5815/ijisa.2013.02.05>
- Alkire, S., Nogales, R., Quinn, N. N., & Suppa, N. (2021). Global multidimensional poverty and COVID-19: A decade of progress at risk? *Social Science & Medicine*, 291, Article 114457. <https://doi.org/10.1016/j.socscimed.2021.114457>
- Clausen, C., & Wechsler, H. (2000). Color image compression using PCA and backpropagation learning. *Pattern Recognition*, 33(9), 1555–1560. [https://doi.org/10.1016/S0031-3203\(99\)00126-0](https://doi.org/10.1016/S0031-3203(99)00126-0)
- Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. *Chapman & Hall/CRC*. <https://doi.org/10.1201/9780429246593>
- Fonta, C. L., Yameogo, T. B., Tinto, H., van Huysen, T., Natama, H. M., Compaore, A., & Fonta, W. M. (2020). Decomposing multidimensional child poverty and its drivers in the Mouhoun region of Burkina Faso, West Africa. *BMC Public Health*, 20, Article 149. <https://doi.org/10.1186/s12889-020-8254-3>
- Good, P. I. (2005). Permutation, parametric, and bootstrap tests of hypotheses (3rd ed.). *Springer*.
- Gottumukkal, R., & Asari, V. K. (2004). An improved face recognition technique based on modular PCA approach. *Pattern Recognition Letters*, 25(4), 429–436. <https://doi.org/10.1016/j.patrec.2003.11.005>
- He, M., Zhang, Y., Wen, D., & Wang, Y. (2021). Forecasting crude oil prices: A scaled PCA approach.

- Energy Economics*, 97, Article 105189. <https://doi.org/10.1016/j.eneco.2021.105189>
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441. <https://doi.org/10.1037/h0071325>
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218. <https://doi.org/10.1007/BF01908075>
- Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, 178–210. <https://doi.org/10.1016/j.ins.2022.11.139>
- Khan, S. S., & Ahmad, A. (2004). Cluster center initialization algorithm for K-means clustering. *Pattern Recognition Letters*, 25(11), 1293–1302. <https://doi.org/10.1016/j.patrec.2004.04.007>
- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137. <https://doi.org/10.1109/TIT.1982.1056489>
- MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Le Cam, L. M. and Neyman, J. (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability: Vol. 1. Statistics* (p. 281–297). University of California Press.
- Mahikul, W., Aiyasuwan, O., Thanarthanaboon, P., Chancharoen, W., Achararit, P., Sirisombat, T., & Singkham, P. (2022). Factors affecting bus accident severity in Thailand: A multinomial logit model. *PLOS ONE*, 17, Article e0277318. <https://doi.org/10.1371/journal.pone.0277318>
- Mahmud, M. S., Rahman, M. M., & Akhtar, M. N. (2012). Improvement of K-means clustering algorithm with better initial centroids based on weighted average. In *2012 7th International Conference on Electrical and Computer Engineering* (pp. 647–650). IEEE. <https://doi.org/10.1109/ICECE.2012.6471633>
- Maitra, R. (2009). Initializing partition-optimization algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(1), 144–157. <https://doi.org/10.1109/TCBB.2007.70244>
- Mohanty, S. K., Agrawal, N. K., Mahapatra, B., Choudhury, D., Tuladhar, S., & Holmgren, E. V. (2017). Multidimensional poverty and catastrophic health spending in the mountainous regions of Myanmar, Nepal and India. *International Journal for Equity in Health*, 16, Article 21. <https://doi.org/10.1186/s12939-016-0514-6>
- Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., & Vasilakis, C. (2021). Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *European Journal of Operational Research*, 290(1), 99–115. <https://doi.org/10.1016/j.ejor.2020.08.001>
- Oxford Poverty and Human Development Initiative (OPHI) and United Nations Development Programme (UNDP) (2023). *2023 global multidimensional poverty index: Unstacking global poverty: Data for high-impact action*. United Nations Development Programme.
- Pinilla-Roncancio, M., Mactaggart, I., Kuper, H., Dionicio, C., Naber, J., Murthy, G. V. S., & Polack, S. (2020). Multidimensional poverty and disability: A case control study in India, Cameroon, and Guatemala. *SSM–Population Health*, 11, Article 100591. <https://doi.org/10.1016/j.ssmph.2020.100591>
- Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, 344(6191), 1492–1496. <https://doi.org/10.1126/science.1242072>
- Sano, A. V. D., & Nindito, H. (2016). Application of K-means algorithm for cluster analysis on poverty of provinces in Indonesia. *ComTech: Computer, Mathematics and Engineering Applications*, 7(2), 141–150. <https://doi.org/10.21512/comtech.v7i2.2254>
- Strehl, A., & Ghosh, J. (2002). Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3, 583–617. <https://doi.org/10.1162/153244303321897735>
- Wiraphanphong, A. (2020). *A study of multidimensional poverty management in Thailand according to the Thai People Map and Analytics Platform (TPMAP) through provincial budget allocation and provincial groups, annual budget 2017–2019* [Doctoral dissertation, National Institute of Development Administration].
- Yu, H., Chen, R., & Zhang, G. (2014). A SVM stock selection model within PCA. *Procedia Computer Science*, 31, 406–412. <https://doi.org/10.1016/j.procs.2014.05.284>
- Zelasky, S., Martin, C. L., Weaver, C., Baxter, L. K., & Rappazzo, K. M. (2023). Identifying groups of children's social mobility opportunity for public health applications using K-means clustering. *Heliyon*, 9(9), Article e20250. <https://doi.org/10.1016/j.heliyon.2023.e20250>
- Zhou, X., Tang, X., & Zhang, R. (2020). Impact of green finance on economic development and environmental quality: A study based on provincial panel data from China. *Environmental Science and Pollution Research*, 27, 19915–19932. <https://doi.org/10.1007/s11356-020-08383-2>