



The Efficiency of Hybrid K-Means and Density Peak Clustering Algorithms for Multidimensional Poverty Clustering Using Thai People Map and Analytics Platform (TPMAP) Database

Narintra Mingolo¹ and Weerachai Sarakorn^{2,*}

¹ Department of Mathematics and English, Faculty of Education and Human Development, Roi Et Rajabhat University, Roi Et, 45120, Thailand

² Department of Mathematics, Faculty of Science, Khon Kaen University, Khon Kaen, 40002, Thailand

*Corresponding author: wsarakorn@kku.ac.th

Received 5 August 2025

Revised 10 March 2026

Accepted 23 April 2026

Abstract

Multidimensional poverty remains a critical challenge for sustainable development in Thailand, requiring analytical approaches that can reveal complex regional patterns of deprivation. This study evaluated the effectiveness of a hybrid clustering algorithm that integrated the K-Means and Density Peak Clustering (DPC) algorithms to categorize Thai provinces based on multidimensional poverty characteristics. The analysis used data from the Thai People Map and Analytics Platform (TPMAP), incorporating seventeen indicators across five key dimensions: living conditions, health, education, income, and access to basic services. The proposed hybrid method combined the computational efficiency of K-Means with the density-based cluster detection capability of DPC to improve clustering accuracy. The clustering results were evaluated using the Silhouette Score and visualized by scatter plots, bar charts, pie charts, and spatial maps to illustrate regional poverty patterns. The results indicated that the hybrid approach produced more coherent and well-separated clusters compared to the traditional K-Means algorithm, revealing clear inter-regional disparities in multidimensional poverty across Thai provinces. The findings suggested that poverty characteristics differ significantly by dimension and region, highlighting the importance of dimension-specific and regionally targeted policy interventions. This study demonstrated the potential of hybrid clustering algorithms as an effective data-driven tool to support evidence-based policymaking for reducing social inequality.

Keywords: Hybrid Clustering, K-Means, Density Peak Clustering, Multidimensional Poverty, TPMAP, Thailand

1. Introduction

Poverty remains one of the most critical challenges in global development and continues to receive significant attention in international policy frameworks, particularly within the United Nations Sustainable Development Goals (SDGs), where Goal 1 aims to eradicate poverty in all its forms (United Nations, 2015). In recent decades, poverty research has evolved beyond traditional income-based measures toward multidimensional perspectives that capture broader aspects of human deprivation. The multidimensional poverty framework proposed by Alkire and Foster (2011) [1] has become one of the most influential approaches for measuring poverty across multiple domains such as education, health, and living standards. This framework has been widely adopted by international organizations, including the United Nations Development Programme (UNDP) and the Oxford Poverty and Human Development Initiative (OPHI), to better understand the complexity of poverty and support targeted social policy design [2].

In Thailand, the Thai People Map and Analytics Platform (TPMAP) [3] was developed as a national data-driven system to monitor and analyze poverty at both household and subdistrict levels. The TPMAP integrates multidimensional indicators covering key aspects of well-being, including education, health, income, and access to basic public services. The availability of such large-scale and multidimensional datasets provides valuable opportunities for advanced data analytics to support evidence-based policymaking. However, identifying

meaningful patterns and similarities among regions or populations remains challenging due to the high dimensionality and complexity of poverty data.

Machine learning techniques, particularly clustering methods, have been increasingly applied to socioeconomic data analysis to reveal hidden patterns and structural inequalities within complex datasets [4-7]. Clustering allows the grouping of regions or populations with similar characteristics without requiring predefined labels, making it a useful tool for exploratory analysis and policy design. Among various clustering algorithms, the K-Means is one of the most widely used methods due to its conceptual simplicity and computational efficiency [6,8]. However, K-Means suffers from several well-known limitations, including sensitivity to centroid initialization and the requirement to predefine the number of clusters, possibly leading to unstable clustering results when data distributions are irregular or heterogeneous.

To address these limitations, Rodriguez and Laio (2014) introduced the Density Peak Clustering (DPC) algorithm, which identifies cluster centers as points having both high local density and large distances from points with higher density. Unlike traditional centroid-based approaches, DPC can detect clusters with arbitrary shapes and varying densities. However, the algorithm has high computational costs and can encounter difficulties in assigning cluster membership when applied to large datasets [9,10].

Previous studies have applied clustering techniques to poverty and socioeconomic datasets to identify regional disparities and support targeted policy interventions, with machine learning and remote sensing data used to estimate and predict poverty patterns in developing regions [11,12]. Other studies have applied clustering and spatial analysis to examine regional poverty characteristics and socioeconomic inequality [12,13]. Recent research has explored improved or hybrid clustering algorithms to enhance clustering performance when dealing with complex data structures [10,13]. However, many of these studies rely primarily on single clustering algorithms such as K-Means or hierarchical clustering, which may not fully capture the complex structure of multidimensional poverty data.

Despite recent progress, limited research has explored hybrid clustering approaches that integrate density-based and centroid-based algorithms for analyzing multidimensional poverty data using large-scale administrative datasets such as TPMAF, while the application of hybrid clustering techniques for identifying regional poverty patterns in Thailand remains underexplored.

Therefore, this study developed and evaluated a hybrid clustering algorithm that integrated DPC with K-Means to analyze multidimensional poverty patterns across Thai provinces using TPMAF data. The proposed framework used DPC to identify reliable initial cluster centers and cluster structures, while K-Means was applied to iteratively refine cluster assignments. The novelty of this research lies in the integration of density-based and centroid-based clustering techniques for multidimensional poverty analysis at the provincial level in Thailand. By combining the strengths of both approaches, the proposed hybrid algorithm improved clustering accuracy, stability, and interpretability. The clustering performance was evaluated using the Silhouette Score, which measures both intra-cluster cohesion and inter-cluster separation. The results provide clearer insights into regional poverty patterns and contribute to more effective data-driven policymaking for poverty alleviation and social development.

2. Materials and methods

2.1 Traditional Clustering Algorithms

Traditional clustering algorithms are essential techniques in unsupervised learning, designed to group data points based on similarity without requiring prior labels. Among these, the K-Means (KM) and DPC algorithms are two widely used approaches that rely on different clustering principles. KM performs centroid-based partitioning and is known for its simplicity and computational efficiency, making it suitable for large datasets with continuous numerical variables. By contrast, DPC identifies cluster centers by detecting points with high local density that are significantly distant from points with higher density, allowing it to capture clusters with irregular shapes and varying densities.

In this study, the multidimensional poverty indicators obtained from the TPMAF dataset were continuous numerical variables that were normalized before clustering to ensure comparable scales across indicators. Under this assumption, distance-based similarity measures such as Euclidean distance effectively represent the relationships among data points. The moderate size of the provincial dataset and the potential presence of heterogeneous poverty patterns across regions make both centroid-based and density-based clustering approaches suitable for analysis. Therefore, KM provided computational efficiency for iterative clustering, while DPC identified reliable cluster centers and captured complex data structures. Understanding the complementary strengths and limitations of these algorithms provided a strong foundation for developing hybrid clustering approaches, particularly for multidimensional poverty datasets where both clustering accuracy and interpretability are important.

2.1.1 KM Clustering

The KM clustering algorithm, initially introduced by MacQueen in 1967 [8], is widely used to partition a set of n data points into K non-overlapping clusters. The algorithm minimizes the within-cluster sum of squares (WCSS), which quantifies the compactness of the clusters. Given a dataset $X = \{x_1, x_2, x_3, \dots, x_n\}$, the objective function of KM is defined as:

$$J = \sum_{j=0}^k \sum_{i=1}^n |x_i^{(j)} - c_j|^2$$

where J is the sum of square error, and $|x_i^{(j)} - c_j|^2$ is the distance between a data point $x_i^{(j)}$ and centroid c_j .

The algorithm typically follows an iterative process as:

1. Initialize K centroids randomly or by heuristic methods.
2. Assign each data point to the cluster with the nearest centroid.
3. Update the centroid of each cluster based on the mean of the assigned points.
4. Repeat steps 2 and 3 until convergence.

Despite its simplicity and computational efficiency, KM is sensitive to the initial centroid positions and assumes that the clusters are spherical and of similar size and density, which limits its effectiveness in handling arbitrary-shaped clusters.

To address these known drawbacks of the traditional KM algorithm, a variety of improved approaches have been proposed. One such enhancement was introduced by Kanungo et al. (2002) [14], who developed the Filtering Algorithm. This technique employed a kd-tree data structure along with geometric filtering to discard redundant candidate centroids, thereby reducing computation time and accelerating the clustering process—particularly for datasets with well-separated clusters. Another advancement was suggested by Bradley et al. (2000) [15]. They presented a constrained KM variant that included minimum cluster size constraints, thereby preventing the formation of empty or excessively small clusters, which often occur in high-dimensional data. Bradley and Fayyad (1998) [16] focused on improving the initialization of centroids. They generated multiple subsamples to identify more reliable starting points to reduce the likelihood of the algorithm settling in suboptimal local minima.

More recently, Sinaga and Yang (2020) [17] introduced the U-KM algorithm, which performs clustering in a fully unsupervised manner. Their method integrates an entropy-based penalty mechanism to automatically determine the most appropriate number of clusters, eliminating the need for pre-specification and making the method more adaptable to various types of data.

2.1.2 Density Peak Clustering

Density Peak Clustering (DPC), proposed by Rodriguez and Laio (2014) [9], identifies cluster centers as data points that are surrounded by neighbors with lower local density and are located at a large distance from any other point with higher density. The two main concepts used in DPC are local density ρ_i and minimum distance δ_i to a point of higher density. The local density ρ_i of point x_i is defined as:

$$\rho_i = \sum_j e^{-\frac{d_{ij}^2}{\sigma^2}}$$

where d_{ij} is the distance between points x_i and x_j , and σ is a cut-off distance. The distance δ_i is defined as:

$$\delta_i = \min_{j: \rho_j > \rho_i} d_{ij}$$

The maximum distance $\delta_i = \max_j d_{ij}$ is assigned to the point with the highest density. The distance δ_i represents the minimum distance between point x_i and any other point with a higher local density. This measure is used to identify potential cluster centers in the dataset. In the DPC algorithm, cluster centers are typically characterized by having both high local density ρ_i and large distance δ_i . Therefore, δ_i helps distinguish points that are far from other high-density points, indicating that they are likely to serve as cluster centers.

Cluster centers are typically identified by plotting δ_i versus ρ_i and selecting points with both high density and high distance. Other points are assigned to the cluster of their nearest neighbor with higher density. DPC does not require the number of clusters to be specified in advance and is capable of detecting clusters of arbitrary shape, but its performance depends heavily on the choice of the cut-off distance σ .

The main steps are as follows:

1. Calculate the Euclidean distance between all pairs of data points.
2. Select a cut-off distance such that each point has 1–2% of the total number of data points as neighbors.
3. Estimate local density ρ_i
4. Compute the minimum distance to higher density δ_i

5. Calculate the decision value $\rho_i \times \delta_i$
6. Identify the data points with the highest values as cluster centers.
7. Assign each point to the same cluster as its nearest neighbor with a higher density.
8. (Optional) Identify outliers. Points with low local density and large δ values may be considered as outliers.

Several enhancements have been proposed to improve DPC in terms of accuracy, efficiency, and adaptability to complex data. Enhanced DPC (EDPC) by Cao et al. (2016) [18] replaces count-based density estimation with a Gaussian kernel, reducing sensitivity to the cut-off distance and improving cluster center detection, especially in asymmetric or overlapping data. Fast DPC (FDP), introduced by Wang et al. (2017) [19], eliminates manual decision graph analysis and uses a heuristic approach for center selection, making it scalable for large datasets while preserving clustering quality. Manifold DPC (MDPC) by Tian et al. (2019) [20] incorporates manifold learning techniques like Laplacian Eigenmaps to better capture the intrinsic geometry of data, enabling accurate clustering of non-Euclidean structures such as rings or nested patterns.

2.2 Clustering in Poverty Data Analysis

Multidimensional poverty data from sources like the TPMAP [3] include diverse indicators across education, health, economy, and accessibility, making them complex and high-dimensional. Clustering is essential for identifying regions or populations with similar poverty profiles to support targeted policy interventions. The KM algorithm is widely used for its simplicity, but it struggles with irregular structures in the data. By contrast, DPC handles non-convex shapes and varying densities more effectively. To address the limitations of each method, recent studies have proposed hybrid approaches that use DPC to initialize KM centroids based on high-density peaks, improving robustness and avoiding poor local optima. These hybrid models consistently outperform standalone methods, especially with complex data, emphasizing the value of algorithm selection and hybridization in producing accurate, interpretable results for evidence-based policymaking.

This study developed and evaluated the effectiveness of a hybrid clustering algorithm by combining KM and DPC for clustering multidimensional poverty data from the TPMAP to provide comprehensive indicators related to education, health, income, and accessibility to basic services.

2.2.1 Data Preparation

The TPMAP dataset was initially examined to handle missing values and outliers. All features were then normalized to ensure that distance calculations during clustering were accurate and stable.

2.2.2 A Hybrid K-Means and Density Peak Clustering Algorithm

To enhance initialization in clustering, this study adopted DPC to determine the optimal number of clusters (K) and to select initial centroids for the KM algorithm. DPC calculates two key metrics as local density (ρ), calculated as $\rho_i = \sum_j e^{-\frac{d_{ij}^2}{\sigma^2}}$, where σ is the cut-off distance selected as a certain percentile of all pairwise distances, and distance to the nearest higher-density point (δ), defined as $\delta_i = \min_{j: \rho_j > \rho_i} d_{ij}$. Points exhibiting both high ρ and δ values are identified as cluster centers. Once initialized, the standard KM process was performed. The detailed steps of the proposed hybrid KM–DPC algorithm are illustrated as follows:

Algorithm 1 A Hybrid K-Means and Density Peak Clustering Algorithm

Require:

- Set of data points $X = \{x_1, x_2, x_3, \dots, x_n\}$

Ensure:

- Number of clusters K , set of cluster centroids $U = \{\mu_1, \mu_2, \mu_3, \dots, \mu_K\}$, and set of clusters $C = \{C_1, C_2, C_3, \dots, C_K\}$

Initialization step:

1. Compute pairwise distances $d_{ij} = ||x_i - x_j||$
2. Determine a cut-off distance σ as a certain percentile of d_{ij}
3. Compute local density: $\rho_i = \sum_j e^{-\frac{d_{ij}^2}{\sigma^2}}$
4. Compute the distance to the nearest higher density point: $\delta_i = \min_{j: \rho_j > \rho_i} d_{ij}$
5. Select K points with the highest $\rho_i \times \delta_i$ as initial centroids U

Assignment step:

6. For each point x_i , compute the distance to all centroids μ_j
7. Assign x_i to the nearest centroid μ_j

Update step:

8. For each cluster $j = 1$ to K :

$$\mu = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i$$

where S_j is the set of data points in cluster j

Convergence:

9. Repeat steps 6 - 8 until the centroids converge or a maximum number of iterations is reached

2.2.3 Cluster Evaluation using Silhouette Score

The clustering results were evaluated using the Silhouette Score [9], which reflects intra-cluster cohesion and inter-cluster separation:

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

where $a(i)$ is the average distance between point i and other points in the same cluster, and $b(i)$ is the average distance between point i and the nearest neighboring cluster. A Silhouette Score near 1 indicates well-clustered data, while negative values imply incorrect clustering. This proposed methodology enhanced the precision of initial centroid selection in KM and mitigated the randomness inherent in traditional initialization, thereby improving clustering performance on multidimensional poverty data.

3. Results and discussion

This section presents the results of provincial clustering in Thailand using a hybrid KM and DPC algorithm. The analysis was applied to 17 multidimensional poverty indicators retrieved from the TPMAP, spanning five core dimensions as Health (4 indicators), Living Conditions (4 indicators), Education (4 indicators), Income (3 indicators), and Access to Basic Services (2 indicators). The distribution of the dataset utilized in this study is illustrated in Figure 1.

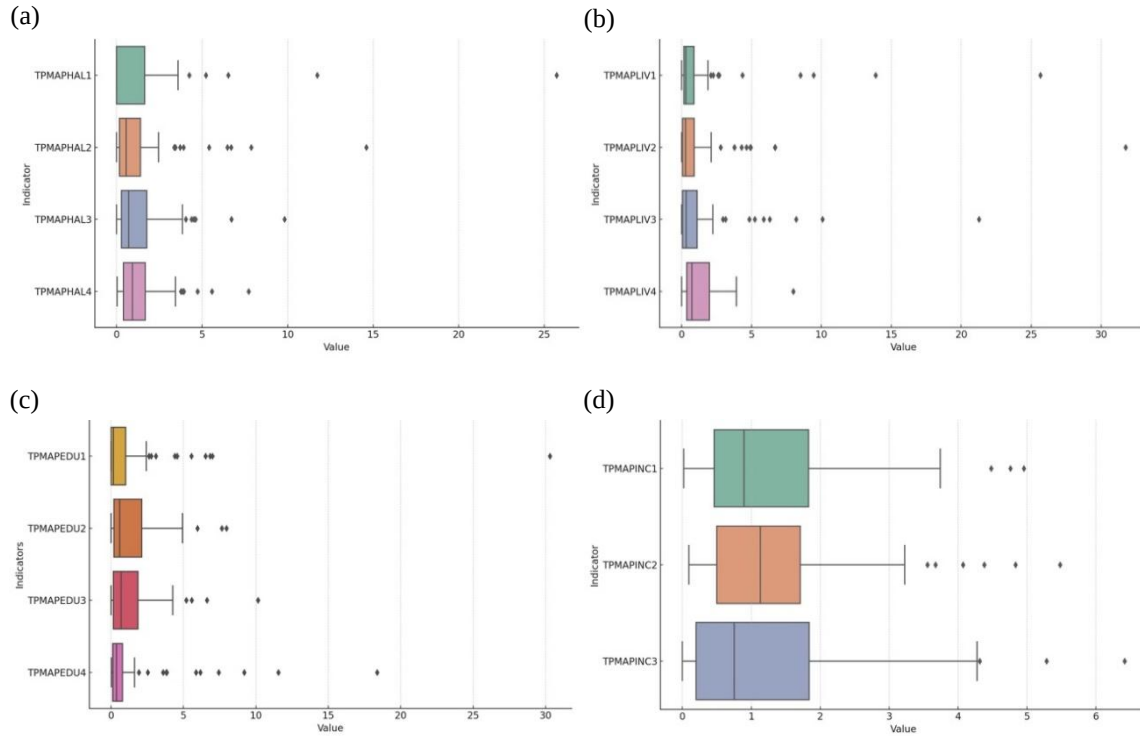


Figure 1 Distribution of the 17 multidimensional poverty indicators from the TPMAP used for provincial clustering, consisting of health (a), living conditions (b), education (c), income (d), access to basic services (e), and percentage of poor people (f). The boxplots show noticeable variability and skewness among several indicators, as well as the presence of outliers in certain provinces. These patterns indicate substantial regional heterogeneity in multidimensional poverty conditions, suggesting that provinces exhibit different deprivation profiles across dimensions. Such variability supports the application of clustering techniques to identify groups of provinces with similar poverty characteristics.

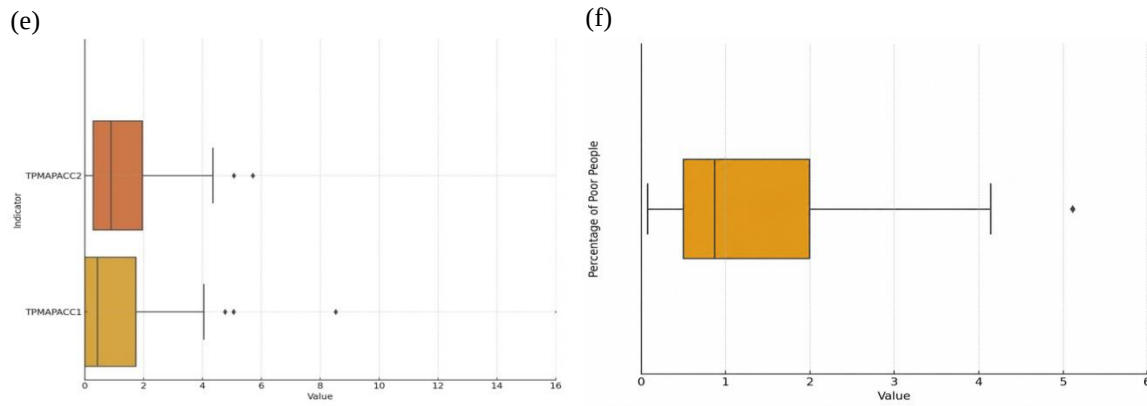


Figure 1 Distribution of the 17 multidimensional poverty indicators from the TPMAP used for provincial clustering, consisting of health (a), living conditions (b), education (c), income (d), access to basic services (e), and percentage of poor people (f). The boxplots show noticeable variability and skewness among several indicators, as well as the presence of outliers in certain provinces. These patterns indicate substantial regional heterogeneity in multidimensional poverty conditions, suggesting that provinces exhibit different deprivation profiles across dimensions. Such variability supports the application of clustering techniques to identify groups of provinces with similar poverty characteristics. (Cont.)

3.1 Health Datasets

Health indicators are essential in multidimensional poverty analysis because they reflect physical well-being and quality of life. This study used TPMAP data to assess four indicators as deaths from non-communicable diseases (TPMAPHAL1), underweight children under five (TPMAPHAL2), lack of universal health coverage (TPMAPHAL3), and chronic illnesses (TPMAPHAL4). These indicators capture various aspects of health vulnerability, including access to healthcare, basic health status, and chronic conditions. A hybrid clustering algorithm combining KM and DPC was applied to group provinces based on these indicators. TPMAPHAL1 is presented as a case example, while the remaining indicators are summarized using average Silhouette Scores in the next section.

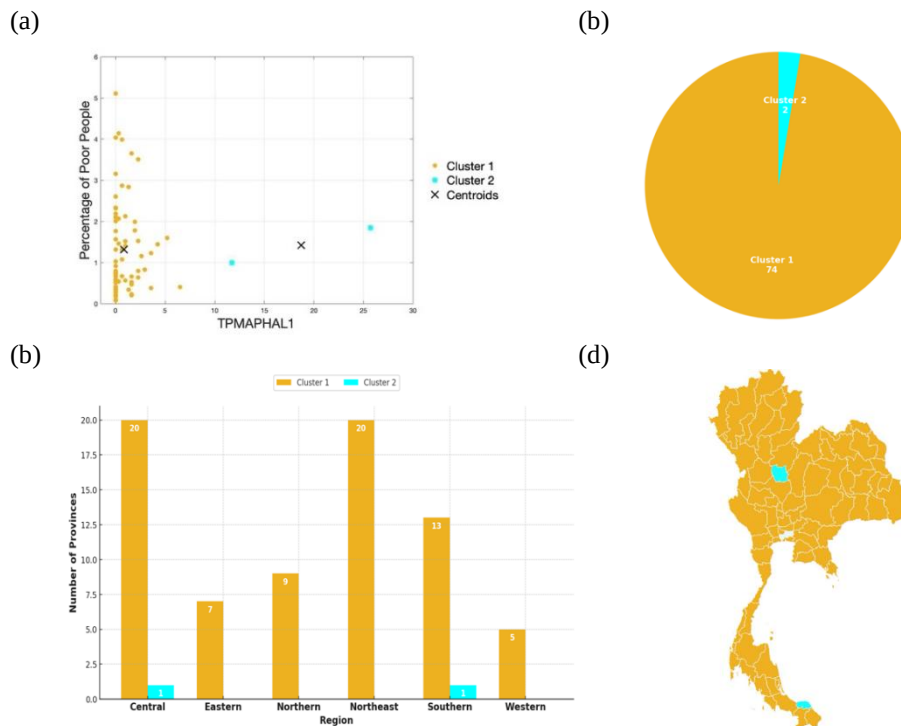


Figure 2 Summary of data clustering to TPMAPHAL1 as the scatter plot of TPMAPHAL1 versus the percentage of poor people (a), the number of provinces classified in each cluster (b), the number of provinces for each cluster in each sub-region (c), and the geographic clustering (d).

Figure 2 shows Thai provinces grouped by health (TPMAPHAL1) using a hybrid KM and DPC algorithm. The provinces fall into two clusters as Cluster 1 (74 provinces) with low TPMAPHAL1 scores and high poverty, mainly in Central and Northeastern Thailand, and Cluster 2 (2 provinces) with better health and reduced poverty in Central and Southern Thailand. The figures reveal an imbalance in cluster size and distribution, highlighting health disparities and the need for targeted policies.

3.2 Living Condition Datasets

Living conditions represent a key dimension in multidimensional poverty analysis, reflecting both the quality and stability of individuals' daily lives. This study employed four indicators from the TPMAP to assess this dimension as the percentage of households without access to clean drinking water (TPMAPLIV1), without proper sanitation (TPMAPLIV2), without electricity (TPMAPLIV3), and with inadequate housing (TPMAPLIV4). These indicators revealed disparities in infrastructure and environmental standards across provinces. The hybrid clustering technique identified distinct patterns of deprivation in living conditions, thereby supporting more precise and equitable policy formulation. The clustering results using TPMAPLIV1 are presented as a case example, while the remaining indicators are summarized by their average Silhouette Scores in the subsequent section.

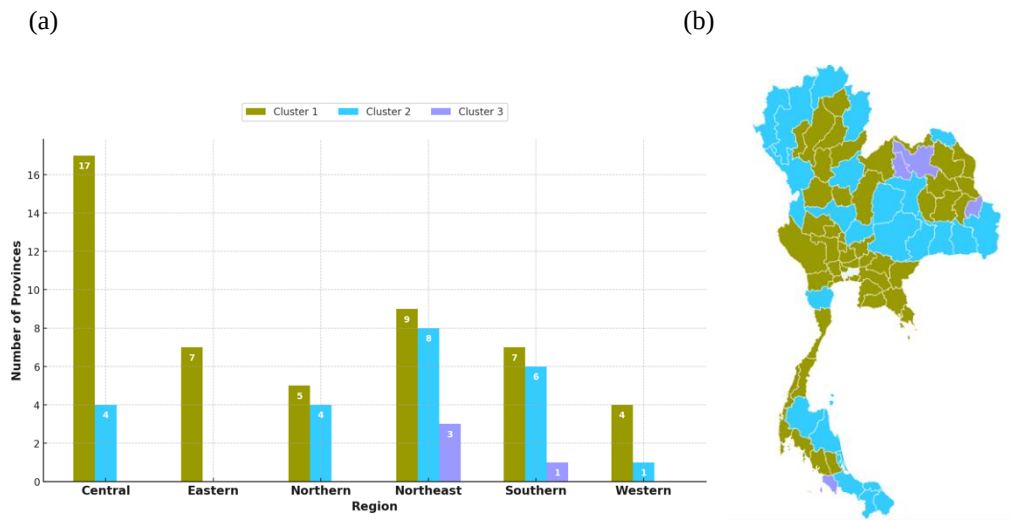


Figure 3 Summary of data clustering to TPMAPLIV1 as the number of provinces for each cluster in each sub-region (a) and the geographic clustering (b).

Figure 3 presents the clustering of Thai provinces based on TPMAPLIV1 using the hybrid KM–DPC algorithm, resulting in three clusters as Cluster 1 (49 provinces), Cluster 2 (23), and Cluster 3 (4). Figure 3a shows Cluster 1 as dominant across all the regions, especially in Central Thailand, while Clusters 2 and 3 are found in provinces mainly in Southern and Northeastern Thailand. Figure 3b illustrates the spatial distribution, highlighting regional disparities in living conditions and the need for tailored, area-specific policies.

3.3 Education Datasets

Education plays a crucial role in long-term poverty reduction by influencing employment opportunities, income, and overall well-being. This study used four TPMAP indicators to examine provincial disparities as the percentage of individuals with no formal education (TPMAPEDU1), with less than primary education (TPMAPEDU2), school-aged children not in school (TPMAPEDU3), and adults lacking basic literacy (TPMAPEDU4). These indicators captured both educational attainment and access, reflecting structural inequality. The hybrid clustering algorithm revealed spatial patterns of educational deprivation. TPMAPEDU1 is presented as a case study, while the remaining indicators are summarized by average Silhouette Scores in the following section.

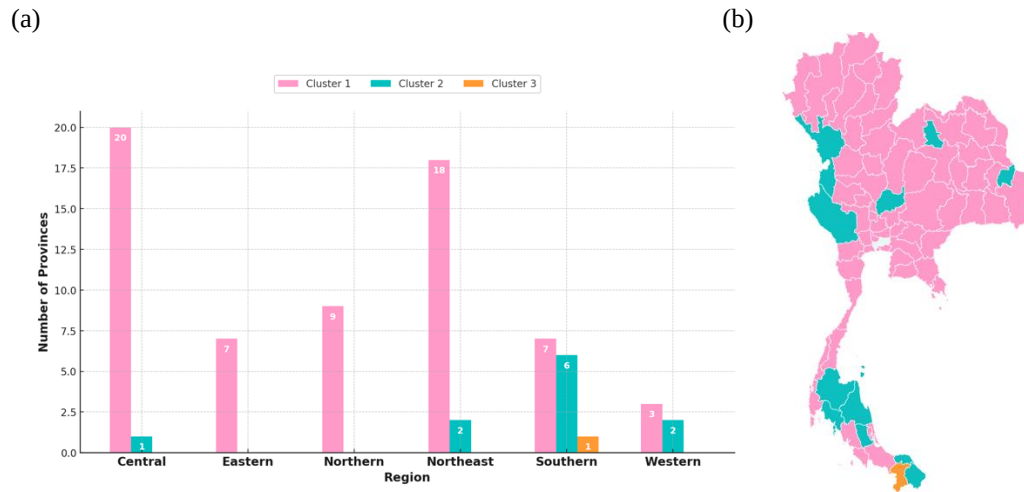


Figure 4 Summary of data clustering to TPMAPEDU1 as the number of provinces for each cluster in each sub-region (a) and the geographic clustering (b).

Figure 4 shows the clustering of Thai provinces based on TPMAPEDU1 using the hybrid KM–DPC algorithm, resulting in three clusters as Cluster 1 (64 provinces) with low scores, Cluster 2 (11) with higher scores, and Cluster 3 (1) as an outlier. Figure 4a shows Cluster 1 spread across all regions, especially in Central and Northeastern Thailand, while Clusters 2 and 3 are more localized. The spatial map (Figure 4b) reveals regional disparities in education, highlighting the need for targeted interventions.

3.4 Income Datasets

Economic vulnerability reflects short-term poverty, particularly in households with unstable income sources. Unlike structural factors such as education or housing, income directly indicates a household's current ability to sustain daily living. This study used three TPMAP indicators as low-income households (TPMAPINC1), unemployed individuals (TPMAPINC2), and households without stable income (TPMAPINC3), which revealed financial fragility, labor exclusion, and income insecurity. The hybrid clustering algorithm classified provinces by economic stress to support timely and targeted policy responses. TPMAPINC1 is used as a case example, while the remaining indicators are summarized by average Silhouette Scores in the following section.

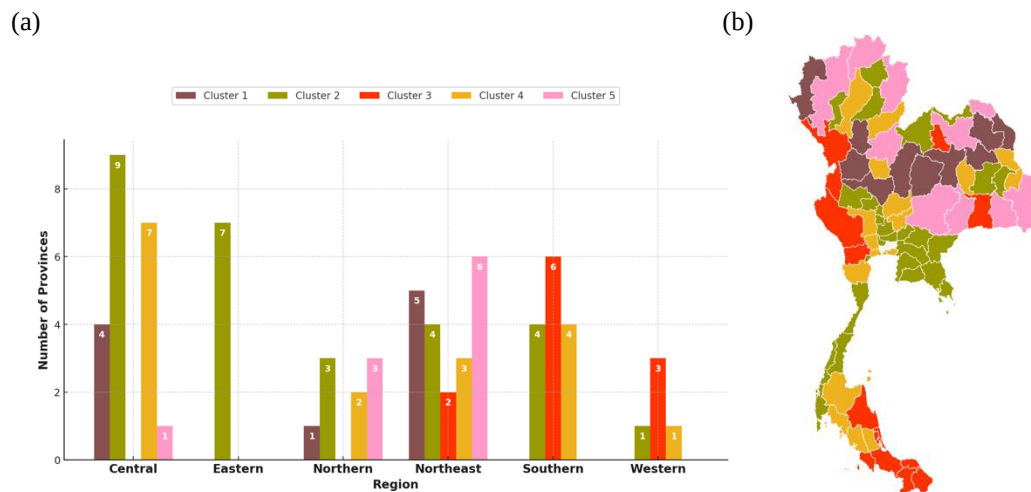


Figure 5 Summary of data clustering to TPMAPINC1 as the number of provinces for each cluster in each sub-region (a) and the geographic clustering (b).

Figure 5 shows Thai provinces grouped into five clusters based on TPMAPINC1 income using a hybrid KM and DPC algorithm. Cluster 1 comprised 10 provinces, Cluster 2 had 28 with low income across regions, Cluster 3 had 11 with high income but some poverty, Cluster 4 had 17, and Cluster 5 had 10. Figure 5a maps the clusters,

with Cluster 2 mainly in Central and Eastern Thailand, and the other clusters in limited areas. Figure 5b highlights regional income disparities, emphasizing the need for area-specific policies.

3.5 Access to Basic Services Datasets

Equitable access to public services is crucial for social inclusion, as limited transportation and internet access deepen inequality. This study used two TPMAPI indicators as households without public transportation (TPMAPACC1) and households without internet access (TPMAPACC2) to examine spatial and digital exclusion. The hybrid clustering algorithm identified regional patterns to guide infrastructure and digital policy. This section presents TMAPACC1 as a case example, while TMAPACC2 is summarized by its average Silhouette Score in the next section.

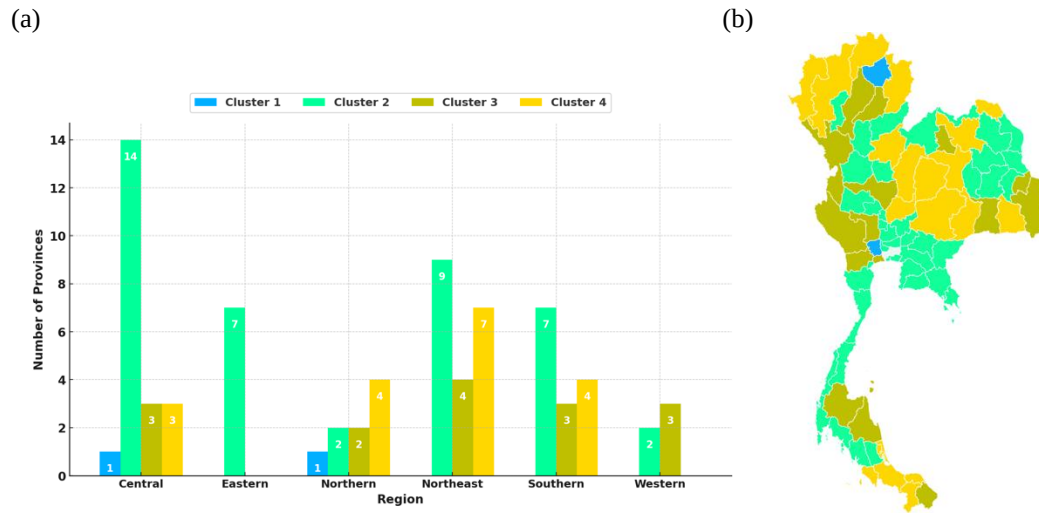


Figure 6 Summary of data clustering to TMAPACC1 as the number of provinces for each cluster in each sub-region (a) and the geographic clustering (b).

Figure 6 shows Thai provinces grouped into four clusters based on the TMAPACC1 indicator, which measures access to basic services using a hybrid KM–DPC algorithm. Cluster 2, with 41 provinces, spans a wide range of poverty levels, mainly in Central and Northeastern Thailand. Cluster 4, common in Northeastern and Southern Thailand, highlights geographic disparities in access, emphasizing the need for targeted policies to address inequalities.

3.6 Average Silhouette Scores

To demonstrate the effectiveness of the proposed clustering approach, this section presents a comparison of the average Silhouette Scores obtained from three algorithms as traditional KM, DPC, and the hybrid KM–DPC method. The Silhouette Score is widely used as an internal validation metric that reflects both intra-cluster cohesion and inter-cluster separation. Higher values indicate clearer and more well-defined clusters. Table 1 summarizes the average Silhouette Scores across all the indicators, highlighting the comparative performance of each algorithm.

Table 1 Comparison of average Silhouette Scores for KM, DPC, and the hybrid KM–DPC clustering algorithm.

Indicator	Average Silhouette Scores		
	KM	DPC	Hybrid KM–DPC
TPMAPHAL1	0.9597	0.9306	0.9597
TPMAPHAL2	0.8977	0.8575	0.9089
TPMAPHAL3	0.6796	0.7291	0.7582
TPMAPHAL4	0.6682	0.4772	0.6642
TPMAPLIV1	0.9562	0.9562	0.9562
TPMAPLIV2	0.9917	0.848	0.9917
TPMAPLIV3	0.9019	0.8599	0.9019
TPMAPLIV4	0.6528	0.6037	0.6528
TPMAPEDU1	0.8458	0.8049	0.8551
TPMAPEDU2	0.7239	0.6962	0.7162
TPMAPEDU3	0.6783	0.6667	0.6783
TPMAPEDU4	0.7355	0.8993	0.9292
TPMAPINC1	0.6636	0.5929	0.7484
TPMAPINC2	0.5858	0.6174	0.6196
TPMAPINC3	0.6868	0.5649	0.6868
TPMAPACC1	0.618	0.856	0.8432
TPMAPACC2	0.6531	0.2633	0.6531

Table 1 presents the average Silhouette Scores from clustering 17 multidimensional poverty indicators using the KM, DPC, and the hybrid KM–DPC algorithms. The results highlight the outstanding performance of the hybrid algorithm, which achieved the highest scores in 9 out of 17 indicators, particularly in the dimensions of Health (TPMAPHAL2, TPMAPHAL3), Education (TPMAPEDU1, TMAPEDU4), and Income (TPMAPINC1), demonstrating its ability to effectively handle complex and heterogeneous data structures. The hybrid method matched the best performance of KM in several indicators such as TPMAPHAL1, TMAPLIV1, and TMAPLIV2, indicating its robustness. By contrast, the standalone DPC outperformed the others in only a few cases, such as TPMAPACC1. Overall, the hybrid KM–DPC algorithm successfully integrated the strengths of both KM and DPC, delivering higher clustering accuracy as a highly suitable approach for in-depth, multidimensional poverty analysis at the provincial level.

4. Conclusions

This study introduced a hybrid clustering algorithm combining KM and DPC to analyze multidimensional poverty in Thailand using TPMAP data. This method improved cluster quality using DPC to initialize KM centroids. Evaluation across 17 indicators using Silhouette Scores confirmed the superior performance of the hybrid algorithm over traditional methods. Analysis across five poverty dimensions revealed spatial disparities and highlighted provinces with either broad or specific deprivations. Our findings support the hybrid model as a valuable tool for informing targeted and equitable policy interventions.

5. Ethical Approval

This study did not involve human participants or animals; therefore, ethical approval was not required.

6. Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The authors used Grammarly to enhance the clarity and grammar of the manuscript. All content was critically reviewed and approved by the authors, who bear full responsibility for the integrity, accuracy, and originality of this study.

7. Acknowledgements

The authors gratefully acknowledge the Thai People Map and Analytics Platform (TPMAP) for providing the multidimensional poverty data used in this study. The TPMAP was developed through the collaboration of the Office of the National Economic and Social Development Council (NESDC) and the National Electronics and Computer Technology Center (NECTEC), Thailand. The availability of this data greatly contributed to the analysis of multidimensional poverty patterns presented in this research.

8. Author Contributions

Narintra Mingolo contributed to conceptualization, methodology development, software implementation, data curation, formal analysis, investigation, visualization, and writing of the original draft. Weerachai Sarakorn

contributed to supervision, validation of the results, critical review and editing of the manuscript, and overall project administration.

9. Conflicts of interest

The authors declare no conflicts of interest.

10. References

- [1] Alkire S, Foster J. Counting and multidimensional poverty measurement. *J Public Econ*. 2011;95(7–8):476–487.
- [2] United Nations Development Programme (UNDP) and Oxford Poverty and Human Development Initiative (OPHI). *Global Multidimensional Poverty Index 2023: Unstacking global poverty*. New York: UNDP; 2023.
- [3] The Office of National Economic and Social Development Board (NESDB) and National Electronics and Computer Technology Center (NECTEC), National Science and Technology Development Agency (NSTDA), Ministry of Science and Technologies. *The Thai People Map and Analytics Platform (TPMAP)*. <https://www.tpmmap.in.th>.
- [4] Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recognit Lett*. 2010;31(8):651–666.
- [5] Sarker IH. Machine learning: Algorithms, real-world applications and research directions. *SN Comput Sci*. 2021;2(3):160.
- [6] Xu D, Tian Y. A comprehensive survey of clustering algorithms. *Ann Data Sci*. 2015;2(2):165–193.
- [7] Basu S, Banerjee A, Mooney RJ. Semi-supervised clustering by seeding. In: *Proceedings of the 19th International Conference on Machine Learning (ICML 2002)*; 2002. p. 19–26.
- [8] MacQueen J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley (CA): University of California Press; 1967. p. 281–297.
- [9] Rodriguez A, Laio A. Clustering by fast search and find of density peaks. *Science*. 2014;344(6191):1492–1496.
- [10] Sun L, Qin X, Ding W, Xu J, Zhang S. Density peaks clustering based on k-nearest neighbors and self-recommendation. *Int J Mach Learn Cybern*. 2021;12(7):1913–1938.
- [11] Jean N, Burke M, Xie M, Alampay Davis WM, Lobell DB, Ermon S. Combining satellite imagery and machine learning to predict poverty. *Science*. 2016;353(6301):790–794.
- [12] Rahman MA, Alam MR, Hossain MA. A clustering approach to identify multidimensional poverty using K-means with cosine similarity. *PLOS ONE*. 2021;16(7):e0255312.
- [13] Zhang M, Ma Y, Li J, Zhang J. A density connection weight-based clustering approach for dataset with density-sparse region. *Expert Syst Appl*. 2023;230:120633.
- [14] Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans Pattern Anal Mach Intell*. 2002;24(7):881–892.
- [15] Bradley PS, Bennett KP, Demiriz A. Constrained K-means clustering. Technical Report MSR-TR-2000-65. Redmond (WA): Microsoft Research; 2000.
- [16] Bradley PS, Fayyad UM. Refining initial points for K-means clustering. In: *Proceedings of the Fifteenth International Conference on Machine Learning*; 1998 Jul 24–27; Madison, WI. p. 91–99.
- [17] Sinaga KP, Yang M-S. Unsupervised K-means clustering algorithm. *IEEE Access*. 2020;8:80716–80727.
- [18] Cao L, Liu Y, Wang D, Wang T, Fu C. A novel density peak fuzzy clustering algorithm for moving vehicles using traffic radar. *Electronics*. 2020;9(1):46.
- [19] Wang Y, Wei Z, Yang J. Feature trend extraction and adaptive density peaks search for intelligent fault diagnosis of machines. *IEEE Trans Ind Inform*. 2019;15(1):105–115.
- [20] Tian J, Zhao J, Zheng C. Clustering of cancer data based on Stiefel manifold for multiple views. *BMC Bioinformatics*. 2021;22:268.