Data-Driven Spatial Typologies of Graduate Employability: Unsupervised Learning Analysis of the Malaysian Labor Market

Saiful Naim Sulaiman[1*], Mohd Fariduddin Mukhtar[2], Widodo W.S.[1], Mohamed Saiful Firdaus Hussin[1]

[1]Faculty of Mechanical Technology and Engineering (FTKM), Universiti Teknikal Malaysia Melaka (UTeM), Durian Tunggal, Melaka, Malaysia

[2]Faculty of Technology Management and Technopreneurship (FPTT), Universiti Teknikal Malaysia Melaka (UTeM), Durian Tunggal, Melaka, Malaysia

DOI: https://doi.org/10.47772/IJRISS.2026.101000572

Received: 27 October 2026; Accepted: 02 November 2026; Published: 10 November 2026

ABSTRACT

National unemployment statistics in Malaysia show overall macroeconomic stability but hide severe localized labor imbalances and educational mismatches. This study investigates this spatial disequilibrium by deploying a multiscale analytical framework to cluster micro-spatial labor markets and evaluate their structural effect on macro-regional graduate employability. An unsupervised K-Means machine learning algorithm was applied as a distance-based multidimensional partitioning tool to cluster a longitudinal constituency-level dataset ($N = 613$) from the Department of Statistics Malaysia (2021–2023). The bivariate attribute-space features—Unemployment Rate and Labor Force Participation Rate—were standardized via Z-score normalization. Validation metrics (Elbow Method and Silhouette Score) alongside sensitivity comparisons against Agglomerative Hierarchical Clustering and Gaussian Mixture Models identified a moderate partition of $K = 3$ socio-economic cluster typologies across a continuous macroeconomic gradient: Traditional/Informal Economies (Cluster 0), Dynamic Industrial Hubs (Cluster 1), and Vulnerable/Frictional Zones (Cluster 2). Longitudinal tracking reveals a major post-pandemic recovery, in which the Vulnerable tier recorded a net contraction of 100 constituencies alongside a corresponding expansion in Dynamic Industrial Hubs by 2023. Cross-analyzing these clusters against state-level graduate unemployment via the Diploma Penalty Gap (DPG) reveals three critical structural dynamics. First, metropolitan knowledge hubs within Cluster 1 (Selangor, Pulau Pinang, Kuala Lumpur) exhibit a positive DPG (+1.0% to +1.7%), consistent with potential credential inflation, in which degree holders may compete with Technical and Vocational Education and Training (TVET) diploma graduates. Second, secondary manufacturing states within Cluster 1 (Johor, Melaka) and Traditional Economies in Cluster 0 (Perak, Kedah) exhibit negative DPGs (-0.6% to -3.0%), suggesting a structural divergence between local industrial skill demand and tertiary qualifications, alongside contextually observed out-migration pressures. Correlation analyses ($R = 0.506$, $p = 0.045$, $N = 16$) confirm a significant association between regional employment capacity and educational mismatch, highlighting the necessity for spatially targeted, demand-side industrial policies to complement workforce development.

Keywords: Graduate Employability, Unsupervised Machine Learning, Educational Mismatch.

INTRODUCTION

The transition of graduates into the labor market remains a major challenge for developing economies, as university enrollments increase the supply of highly qualified labor while industrial capacity fails to expand at an adequate rate (Gooptu & Mukherjee, 2023; Yen et al., 2023). As a result, degree and diploma holders face high youth unemployment, underemployment, and skill underutilization (Omar et al., 2012). In the post-pandemic economy, digital transformation and changing skill requirements have worsened these issues, leaving many graduates jobless or underutilized shortly after graduation (Zaid & Kamin, 2026).

Analyzing the graduate labor market requires examining "Spatial Mismatch" and the "Degree Penalty". Spatial Mismatch occurs when the physical distance between residential areas and job centers limits employment opportunities, causing a localized imbalance in labor supply and demand that heavily affects diploma holders who lack mobility (Eom, 2022; Jin et al., 2022). Concurrently, the rapid increase in tertiary education creates an educational mismatch that forces university graduates to take lower-level jobs (Capsada-Munsech & Ortiz-Gervasi, 2025; Chuang & Liang, 2022). This oversupply results in a "degree penalty" characterized by lower earnings and productivity, while simultaneously creating a crowding-out effect that pushes TVET and diploma graduates into unstable employment (Capsada-Munsech & Ortiz-Gervasi, 2025; Ma et al., 2022; Tobler & Leesch, 2026).

Despite these labor market issues, traditional policies and studies typically rely on national or state-level aggregate statistics, which often hide localized problems such as regional stagnation, poverty traps, and cross-border human capital flight or "brain drain" (Abdila et al., 2025; Jauhar et al., 2015; Nistor, 2009). Furthermore, standard econometric methods often fail to capture multidimensional structural disparities and complex socio-economic distributional patterns across geographic units (Anselin, 2010). As a result, uniform national policies fail to address the unique conditions of regional labor markets, leading to misallocated investments and persistent inequality (Asadullah et al., 2023).

To address these gaps, this study applies Unsupervised Machine Learning—specifically K-Means clustering—on constituency-level macroeconomic and labor market data in Malaysia. As an unconstrained, distance-based partitioning tool, this method was deliberately selected over spatially constrained models to group demographically similar regions regardless of physical adjacency. This data-driven method groups fine-grained administrative electoral constituencies into distinct socio-economic typologies based strictly on attribute-space similarities, thereby eliminating state-level label bias.

The primary objectives of this study are to track the longitudinal changes in spatial graduate employability clusters from 2021 to 2023, and to quantify the localized "Diploma vs. Degree Penalty" across different regions. Ultimately, by identifying distinct spatial clusters ranging from metropolitan hubs to stagnant regions, this study provides evidence for targeted regional educational policies, labor market interventions, and workforce alignment.

LITERATURE REVIEW

Spatial Heterogeneity and Regional Labor Markets

Regional labor market performance depends heavily on location, geographic dependencies, local industry structures, and spatial non-stationarity (Kosfeld & Dreger, 2006; Rusanovskiy & Markov, 2016). While classical economic models traditionally assume spatial neutrality, these models fail because geography directly dictates employment access and wage outcomes (Filiztekin, 2009; Simeonov, 2024). Consequently, studies show significant positive spatial autocorrelation in unemployment distributions, where high-unemployment areas cluster next to other depressed regions and low-unemployment areas form prosperous spatial clusters (Filiztekin, 2009; Ogryzek & Jaskulski, 2025). These spatial divides are often intensified during macroeconomic shocks, as post-pandemic recoveries favor core agglomerations with high structural adaptability while peripheral zones remain in persistent stagnation (Brada et al., 2021).

The Spatial Mismatch Hypothesis explains this geographical polarization, stating that physical distance and transport friction between residential areas and industrial job centers create severe barriers for job seekers (Eom, 2022; Jin et al., 2022). For instance, in suburban and peripheral areas, poor public transit and long commutes isolate lower-income workers and new graduates from high-density urban jobs (Eom, 2022; Jin et al., 2022). Furthermore, studies using Exploratory Spatial Data Analysis (ESDA) and Local Indicators of Spatial Association (LISA) confirm that spatial heterogeneity operates at sub-national levels (Abdila et al., 2025; Ogryzek & Jaskulski, 2025). As a result, macro-level aggregation, such as state or national reporting, hides localized unemployment "hot spots" and spatial poverty traps, rendering uniform central government policies ineffective across heterogeneous regional labor markets (Abdila et al., 2025; Simeonov, 2024).

Educational Over-qualification and the "Degree Penalty"

The global expansion of higher education has created a structural imbalance between the supply of tertiary graduates and the absorptive capacity of industrial economies (Gooptu & Mukherjee, 2023). Theoretical models, specifically the Overeducation, Required, Under-education (ORU) framework and Nash bargaining formulations, demonstrate that when educated labor supply exceeds industrial demand, job-matching mechanisms break down. Consequently, overqualified workers face substantial wage penalties compared to correctly matched peers (Sattinger & Hartog, 2013). Empirical data confirms that this overeducation causes persistent earning deficits and reduces job satisfaction, as university degree holders are forced into occupations that do not utilize their specialized training (Chuang & Liang, 2022; Ma et al., 2022).

Furthermore, this structural oversupply creates a "crowding-out" effect within labor queues (Capsada-Munsech & Ortiz-Gervasi, 2025; Tobler & Leesch, 2026). When university degree holders accept semi-skilled or non-tertiary roles, they displace intermediate workers, such as Technical and Vocational Education and Training (TVET) diploma holders, into lower-tier employment (Capsada-Munsech & Ortiz-Gervasi, 2025). In emerging industrial economies such as Malaysia, this mismatch is worsened by rigid curricula and industry skill gaps (Hasnan et al., 2019; Rodzalan et al., 2022; Zaid & Kamin, 2026). Specifically, industrial manufacturing sectors prioritize immediate operational competencies, digital skills, and hands-on vocational capabilities over general academic degrees (Zaid & Kamin, 2026). As a result, university degree holders frequently face a localized "degree penalty" characterized by prolonged job searches and wage discounts. Conversely, diploma and TVET graduates experience faster employment absorption because their skills align more closely with mid-level industrial technical requirements (Hasnan et al., 2019; Zulkefli & Saad, 2026).

Human Capital Flight and Peripheral Stagnation

Persistent regional inequality drives a continuous cycle of peripheral stagnation and human capital flight (Asadullah et al., 2023; Takahashi et al., 2013). Industrial capital naturally concentrates in core metropolitan regions that offer superior infrastructure and dense market access, effectively leaving peripheral areas economically isolated (Takahashi et al., 2013). Consequently, highly skilled workers and young graduates migrate from lagging states to core economic centers to escape spatial poverty traps and depressed wages. This out-migration, broadly known as regional brain drain, depletes local human capital, reduces the returns on public educational investments, and further discourages private capital inflow (Asadullah et al., 2023; Jauhar et al., 2015; Nistor, 2009)

Furthermore, even within highly industrialized zones, spatial proximity to high-wage international hubs creates unique labor market distortions for university graduates. For example, Rizzo and Glasson (2011) analyzed the Strait Transnational Urban Region (STUR) between Singapore and Southern Johor, Malaysia, where 30,000 to 50,000 workers commute across the Johor Strait daily to access higher wages. Although this cross-border mobility generates foreign currency remittances, it causes severe structural asymmetries in the local Malaysian labor market. Specifically, local industrial sectors struggle to retain skilled talent, public infrastructure experiences extreme congestion, and local firms face significant wage pressures driven by their proximity to a global financial center (Rizzo & Glasson, 2011).

Machine Learning Applications in Socio-Economic Policy

To capture complex, non-linear socio-economic interactions across regions, researchers increasingly use unsupervised machine learning algorithms (Ashmarina & Vochozka, 2019; Mumtaz & Whiteford, 2021). Traditional spatial econometric models rely on linear assumptions and rigid administrative boundaries that fail to reflect functional labor market limits (Kosfeld & Dreger, 2019; Ogryzek & Jaskulski, 2025). Conversely, unsupervised partitioning methods provide robust, data-driven alternatives. These methods are fundamentally divided into spatially constrained algorithms (e.g., SKATER), which strictly enforce physical geographic contiguity, and purely attribute-space algorithms (e.g., K-Means, K-Medoids, Fuzzy C-Means), which prioritize structural cohesion. This allows researchers to group demographically similar regions regardless of physical distance.

Recent applications demonstrate that machine learning models significantly improve targeting accuracy in regional development planning (Mumtaz & Whiteford, 2021). Furthermore, researchers can determine optimal cluster counts and validate spatial partition robustness by evaluating internal cluster validity indices, such as the Davies-Bouldin Index (DBI) and Silhouette scores (Idhom et al., 2025). Advanced frameworks also incorporate feature filtering and hyperparameter optimization (e.g., AutoRFKM) to ensure algorithmic fairness across regional classifications (Andena et al., 2025).

Despite these algorithmic advances, a critical empirical gap remains in addressing the Modifiable Areal Unit Problem (MAUP) when linking granular socio-economic models to broader educational metrics. Existing literature rarely bridges unsupervised K-Means clustering conducted at a fine-grained electoral constituency level with macro-level longitudinal educational mismatches. Consequently, this study adopts a multiscale operational compromise, driven by data availability, and explicitly cross-analyzes Euclidean K-Means clusters of highly granular Malaysian constituency data with macro-level educational metrics. While this introduces inherent constraints on ecological inference, this methodology establishes a necessary longitudinal framework to track structural cluster trajectories and uses state-level macroeconomic data to evaluate localized diploma versus degree penalties.

METHODOLOGY

The overall analytical pipeline for this study is structured into four sequential phases: data acquisition, unsupervised clustering, statistical validation, and post hoc educational mismatch analysis, as illustrated in Figure 1.

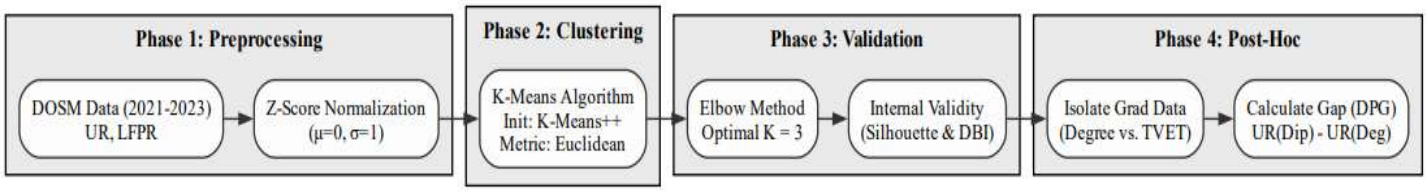

Figure 1: End-to-end methodological flowchart detailing the data acquisition, spatial clustering, and post-hoc educational mismatch analysis pipeline

Data Acquisition and Preprocessing

To capture localized structural variations in labor market performance, this study compiled highly granular subdistrict macroeconomic data spanning 613 electoral constituencies (comprising 600 state legislative assembly seats and 13 federal territory parliamentary seats) from 2021 to 2023. The dataset was obtained from official statistical repositories published by the Department of Statistics Malaysia (DOSM). The initial primary indicators selected for empirical modeling comprise three key macroeconomic features: (1) the Unemployment Rate (UR_i), measuring open labor market slack; (2) the Labor Force Participation Rate ($LFPR_i$), capturing overall economic activity and workforce engagement; and (3) the Employment-to-Population Ratio (EP_i), reflecting the economy's capacity to generate gainful employment relative to the working-age population. Although the raw DOSM repository also reports the Employment-to-Population Ratio (EP_i), a preliminary data integrity and sensitivity audit revealed severe clerical anomalies in the raw EP_i series alongside strict mathematical collinearity with $LFPR_i$ and UR_i ($EP_i = LFPR_i \times [1 - UR_i]$). Consequently, EP_i was excluded from the active K-Means feature space to prevent distance distortion, though mathematically verified true EP_i values are retained in post-hoc reporting for macroeconomic context.

Because raw macroeconomic features exhibit varying numerical ranges and distributional variances, an unscaled feature space can cause distance-based clustering algorithms to bias centroid placement toward variables with larger absolute magnitudes (Hemavathy & Chinnadurai, 2022). To prevent artificial feature weighting and ensure equal geometric contribution during spatial partitioning, all variables were standardized using Z-score normalization prior to algorithm execution:

$$Z_{i,j} = \frac{X_{i,j} - \mu_j}{\sigma_j}$$

where $X_{i,j}$ represents the raw observed value of feature j for constituency i, μ_j is the arithmetic mean of feature j, and σ_j is the corresponding standard deviation across all observations. This transformation projects all inputs onto a standardized scale with a mean of zero ($\mu = 0$) and unit variance ($\sigma^2 = 1$), establishing an unbiased feature space for multidimensional distance calculations (Hemavathy & Chinnadurai, 2022).

Unsupervised Attribute-Space Clustering Algorithm

To partition regional labor markets based on multi-attribute socio-economic similarity, this study deployed K-Means clustering, an unsupervised partitioning technique widely validated in socio-economic and regional economic policy literature (Ashmarina & Vochozka, 2019; Mumtaz & Whiteford, 2021; Reza et al., 2025). Traditional regional labor analyses rely heavily on rigid, predefined administrative boundaries (e.g., states or districts) that fail to capture functional economic zones or cross-boundary labor flows (Ogryzek & Jaskulski, 2025). Unsupervised attribute-space clustering overcomes these administrative limitations by grouping geographic units solely on intrinsic data distributions, thereby revealing data-driven socio-economic typologies and hyper-local structural anomalies without a priori label bias (Andena et al., 2025; Mumtaz & Whiteford, 2021).

Mathematically, the K-Means algorithm partitions N observations across D-dimensional space into K distinct, non-overlapping clusters $\boldsymbol{C} = \{C_1, C_2, \dots, C_K\}$. By utilizing standard Euclidean distance, this approach treats observations as independent points in attribute space ($\mathbb{R}^2$), explicitly prioritizing structural economic cohesion over geographic adjacency constraints.

The objective function aims to minimize the within-cluster sum of squares (WCSS), defined as the sum of squared Euclidean distances between each data point $\boldsymbol{z}_i$ and its assigned cluster centroid $\boldsymbol{\mu}_k$:

$$\underset{\boldsymbol{C}}{argmin} \sum_{k=1}^{K} \sum_{\boldsymbol{z}_i \in C_k} \| \boldsymbol{z}_i - \boldsymbol{\mu}_k \|^2$$

where $\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{\boldsymbol{z}_i \in C_k} \boldsymbol{z}_i$ denotes the mean vector of all observations belonging to cluster C_k. Centroids were initialized using the K-Means++ algorithm to accelerate convergence and avoid sub-optimal local minima (Reza et al., 2025).

Cluster Optimization and Validation

Determining the appropriate number of clusters (K) is critical to balance model parsimony with spatial granularity. In this study, the optimal partition count was established using the Elbow Method by evaluating the reduction in Within-Cluster Sum of Squares (WCSS) across candidate values ranging from $K = 2$ to $K = 8$. The inflection point ("elbow") in the WCSS trajectory, corroborated simultaneously by the peak Silhouette Score (0.3934), identified $K = 3$ as the optimal partition parameter (see Figure 2). This configuration demonstrates a moderate, yet structurally stable geometric separation that effectively balances structural cohesion with policy interpretability.

To rigorously validate the separation and internal compactness of the $K = 3$ spatial typologies, quantitative validation metrics were evaluated in accordance with benchmark machine learning protocols. Specifically, the Silhouette Coefficient (S_i) and the Davies-Bouldin Index (DBI) were calculated. The Silhouette Coefficient measures individual point cohesion relative to neighbor separation:

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

where a_i is the mean intra-cluster distance for point i, and b_i is the mean nearest-cluster distance. Concurrently, the Davies-Bouldin Index evaluates the ratio of intra-cluster dispersion to inter-cluster separation:

$$\mathrm{DBI} = \frac{1}{\mathrm{K}}\sum_{\mathrm{k}=1}^{\mathrm{K}} \max_{\mathrm{l}\neq\mathrm{k}} \left(\frac{\mathrm{s_k} + \mathrm{s_l}}{\mathrm{d}(\boldsymbol{\mu}_\mathrm{k}, \boldsymbol{\mu}_\mathrm{l})}\right)$$

where $\mathrm{s_k}$ is the average distance of points in cluster k to centroid $\boldsymbol{\mu}_\mathrm{k}$, and $\mathrm{d}(\boldsymbol{\mu}_\mathrm{k}, \boldsymbol{\mu}_\mathrm{l})$ is the distance between centroids. Empirical simulation studies confirm that high mean Silhouette scores, combined with minimized DBI values, indicate cluster stability and statistical separation (Idhom et al., 2025). For the selected $K = 3$ partition, the DBI converged at an optimized value of 0.8531, mathematically confirming sufficient inter-cluster separation and intra-cluster compactness across the feature space.

Furthermore, to evaluate algorithmic sensitivity against alternative topological structures, the K-Means partition was compared against Gaussian Mixture Models (GMM) and Agglomerative Hierarchical Clustering. The Adjusted Rand Index (ARI) yielded moderate agreement scores (0.53 and 0.66, respectively), accurately reflecting the continuous, overlapping nature of the macroeconomic data gradient. Despite these expected boundary variations across different clustering algorithms, K-Means was retained as the primary operational model and validated across 15 random initializations using K-Means++ to ensure stability against sub-optimal local minima. Its variance-minimizing objective function generates the most structurally balanced and interpretable spatial typologies required for actionable regional labor policy.

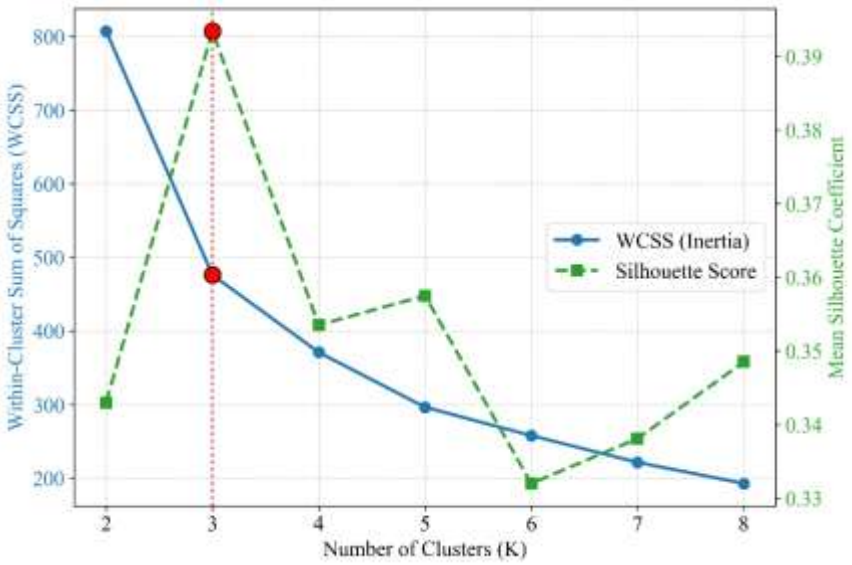

Figure 2: Dual-axis cluster validation indicating an optimal partition at $K = 3$. The primary axis demonstrates the WCSS inflection point (Elbow Method), while the secondary axis shows the relative peak in partition separation (Peak Silhouette Score = 0.3934) along the continuous feature space.

Post-Hoc Educational Mismatch Analysis

To bridge macroeconomic attribute clustering and educational policy and workforce development, a secondary analytical phase was conducted to evaluate localized educational mismatches across subnational jurisdictions. To operationalize educational policy analysis despite the Modifiable Areal Unit Problem (MAUP), this study implemented a multiscale analytical framework. While K-Means established micro-spatial macroeconomic typologies at the constituency level, state-level indices were utilized as a macro-regional context vector. Although classifying a state under a single dominant cluster masks intra-state variance and introduces ecological aggregation bias, this compromise is mathematically necessary to assess qualification-specific structural friction. Therefore, utilizing baseline graduate employment data from the Graduates Statistics 2024

report (Department of Statistics Malaysia, 2025), a longitudinal "Diploma Penalty Gap" ($DPG_{s,t}$) was formulated for each state s across the four-year observation window ($t \in \{2020,2021,2022,2023\}$):

$$DPG_{s,t} = UR_{s,t}^{\text{Diploma/TVET}} - UR_{s,t}^{\text{Bachelor's Degree}}$$

where $UR_{s,t}^{\text{Diploma/TVET}}$ denotes the open unemployment rate among sub-degree and TVET diploma holders, and $UR_{s,t}^{\text{Bachelor's Degree}}$ represents the unemployment rate among graduates with tertiary degrees.

The mathematical sign and magnitude of the $DPG_{s,t}$ metric provide explicit diagnostic insights into regional labor absorption capacity. A positive value ($DPG_{s,t} > 0$) indicates a labor market that penalizes sub-degree qualifications, where diploma and TVET graduates experience higher relative unemployment due to employer credentialism or structural skill deficits. Conversely, a negative value ($DPG_{s,t} < 0$) signifies a structural "Degree Penalty," wherein university degree holders face higher unemployment rates than diploma holders. This negative gap reflects regional industrial stagnation and an inability to absorb highly educated human capital, leading to overeducation, underemployment, and graduate crowding-out dynamics in lagging economic zones (Capsada-Munsech & Ortiz-Gervasi, 2025; Chuang & Liang, 2022).

RESULTS AND DISCUSSION

Multi-Dimensional Profiles of Spatial Typologies ($K = 3$)

The unsupervised K-Means clustering algorithm partitioned the 613 parliamentary and state constituencies into three distinct macroeconomic spatial typologies. The final cluster centroids, representing the mean socio-economic signature of each typology for the year 2023, are detailed in Table 1.

Table 1: Final Cluster Profiles and Centroid Coordinates ($K = 3$)

Cluster Typology	Unemployment Rate (%)	Participation Rate (%)	True EP Ratio (%)	Constituency Count
0 (Traditional/Informal Economies)	3.11	62.09	60.16	243
1 (Dynamic Industrial Hubs)	2.48	72.57	70.77	256
2 (Vulnerable/Frictional Zones)	6.95	68.50	63.74	114

The spatial partitioning of these constituencies across the two-dimensional feature space is visualized in Figure 3. Consistent with the Silhouette Score of 0.3934, the visualization illustrates a moderate, yet structurally stable geometric separation between the three typologies. Rather than displaying artificially isolated clusters, the plot accurately reflects a continuous macroeconomic spectrum, in which the algorithm establishes optimal structural boundaries across overlapping regional labor gradients.

Based on these coordinates, each cluster exhibits a unique structural reality (Figure 4). Cluster 1 represents the "Dynamic Industrial Hubs" core, characterized by the lowest unemployment rate (2.48%) and the highest capacity for labor absorption (a true EP ratio of 70.77%). Conversely, Cluster 2 is classified as "Vulnerable/Frictional Zones," suffering severe labor market friction and a disproportionately high unemployment rate (6.95%) despite robust workforce participation (68.50%). Additionally, Cluster 0 represents "Traditional/Informal Economies," which exhibit the lowest labor force participation (62.09%)

alongside moderate unemployment (3.11%), indicating a high prevalence of informal sectors or populations opting out of the formal labor market entirely.

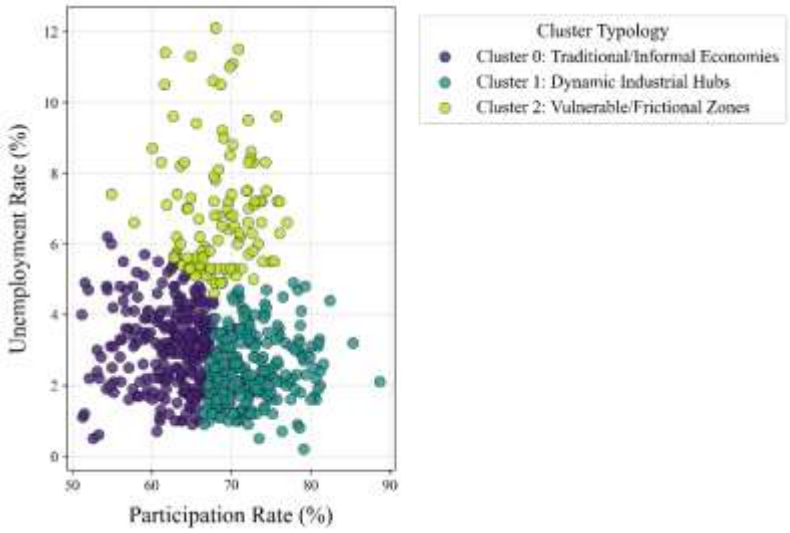

Figure 3: Scatter plot visualizing the K-Means spatial clustering (*K*=3) of constituencies based on 2023 macroeconomic indicators.

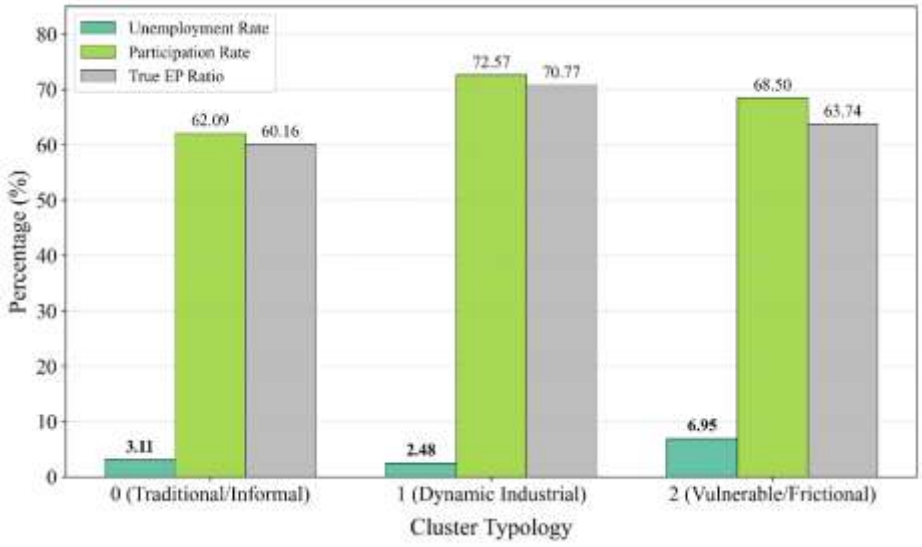

Figure 4: Average socio-economic indicators across the three identified spatial typologies (year 2023).

Geographical Distribution and Post-Pandemic Recovery

Evaluating the longitudinal cluster assignments from 2021 to 2023 reveals the trajectory of regional post-pandemic recovery and structural realignment, as detailed in Table 2. For spatial clarity, the geographic distribution highlights only the top three dominant states per cluster, representing the highest concentration of constituencies rather than an exhaustive national inventory.

Table 2: Longitudinal Cluster Evolution (2021–2023) and Primary Geographic Distribution

Cluster Typology	2021	2022	2023	Net Shift	Dominant States (2023 Constituency Count)
0 (Traditional/Informal)	239	264	243	+4	Kelantan (43), Perak (32), Terengganu (30)

1 (Dynamic Industrial)	160	213	256	+96	Selangor (55), Sarawak (45), Johor (43)
2 (Vulnerable/Frictional)	214	136	114	-100	Sabah (69), Perak (15), Sarawak (14)

The Vulnerable/Frictional typology (Cluster 2) experienced a rapid and substantial contraction, declining from 214 constituencies during the pandemic's baseline in 2021 to 114 by 2023 (a net shift of -100). Spatially, these remaining vulnerable jurisdictions are predominantly concentrated in East Malaysia (Sabah, 69 constituencies) alongside peripheral pockets in Perak (15 constituencies) and Sarawak (14 constituencies). Concurrently, the Dynamic Industrial Hubs (Cluster 1) demonstrated robust expansion, scaling from 160 to 256 constituencies (+96 net shift). This growth is anchored by core economic powerhouses, namely Selangor (55 constituencies), Sarawak (45 constituencies), and Johor (43 constituencies), capturing active industrial absorption capacity following the removal of data-artifact constraints. Meanwhile, the Traditional/Informal Economies (Cluster 0) exhibited remarkable structural stability throughout the observation window, shifting marginally from 239 to 243 constituencies (+4 net shift). These traditional jurisdictions remain heavily entrenched along the East Coast and northern peninsular perimeter, specifically in Kelantan (43 constituencies), Perak (32 constituencies), and Terengganu (30 constituencies), reflecting enduring structural insulation from macro-industrial recovery cycles.

The aggregate net reallocation of constituencies from the Vulnerable tier to the Dynamic Industrial typology reflects a centralized, infrastructure-driven macroeconomic recovery. Unlike classical models that predict uniform national rebounds, this trajectory aligns with Brada et al. (2021), who establish that regional recoveries are heavily dictated by structural adaptability and positive spatial spillovers that cluster high-performing regions geographically. In the Malaysian context, post-pandemic labor absorption was aggressively captured by states with pre-existing industrial ecosystems and capital infrastructure, notably Selangor, Sarawak, and Johor. As formalized in spatial search frameworks, these dense core markets leverage agglomeration economies to achieve superior vacancy-filling efficiency compared to their peripheral counterparts (vom Berge, 2013). Consequently, the data indicate that Johor's labor market maintains high aggregate labor absorption, characteristic of an expanding manufacturing corridor.

Conversely, the profound structural rigidity of the Traditional/Informal economies (Cluster 0) highlights a dual-track national labor market. The insulation of the East Coast and northern perimeter from these macro-industrial recovery cycles suggests a heavy reliance on inelastic sectors and informal employment. As demonstrated by Giannakis & Bruggeman (2017), rural and peripheral regions often display a form of "defensive resilience" through primary activities during systemic shocks. However, they ultimately lack the structural diversification capacity required for rapid post-crisis expansion. Furthermore, while the informal economy in these regions acts as a vital macroeconomic stabilizer that cushions aggregate output loss, this stabilization inherently operates via severe wage and skill downgrading as displaced formal labor crowds into flexible informal production (Marjit et al., 2023). Therefore, while these traditional structures shielded local workforces from the most severe pandemic-era employment shocks, they simultaneously lack the high-value industrial capacity to capitalize on the subsequent national economic expansion.

Finally, the geographic entrenchment of the remaining Vulnerable/Frictional zones in East Malaysia, particularly Sabah, signals deep-rooted, structural spatial friction. Despite the broader national recovery, these regions continue to experience a disproportionately high unemployment-to-participation ratio. This persistent localized friction is driven by systemic geographical isolation and infrastructural bottlenecks.

As Almaden (2026) demonstrates through a compounded disadvantages framework, the intersection of geographic isolation and institutional voids constitutes a core condition that locks rural and peripheral areas into systemic underdevelopment traps. In East Malaysia, this spatial penalty prevents the effective matching of active labor supply with viable domestic industrial demand, indicating that resolving graduate

unemployment in these regions requires targeted, supply-side developmental and infrastructural interventions rather than broad macroeconomic stimulus.

Educational Mismatch and the TVET Credentialism Trap

To bridge these macroeconomic typologies and localized educational policy, the Diploma Penalty Gap (DPG) was calculated across all 16 states and federal territories from 2020 to 2023. A positive DPG indicates that the labor market penalizes sub-degree credentials (higher-diploma unemployment), whereas a negative DPG signifies a structural "Degree Penalty" (higher university graduate unemployment). Table 3 summarizes the longitudinal evolution of this mismatch across all 16 states and federal territories. To highlight intra-cluster structural variations, jurisdictions are grouped by their 2023 spatial cluster and ranked in descending order by their 2023 DPG values, thereby visually delineating transitions from positive Diploma Penalties to negative Degree Penalties.

Table 3: Diploma Penalty Gap (DPG) by State and Dominant Cluster across the years 2020–2023

Cluster Typology	State / Federal Territory	2020 (%)	2021 (%)	2022 (%)	2023 (%)
1 (Dynamic Industrial)	Selangor	0.0	0.0	0.0	+1.7
	Pulau Pinang	+0.8	+1.1	+0.9	+1.3
	W.P. Kuala Lumpur	+1.3	+0.9	+1.2	+1.0
	W.P. Putrajaya	+0.4	+0.7	+0.2	+0.5
	Sarawak	+0.4	+0.2	+0.2	+0.4
	Melaka	-1.1	-1.0	-1.1	-0.6
	Johor	-1.8	-2.2	-1.4	-1.3
0 (Traditional/Informal)	Pahang	+0.4	+0.3	+0.4	+0.1
	Terengganu	+0.5	+0.4	+0.5	-0.1
	Kelantan	-0.2	-0.4	-0.2	-0.6
	Perlis	-1.0	-1.5	-1.5	-1.0
	Negeri Sembilan	-1.0	-0.6	-1.1	-1.3
	Kedah	-2.4	-2.5	-2.2	-2.0
	Perak	-3.0	-3.7	-3.0	-3.0
2 (Vulnerable/Frictional)	W.P. Labuan	+0.5	-0.1	+0.3	+0.1
	Sabah	-0.7	-1.1	-1.1	-0.2

The gap analysis (Figure 5), derived from recent Department of Statistics Malaysia (2025) graduate statistics, reveals stark structural polarization across the three macro-clusters. Specifically, the Dynamic Industrial Hubs (Cluster 1) exhibit an upward trajectory culminating in a net positive average gap (+0.43% in 2023), anchored by core metropolitan centers such as Selangor (+1.7%), Pulau Pinang (+1.3%), and W.P. Kuala Lumpur

(+1.0%). However, this dynamic simultaneously exposes a "credentialism trap" within primary urban hubs. As these metropolitan economies transition toward knowledge-intensive service sectors, employers increasingly rely on university degrees as a baseline screening mechanism for technical roles. Consequently, in line with theoretical crowding-out frameworks identified by Capsada-Munsech and Ortiz-Gervasi (2025), this pattern suggests that credential inflation may place TVET diploma holders at a relative disadvantage in hiring queues.

Notably, Table 3 also demonstrates intra-cluster divergence within Cluster 1 between metropolitan knowledge hubs and secondary manufacturing corridors. While southern industrial states such as Johor (-1.3%) and Melaka (-0.6%) achieve high aggregate employment participation, placing them within the Dynamic Industrial tier, they simultaneously exhibit a persistent negative DPG (a Degree Penalty). Because their industrial bases prioritize mid-tier manufacturing and operational production over high-value research and corporate services, local factories rapidly absorb TVET technicians while leaving an oversupply of university degree holders. This localized gap between manufacturing demand and degree output provides contextual support for the 'Singapore Pull Factor' hypothesis (Jauhar et al., 2015; Rizzo & Glasson, 2011), wherein cross-border wage differentials may contribute to selective human capital outflows.

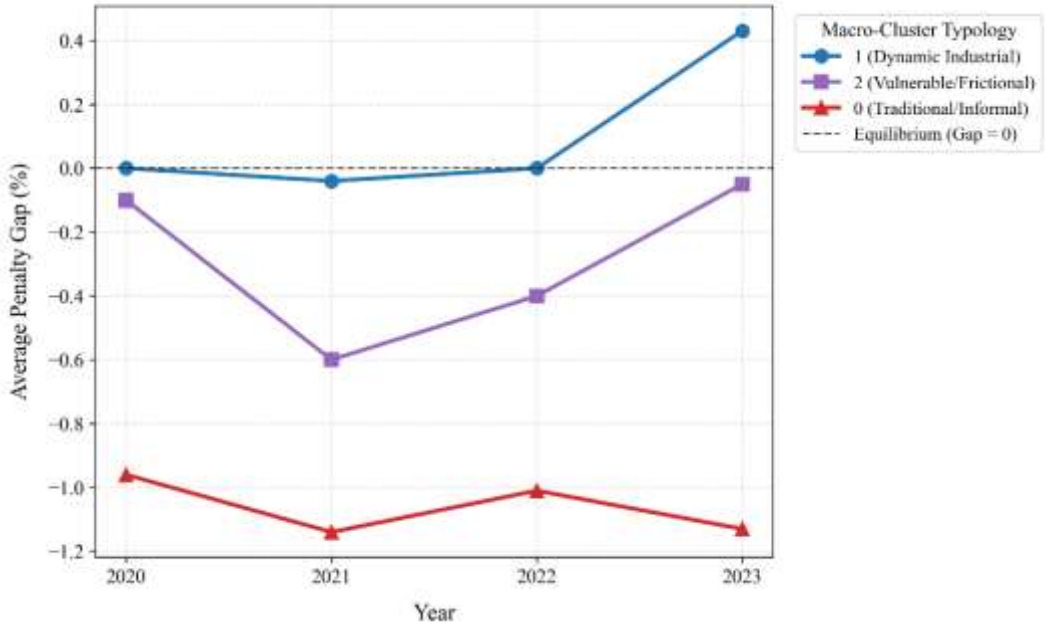

Figure 5: Evolution of the Diploma vs. Degree Unemployment Gap across the three macroeconomic spatial typologies (2020–2023). The dashed line represents the educational equilibrium (DPG = 0).

Conversely, the Traditional/Informal Economies (Cluster 0) suffer a persistent and severe structural Degree Penalty, averaging -1.13% in 2023 and reaching extremes in Perak (-3.0%), Kedah (-2.0%), and Negeri Sembilan (-1.3%). Meanwhile, the Vulnerable/Frictional Zones (Cluster 2), dominated by Sabah (-0.2%) and W.P. Labuan (+0.1%), hover near educational equilibrium (-0.05% average) simply because elevated frictional unemployment affects both qualification tiers almost equally.

To formally quantify the relationship between macroeconomic absorption capacity and qualification mismatch, a bivariate correlation analysis was conducted between each state's true Employment-to-Population (EP) Ratio and its 2023 DPG (Figure 6). The results show a statistically significant positive bivariate correlation ($R = 0.506$, $p = 0.0454$, $N=16$, 95% CI [0.014, 0.801]), which is further corroborated by a non-parametric Spearman rank robustness check ($\rho = 0.539$, $p = 0.031$). This indicates that lower regional employment generation capacity is associated with a more pronounced relative penalty for university degree holders, though the wide confidence interval underscores the exploratory nature of this state-level aggregation.

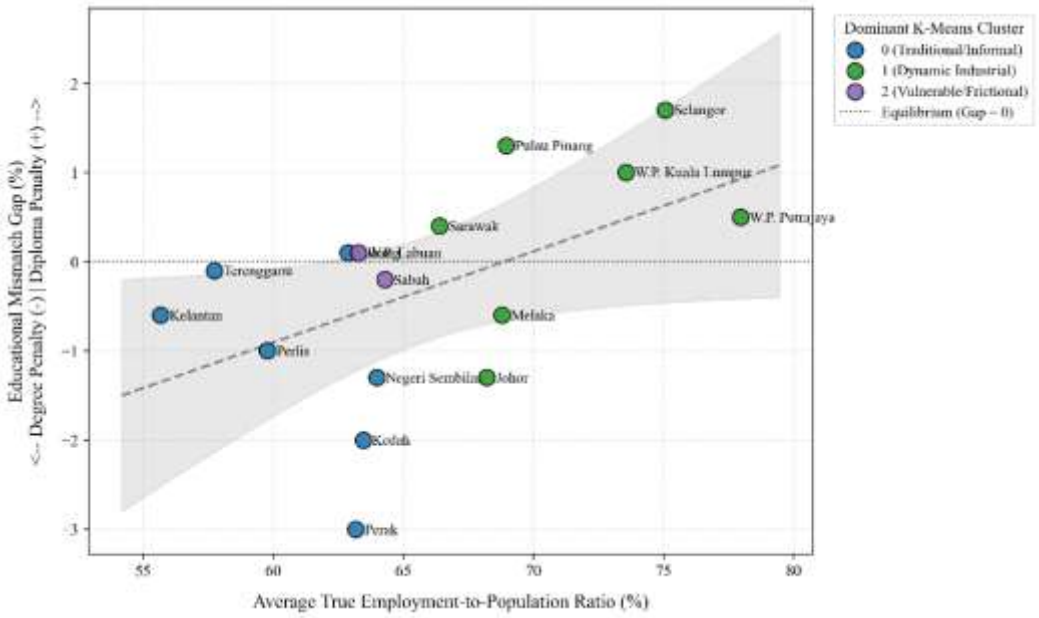

Figure 6: Pearson correlation analysis (R=0.506) demonstrates a linear relationship between a region's Employment-to-Population Ratio and its resulting Educational Mismatch Gap for 2023.

CONCLUSION

This study deployed an unsupervised machine learning approach to deconstruct the spatial heterogeneities of the Malaysian labor market, challenging the traditional reliance on aggregate national unemployment statistics. By utilizing K-Means clustering on granular macroeconomic data from 2021 to 2023, the analysis identified three distinct regional typologies. Consequently, this provides robust empirical evidence indicating that the national labor landscape operates as a fractured ecosystem of localized economies rather than a homogeneous entity.

The empirical findings highlight critical structural divergences in post-pandemic recovery and graduate employability. While the Vulnerable tier recorded a net reduction of 100 constituencies in favor of Dynamic Industrial Hubs between 2021 and 2023, East Coast and northern peripheral constituencies remained entrenched in Traditional/Informal structures. Furthermore, the Diploma Penalty Gap (DPG) analysis exposes a dual crisis in tertiary labor absorption. First, primary metropolitan cores within Cluster 1 (Selangor, Kuala Lumpur, and Pulau Pinang) exhibit a positive Diploma Penalty Gap, consistent with patterns where higher credential requirements may elevate entry barriers for TVET and diploma graduates. Second, Traditional Economies (Cluster 0) and secondary manufacturing corridors (such as Johor and Melaka within Cluster 1) suffer from a persistent "Degree Penalty". In these regions, the inability of local industrial structures to generate high-value, knowledge-intensive positions leaves university degree holders vulnerable to structural underemployment and contextually indicated cross-border human capital flight.

Finally, the observed correlation between regional employment capacity and educational mismatch ($R = 0.506$, $p = 0.0454$) illustrates that educational outcomes are closely linked to localized industrial absorption capacity. Therefore, uniform, "one-size-fits-all" national education policies are insufficient; thus, resolving these structural imbalances requires a shift toward spatially targeted, demand-side industrial strategies.

LIMITATIONS AND FUTURE RESEARCH

Despite these empirical contributions, this study acknowledges several methodological limitations. First, the multiscale analytical framework bridges fine-grained constituency-level macroeconomic clustering with macro-regional state-level educational metrics. This spatial aggregation inherently introduces ecological inference constraints associated with the Modifiable Areal Unit Problem (MAUP). Second, the study relies on

aggregate regional labor statistics; the absence of micro-level graduate tracer datasets and localized cross-border commuting logs restricts the precise quantification of human capital flight. To advance this framework, future researchers should integrate primary tracer survey data and mobile network signaling data to capture actual workforce mobility. Methodologically, subsequent studies must incorporate spatially constrained clustering algorithms (e.g., SKATER) and spatial autoregressive econometric models (such as Spatial Durbin Models) to explicitly account for geographic adjacency and spatial weight matrices (W).

Data Availability Statement

The constituency-level macroeconomic and labor market data (2021–2023) that form the core dataset for the unsupervised machine learning analysis in this study are publicly available. All primary data used in the K-Means clustering methodology were retrieved from the Department of Statistics Malaysia (DOSM) and are available on Malaysia's Official Open Data Portal at https://data.gov.my/data-catalogue/lfs_dun. The custom Python scripts developed for automated data extraction, preprocessing, and spatial cluster generation are available from the corresponding author upon reasonable request.

REFERENCES

1. Abdila, N. S., Fitriani, R., & Pratama, M. L. (2025). Clustering With Skater Methods And Utilization Of Lisa On Unemployment Rate. BAREKENG: Jurnal Ilmu Matematika Dan Terapan, 19(4), 2633–2646. https://doi.org/10.30598/barekengvol19iss4pp2633-2646
2. Almaden, C. R. C. (2026). Compounded disadvantages in rural regions: A configurational analysis for understanding development deficits in southern Philippines. Journal of Rural Studies, 124, 104193. https://doi.org/10.1016/J.JRURSTUD.2026.104193
3. Andena, S. R., Kuna, G., & Chappidi, E. (2025). A Scalable Framework for Fair Clustering with Hyperparameter Optimization and Multi-Attribute Constraints. 2025 1st International Conference on Advancement in Futuristic Technologies (ICAFT), 13, 1–11. https://doi.org/10.1109/ICAFT66710.2025.11452743
4. Anselin, L. (2010). Thirty years of spatial econometrics. Papers in Regional Science, 89(1), 3–25. https://doi.org/10.1111/J.1435-5957.2010.00279.X
5. Asadullah, M. N., Joseph, J., & Chin, J. (2023). The Political Economy of Poverty Reduction in Malaysia. Progress in Development Studies, 23(2), 127–151. https://doi.org/10.1177/14649934231151486
6. Ashmarina, S., & Vochozka, M. (2019). Sustainable Growth and Development of Economic Systems Contradictions in the Era of Digitalization and Globalization. In . http://www.springer.com/series/1262
7. Brada, J. C., Gajewski, P., & Kutan, A. M. (2021). Economic resiliency and recovery, lessons from the financial crisis for the COVID-19 pandemic: A regional perspective from Central and Eastern Europe. International Review of Financial Analysis, 74, 101658. https://doi.org/10.1016/J.IRFA.2021.101658
8. Capsada-Munsech, Q., & Ortiz-Gervasi, L. (2025). Crowding out or “one-size-fits-all” occupations? A regional exploration of youth overeducation in Spain. International Labour Review, 164(2). https://doi.org/10.16995/ilr.18846
9. Chuang, Y. chyi, & Liang, C. Y. (2022). Overeducation and skill mismatch of university graduates in Taiwan. Review of Development Economics, 26(3), 1693–1712. https://doi.org/10.1111/rode.12891
10. Department of Statistics Malaysia, D. (2025). Graduates Statistics 2024. Department of Statistics, Malaysia Portal.
11. Eom, H. (2022). Does Job Accessibility Matter in the Suburbs? Black Suburbia, Job Accessibility, and Employment Outcomes. Land, 11(11), 1952. https://doi.org/10.3390/land11111952
12. Filiztekin, A. (2009). Regional unemployment in Turkey. Papers in Regional Science, 88(4), 863–879. https://doi.org/10.1111/j.1435-5957.2009.00237.x
13. Giannakis, E., & Bruggeman, A. (2017). Economic crisis and regional resilience: Evidence from Greece. Papers in Regional Science, 96(3), 451–477. https://doi.org/10.1111/pirs.12206
14. Gooptu, S., & Mukherjee, V. (2023). School dropout and overeducation in developing economies: Feasibility of a budgetary solution †. Review of Development Economics, 27(2), 825–852. https://doi.org/10.1111/rode.12964

15. Hasnan, K. A., Zakariya, Z., Yin, K. Y., & Aziz, S. N. A. (2019). Skill mismatch among instructors across technical and vocational education training (TVET) institutions in Malaysia. International Journal of Innovation, Creativity and Change, 10(6), 270–279.
16. Hemavathy, P., & Chinnadurai, M. (2022). Performance validation of clustering algorithms using selection of attributes and application of filters in terms of data reduction. Concurrency and Computation: Practice and Experience, 34(8). https://doi.org/10.1002/CPE.5364
17. Idhom, M., Muhaimin, A., Damaliana, A. T., Trimono, Nasrudin, M., & Wara, S. S. M. (2025). Evaluating Clustering Metrics for Time-Series Data: A Power Test Approach. 2025 IEEE 11th Information Technology International Seminar (ITIS), 580, 368–373. https://doi.org/10.1109/ITIS67966.2025.11308723
18. Jauhar, J., Ghani, A. B. A., Joarder, M. H. R., Subhan, M., & Islam, R. (2015). Brain drain to Singapore: A conceptual framework of Malaysians' diaspora. Social Sciences (Pakistan), 10(6), 702–711.
19. Jin, S., Nie, T., Pun, N., & Xu, D. (2022). Spatial Mismatch, Different Labor Markets and Precarious Employment: The Case of Hong Kong. Social Indicators Research, 161(1), 51–73. https://doi.org/10.1007/s11205-021-02819-z
20. Kosfeld, R., & Dreger, C. (2006). Thresholds for employment and unemployment: A spatial analysis of German regional labour markets, 1992–2000. Papers in Regional Science, 85(4), 523–542. https://doi.org/10.1111/j.1435-5957.2006.00096.x
21. Kosfeld, R., & Dreger, C. (2019). Towards an East German wage curve - NUTS boundaries, labour market regions and unemployment spillovers. Regional Science and Urban Economics, 76(January 2018), 115–124. https://doi.org/10.1016/j.regsciurbeco.2018.01.006
22. Ma, W., Baek, J., Qi, M., Li, J., & Liu, B. (2022). The Influence of Overeducation on Chinese Workers' Job Satisfaction from China Household Tracking Survey (2014–2018). International Journal of Environmental Research and Public Health, 19(23), 16032. https://doi.org/10.3390/ijerph192316032
23. Marjit, S., Bhattacharyya, R., & Yang, L. (2023). The Role of Informality in Moderating the Impact of Adverse Macroeconomic Shocks. In Economic and Political Weekly (Vol. 58, Issue 48, pp. 13–15). https://doi.org/10.2139/ssrn.4527561
24. Mumtaz, Z., & Whiteford, P. (2021). Machine Learning Based Approach for Sustainable Social Protection Policies in Developing Societies. Mobile Networks and Applications, 26(1), 159–173. https://doi.org/10.1007/s11036-020-01696-z
25. Nistor, A. (2009). Assessing the Effectiveness of Human Capital Investments on the Regional Unemployment Rate in the United States: 1990 and 2000. International Regional Science Review, 32(1), 65–91. https://doi.org/10.1177/0160017608325594
26. Ogryzek, M., & Jaskulski, M. (2025). Applying Methods of Exploratory Data Analysis and Methods of Modeling the Unemployment Rate in Spatial Terms in Poland. Applied Sciences, 15(8), 4136. https://doi.org/10.3390/app15084136
27. Omar, N. H., Abdul Manaf, A., Helma Mohd, R., Che Kassim, A., & Abd. Aziz, K. (2012). Graduates' Employability Skills Based on Current Job Demand through Electronic Advertisement. Asian Social Science, 8(9), 103–112. https://doi.org/10.5539/ass.v8n9p103
28. Reza, D. T., Sunarno, Supriadi, Wahyuni, S., Arrizal, H. H., & Sinambela, M. (2025). Comparative Analysis of Earthquake Clustering in West Java Using Unsupervised Learning: K-Means, K-Medoids, and Fuzzy C-Means. 2025 International Conference on Information Technology and Computing (ICITCOM), 104, 489–494. https://doi.org/10.1109/ICITCOM66635.2025.11265106
29. Rizzo, A., & Glasson, J. (2011). Conceiving transit space in Singapore/Johor: a research agenda for the Strait Transnational Urban Region (STUR). International Journal of Urban Sustainable Development, 3(2), 156–167. https://doi.org/10.1080/19463138.2011.590722
30. Rodzalan, S. A., Mohd Noor, N. N., Abdullah, N. H., & Mohamed Saat, M. (2022). TVET Skills Gap Analysis in Electrical and Electronic Industry: Perspectives from Academicians and Industry Players. Journal of Technical Education and Training, 14(1), 158–177. https://doi.org/10.30880/jtet.2022.14.01.014
31. Rusanovskiy, V. A., & Markov, V. A. (2016). Influence of the spatial factor on the regional differentiation of unemployment in the Russian economy. Studies on Russian Economic Development, 27(5), 593–604. https://doi.org/10.1134/S1075700716050129

32. Sattinger, M., & Hartog, J. (2013). Nash bargaining and the wage consequences of educational mismatches. Labour Economics, 23, 50–56. https://doi.org/10.1016/j.labeco.2013.03.002
33. Simeonov, Z. G. (2024). Analyzing the Impact of Workforce Distribution on Regional Economy Using GIS: Spatial Analysis. 2024 59th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), 1–4. https://doi.org/10.1109/ICEST62335.2024.10639670
34. Takahashi, T., Takatsuka, H., & Zeng, D.-Z. (2013). Spatial inequality, globalization, and footloose capital. Economic Theory, 53(1), 213–238. https://doi.org/10.1007/s00199-011-0686-7
35. Tobler, L., & Leesch, J. (2026). Workers, jobs, and how they are matched: A decomposition of US labor market trends in educational mismatch. Social Science Research, 135, 103311. https://doi.org/10.1016/j.ssresearch.2026.103311
36. vom Berge, P. (2013). Search unemployment and new economic geography. The Annals of Regional Science, 50(3), 731–751. https://doi.org/10.1007/s00168-012-0535-z
37. Yen, W. T. M., Yen, Y. Y., Choy, C. S., Nee, T. W., & Ming-Li, T. (2023). Perspectives of employers on graduate employability skills: A case of Malaysia. Asian Development Policy Review, 11(4), 229–243. https://doi.org/10.55493/5008.v11i4.4946
38. Zaid, N. S., & Kamin, Y. (2026). Beyond Digital Skills: Addressing The Technical and Soft Skills Deficit in Malaysian TVET for the Manufacturing Sector. Journal Of Technical Education And Training, 18(1), 360–374. https://doi.org/10.30880/jtet.2026.18.01.025
39. Zulkefli, N., & Saad, N. M. (2026). Evaluating Implementation Effectiveness in Malaysian Higher TVET: A Lean Waste–SERVQUAL Integrated Framework. Applied Research, 5(3), 701–720. https://doi.org/10.1002/appl.70117