Profiling Japanese Universities for Thai Students' Decision Support Using Hierarchical Clustering

Tanaanan Chalermpan, Nakarin Israngkura Na Ayudhaya, and Sarach Tuomchomtam *

Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand
Email: tanaanan.c@ku.th (T.C.); nakarin.is@ku.th (N.I.N.A.); sarach.t@ku.th (S.T.)
*Corresponding author
Manuscript received May 7, 2026; revised June 11, 2026; accepted July 9, 2026; published November 24, 2026

***Abstract*—Selecting an appropriate overseas university is a high-stakes, information-intensive decision increasingly recognized as a core challenge in educational technology and knowledge management. For Thai students targeting Japan, this challenge is acute: the absence of a structured, data-driven decision support tool forces reliance on incomplete rankings and informal advice, resulting in persistent information overload and prestige bias. To address this gap, this study proposes an Educational Data Mining (EDM) framework that models Japanese universities as institutional knowledge archetypes and operationalizes them as a student-facing decision support tool. Applying hierarchical clustering with Ward's Linkage to an equal-weighted, multi-dimensional feature set spanning academic metrics, financial costs, and environmental factors, we analyzed 40 Japanese universities with verified Thai-Japan institutional ties. The analysis yields six archetypal institutional profiles, from the high-cost, high-research "Hyper-Urban Core" and elite "Innovation Arch" to the value-oriented "Traditional Heartland" and technology-specialized "Golden Route" clusters, each capturing a distinct combination of academic focus, financial accessibility, and geographic context. These profiles are operationalized as a strategic decision matrix enabling students to narrow 40 institutions to a targeted shortlist by academic, financial, and lifestyle criteria. Cluster distinctiveness is confirmed via Analysis of Variance (ANOVA) and Chi-squared tests across 15 of 18 features ($p < 0.05$), robust to non-parametric testing and Benjamini-Hochberg correction. Cluster stability is confirmed through external validation against *k*-means clustering (mean Adjusted Rand Index (ARI) = 0.784) and Principal Component Analysis (PCA) robustness checks across three dimensionality reduction criteria (ARI ≥ 0.916). The methodology is data-agnostic and scalable, offering a reproducible blueprint for transparent, student-centered decision support in international higher education.**

***Keywords*—educational data mining, hierarchical clustering, decision support framework, learning analytics, international student mobility, higher education technology**

I. Introduction

Japan has emerged as one of the most sought-after higher education destinations in Asia, attracting students through a combination of world-class research institutions, government scholarship programs administered by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) [1], the Japan Society for the Promotion of Science (JSPS) [2], and the Japan Student Services Organization (JASSO) [3], and a growing network of bilateral academic partnerships. For Thai students in particular, who benefit from formal institutional ties brokered through the JSPS Bangkok Office [4], Japan represents a uniquely accessible yet structurally complex study-abroad landscape. Despite this accessibility, the university selection process remains a poorly supported, high-stakes decision task characterized by information overload, prestige bias, and a reliance on informal peer networks rather than objective, data-driven guidance.

This challenge is well-documented in the educational technology and knowledge management literature. Information overload, in which an excess of unstructured data impairs decision quality rather than improving it, is a recognized failure mode in higher education guidance [5]. Traditional university ranking systems compound this problem: as Sepehr *et al.* [6] demonstrate, global rankings are systematically biased toward institutions in high-income, high-Research and Development (R&D) countries and prioritize research output over the student-relevant attributes (cost of living, disciplinary focus, admission accessibility, and campus environment) that determine practical fit for an individual applicant. The consequence is a decision landscape in which the most visible information is the least useful for a specific student's needs.

Previous research has approached the Thai student-to-Japan mobility challenge from two distinct but complementary directions. Supply-side policy analyses by Sato [7, 8] employed Logical Framework Approach (LFA) [9] and Standardized Mean Difference (SMD) methods to evaluate national policy effectiveness but offer no actionable guidance to individual student decision-makers. Demand-side survey research by Tomita and Mochizuki [10] identified key student preferences, particularly for English-medium instruction, but survey-based methods suffer from sample limitations and an inability to provide the university-specific, multi-dimensional profiles that constitute actionable decision information. Neither tradition has produced a practical tool for student-facing decision support.

The broader Information and Education Technology literature points toward Educational Data Mining (EDM) and learning analytics as natural frameworks for bridging this gap. EDM methods have been applied to cluster students by learning behavior and predict academic achievement outcomes in higher education contexts. In related work, Nuankaew *et al.* [11] demonstrated that machine learning clustering techniques including *k*-means and decision tree classification can yield meaningful student achievement group profiles from institutional academic data, thereby establishing the methodological viability of EDM clustering for educational decision support, a principle this study applies at the institutional-profile level, building on the methodological precedent established by Nuankaew *et al.* [11] in this journal. Additionally, Matlis *et al.* [12] and Maral [13] have demonstrated EDM clustering at the institutional level, while comprehensive portals such as STUDY in JAPAN [14] and Kei-Net Plus [15]

aggregate raw institutional data but provide no analytical or decision support layer.

To close this gap, this study makes three integrated contributions to the educational technology and learning analytics literature. First, we construct a multi-dimensional institutional knowledge model of Japanese higher education by applying hierarchical clustering with Ward's Linkage to a feature set combining academic selectivity, research capacity, disciplinary composition, financial cost, and international program availability across 40 Japanese universities with verified Thai-Japan institutional ties [4]. Second, we characterize the resulting six institutional archetypes through rigorous statistical characterization (Analysis of Variance (ANOVA), Kruskal-Wallis, and Chi-squared tests, with Benjamini-Hochberg correction and external validation via Adjusted Rand Index), establishing that the clusters represent structurally distinct institutional types rather than arbitrary groupings. Third, we operationalize the institutional taxonomy as a strategic decision support matrix, a binary checklist tool that enables students to systematically align their individual profiles with the appropriate institutional archetype, reducing decision complexity from 40 undifferentiated options to a targeted, reasoning-supported shortlist.

II. Related Works

This section maps the existing literature across three converging research streams: international student mobility and policy evaluation, EDM-based university benchmarking and clustering, and student-facing decision support tool design. Table 1 provides a comparative summary including the newly integrated work by Nuankaew *et al.* [11].

Table 1. Comparison of related research areas by analytical level, methodology, and identified gap

Work	Objective	Scope / Method	Key Takeaway / Gap
Sato (2005) [7]	Policy Eval. (LFA)	Policy / Qualitative	Establishes logical basis for policy evaluation; no student-level guidance.
Sato (2012) [8]	Policy Eval. (SMD)	Policy / Quantitative	Quantifies impact via standardized metrics; aggregate-level only.
Tomita & Mochizuki (2016) [10]	Demand-Side Analysis	Student Opinion / Survey	Identifies English programs as critical; sample bias limits generalizability.
Kler *et al.* (2023) [5]	Migration Modeling	Country-Level / Gravity	Confirms structural data predicts student flows; treats Japan as monolithic.
Nuankaew *et al.* (2022) [11]	Achievement Clustering	University / *k*-means + ML	Applies EDM clustering to predict student achievement groups; campus-level context only.
Matlis *et al.* (2024) [12]	Institutional Clustering	University / Unsupervised ML	Demonstrates clustering for institutional grouping; ignores cost and lifestyle dimensions.
Sepehr *et al.* (2025) [6]	Ranking Correlation	Global Rankings / Correlation	Exposes economic bias in global rankings; motivates alternative metrics.
Maral (2026) [13]	Performance Eval. (MCDM)	University / Hybrid clustering	Provides nuanced institutional evaluation; not student-facing.
This study	EDM-based decision support framework	University / HC + ANOVA / Kruskal-Wallis	Multi-dimensional student-centered archetypes + operational Information and Communications Technology (ICT) decision matrix.

LFA = Logical Framework Approach; SMD = Standardized Mean Difference; HC = Hierarchical Clustering; MCDM = Multi-Criteria Decision-Making; ML = Machine Learning.

Existing research on international student mobility to Japan has predominantly operated at the policy or macro-structural level. Sato [7] evaluated Japan's foreign student policy using the Logical Framework Approach (LFA) [9], effectively mapping policy logic but lacking the quantitative depth required to generate actionable student guidance. Sato's subsequent work [8] introduced SMD statistics but remained focused on aggregate policy outcomes rather than institutional profiles. Macro-level econometric studies using gravity models [16] and push-pull factor frameworks [5] identified national-level migration determinants but treated "Japan" as a homogeneous destination, obscuring critical variance in campus cultures, research intensity, admission difficulty, and living costs.

At the institutional analysis level, Matlis *et al.* [12] successfully segmented universities using unsupervised learning on academic performance rankings, demonstrating the methodological viability of clustering for institutional profiling. However, their reliance exclusively on research output and reputation metrics ignores multi-dimensional practicalities such as financial constraints, English-language program availability, and regional lifestyle characteristics that constitute the decision-relevant attributes for an international student applicant. Maral [13] extended institutional evaluation by pairing clustering with Multi-Criteria Decision-Making (MCDM), producing more nuanced performance profiles, but the framework remains designed for institutional benchmarking rather than student decision support. Sepehr *et al.* [6] established that standard global rankings systematically favor institutions in high-income, high-R&D countries, motivating the need for alternative, student-centered metrics.

A complementary tradition in educational evaluation employs Multi-Criteria Decision-Making (MCDM) frameworks with expert-elicited criteria weights. At the faculty level, Radovanović *et al.* [17] applied Spherical Fuzzy AHP combined with Grey Measurement of Alternatives and Ranking according to Compromise Solution (MARCOS) to rank university professors across qualitative criteria including teaching quality and ethical behavior. At the institutional strategy level, Bouraima *et al.* [18] used the FullEX method to prioritize sustainability strategies for higher education in developing countries, while fuzzy VIKOR with Entropy-derived weights has been applied to institutional ranking under uncertainty [19]. While these approaches offer interpretable, weight-explicit structures

suited to expert-driven evaluation, they require pre-specified criteria weights that introduce subjectivity when individual student preferences vary substantially. The present study complements this tradition by deriving institutional archetypes directly from observed structural data, enabling student-facing decision support without weight elicitation. Additional MCDM applications in adjacent educational contexts further demonstrate the broad applicability of structured decision frameworks, though these operate at different analytical levels from the institutional profiling focus of the present study [20–23].

Of particular methodological relevance is the work by Nuankaew *et al.* [11], published in this journal, which applied *k*-means clustering combined with four classification tree techniques to profile student achievement groups in an educational technology program. Their Cross-Industry Standard Process for Data Mining (CRISP-DM) pipeline, encompassing unsupervised cluster assignment followed by supervised classification tree prediction (achieving 68.37% accuracy with Random Forest), established a precedent for student profiling research in the educational context of the field of Educational Data Mining (EDM). The present study applies this methodological tradition at the international institutional-profile domain: where Nuankaew *et al.* [11] cluster students within a single institution, we cluster institutions to serve individual students across an international higher education system.

Within the educational technology literature, the design of Information and Communications Technology (ICT)-based knowledge tools for student decision-making has become an active research area. Such tools deliver context-sensitive, structured information to users at the point of need, reducing reliance on expert intermediaries. These systems share a common design principle: transforming raw data into organized, actionable knowledge representations which aligns precisely with the present study's approach. The institutional archetype profiles derived through clustering serve as the knowledge representation layer; the strategic decision matrix constitutes the student-facing delivery mechanism.

III. Materials and Methods

A. Framework Overview

The analytical framework integrates six sequential stages (Fig. 1): (1) multi-source data collection from authoritative institutional databases; (2) feature engineering and pre-processing via Z-score normalization and one-hot encoding; (3) hierarchical clustering with Ward's Linkage, multi-index cluster selection, and Principal Component Analysis (PCA) robustness check; (4) statistical characterization of cluster distinctiveness via ANOVA, Kruskal-Wallis, Chi-squared tests, and Benjamini-Hochberg correction; (5) external cluster validation via *k*-means comparison and Adjusted Rand Index; and (6) institutional knowledge taxonomy construction and decision support operationalization via a strategic decision matrix. All analyses were implemented in Python 3.9 using Scikit-learn 1.3, SciPy 1.11, Statsmodels 0.14, Matplotlib 3.7, Seaborn 0.12, and Geopandas 0.14.

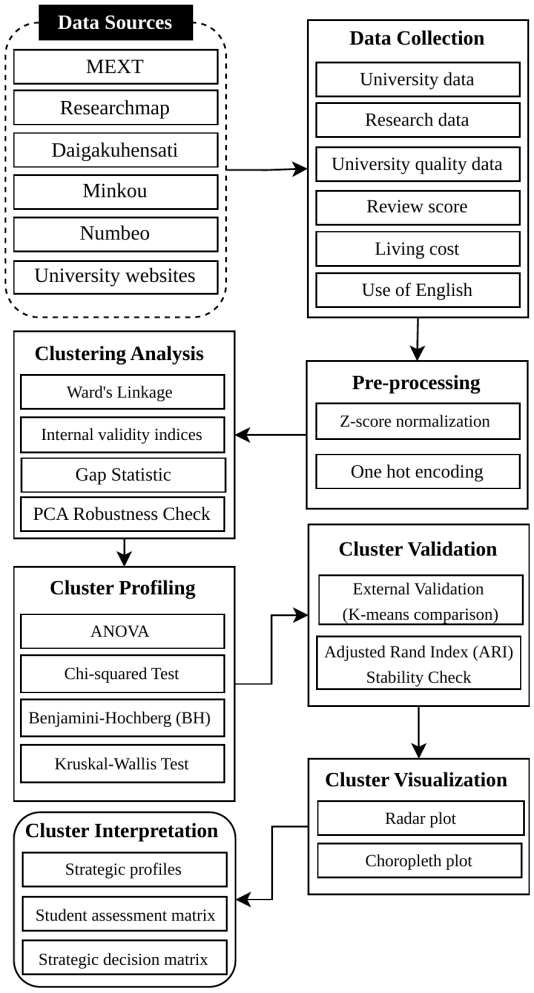

Fig. 1. Six-stage EDM framework for constructing institutional knowledge archetypes of Japanese universities and delivering them as a student-facing decision support tool.

B. Experimental Data

1) Data sources

This study aggregates institution-level structural data from five authoritative, publicly accessible sources: (a) MEXT [1] for official institutional classifications; (b) researchmap, operated by Japan Science and Technology Agency (JST) [24], for research activity metrics including researcher counts, publication volumes, and disciplinary composition; (c) Daigaku-hensati.com for the Hensachi admission selectivity score [25]; (d) Minkou.jp for aggregate student review scores [26]; and (e) Numbeo [27] for location-specific monthly living cost estimates in USD.

2) Data collection and cohort definition

The study cohort is defined by strategic relevance to the Thai-Japan student mobility context: we target all universities registered with the JSPS Bangkok Office [4], whose registration signals a verified institutional commitment to Thai-Japan academic exchange. Of 43 registered institutions, three are excluded: the National Institute of Technology (KOSEN) due to its vocational focus, Shizuoka University and Tamagawa University due to incomplete feature data availability, yielding a final analytical cohort of 40 universities. Full cluster membership details for all 40 institutions, including university name, institutional type, and prefecture, are provided in (Table A1).

Thirteen features organized across six thematic categories are assembled for each institution (Table 2). This feature set is designed to be orthogonal in conceptual scope, spanning academic, research, environmental, financial, and programmatic dimensions so as to ensure that the resulting clusters capture multi-dimensional institutional profiles rather than a single latent dimension such as research prestige.

Table 2. Summary of dataset features by source, category, type, and description

Source	Feature	Type	Range (Examples)	Description
MEXT[a]	Name	Text	Chiba Univ., Tokai Univ.	Full official university name.
	Type	Categorical	National, Private	Official institutional classification.
	State	Categorical	Hokkaido, Osaka	Prefecture of primary campus.
Hensachi[b]	Admission Score	Continuous	35.5 / 52.2 / 65.7	Standardized admission selectivity score (Hensachi).
Minkou[c]	Review Score	Continuous	3.59 / 4.0 / 4.25	Aggregate student community rating (1–5 scale).
researchmap[d]	Research Students	Discrete	0 / 29 / 521	Number of enrolled research (postgraduate) students.
	Researcher Counts	Discrete	1 / 2481 / 11,247	Total faculty and researchers affiliated.
	Publication Counts	Discrete	150 / 240K / 1.5M	Cumulative publications indexed in researchmap.
Websites[e]	Subject Area Counts	Continuous	E.g., 61%, 33%, 12%	Proportion of researchers across 8 disciplines.
	Tuition Discount	Binary	No = 0, Yes = 1	Availability of international tuition discount.
	IDP Availability	Binary	No = 0, Yes = 1	Availability of Integrated Doctoral Program.
	ETP Availability	Binary	No = 0, Yes = 1	Availability of English Taught Program.
Numbeo[f]	Living Cost	Continuous	866 / 1238 / 1698	Estimated monthly living cost in USD.

Superscript labels denote feature categories: (a) Institutional Identity, (b) Academic Selectivity, (c) Campus Environment, (d) Research Capacity, (e) International Accessibility, (f) Financial Context.

C. Data Pre-processing

Feature parity is established through two complementary pre-processing operations. Continuous and discrete numerical features are standardized to zero mean and unit variance via Z-score normalization, preventing features with larger absolute scales from exerting disproportionate influence on Euclidean distance calculations during clustering. Categorical nominal variables (University Type, Prefecture/State) are transformed into binary indicator vectors through one-hot encoding, enabling their inclusion in the distance metric without imposing an artificial ordinal structure.

D. Clustering Algorithm and Optimal Cluster Selection

Hierarchical Clustering (HC) with Ward's [28] minimum variance linkage criterion is selected as the primary analytical method for three methodological reasons: it requires no pre-specification of K (enabling data-driven cluster number selection via dendrogram analysis), it is robust to non-spherical cluster shapes common in institutional data, and it produces a hierarchical structure that reveals both coarse and fine-grained institutional groupings.

Solution quality is evaluated using four complementary internal validation metrics: (1) Inertia (Within-Cluster Sum of Squares, WCSS); (2) Silhouette coefficient [29], which quantifies the ratio of intra-cluster cohesion to inter-cluster separation; (3) Calinski-Harabasz index [30], which maximizes the ratio of between-cluster to within-cluster dispersion; and (4) Davies-Bouldin index [31], which minimizes the average maximum ratio of within-cluster scatter to between-cluster distance.

In cases where internal validation metrics do not converge on a single optimal K, the gap statistic [32] is applied as a formal decision criterion. The gap statistic compares the within-cluster dispersion of the observed data against that expected under a null reference distribution, with the optimal K identified as the smallest K such that $\text{Gap}(K) \geq \text{Gap}(K+1) - \text{sk}(K+1)$, where $\text{sk}(K)$ denotes the simulation error-adjusted standard deviation of the reference distribution. Reference datasets were generated uniformly over the range of each observed feature using $B = 100$ Monte Carlo samples. Application of this decision rule to the present dataset, and the resulting K selection, is discussed in Section IV.A.

The full feature matrix after one-hot encoding comprises 43 dimensions (13 continuous/discrete features, 3 binary program indicators, 2 university type indicators, and 25 prefecture indicators) for $n = 40$ observations, placing the analysis near the boundary identified by Beyer *et al.* [33] where Euclidean distances in high-dimensional spaces may become less discriminative. To formally assess cluster stability under dimensionality reduction, a PCA robustness check was conducted across three standard component-retention criteria. Under the Kaiser criterion (eigenvalue > 1; 4 components; 67.1% variance explained), Ward's $K = 6$ re-clustering yielded an Adjusted Rand Index (ARI) of 0.916 against the full-dimensional partition, indicating substantial agreement. Retaining components at the 80% cumulative variance threshold (7 components; 83.0%) and 95% threshold (13 components; 95.6%) both yielded ARI = 1.000, confirming an identical six-cluster partition. The reported cluster structure is therefore stable across all three retention levels and is not an artifact of high-dimensional Euclidean geometry. The stability of the $K = 6$ partition across these three substantially different feature representations confirms that the cluster structure is not driven by any single variable group, including the research-related indicators that achieve the highest F-statistics in the cluster characterisation.

E. Statistical Characterization of Clusters

Cluster structural distinctiveness is assessed using two complementary inferential tests selected according to feature measurement type. One-way ANOVA is applied to all continuous and discrete numerical features to test whether population means differ significantly across the $K = 6$ clusters ($p < 0.05$). Pearson's Chi-squared test of independence is applied to all binary and categorical features. A significance threshold of $p < 0.05$ is adopted throughout. Although the dataset comprises 13 source features (Table 2), the disciplinary composition feature (Subject Area Counts) decomposes into 8 individual discipline proportions,

expanding the feature set to 20 candidate variables. Of these, Name and State are excluded from statistical testing as non-quantitative identifiers, yielding the 18 testable variables (14 continuous or discrete and 4 categorical) used for statistical characterization. Given the simultaneous testing of 18 features, results are interpreted conservatively; features with $p < 0.01$ or $p < 0.001$ are considered to provide the strongest evidence of cluster distinctiveness. Prior to ANOVA, normality and homogeneity of variance are assessed via Shapiro-Wilk [34] and Levene [35] tests respectively. For features violating these assumptions, Kruskal-Wallis non-parametric tests [36] are conducted as a robustness check. To address multiple comparison concerns, Benjamini-Hochberg correction [37] is applied across all 18 simultaneous tests. To provide algorithm-independent external validation, the Ward's linkage $K = 6$ partition is additionally compared against k-means clustering ($K = 6$) using the Adjusted Rand Index [38].

F. Cluster Visualization

Three complementary visualization strategies are employed. A dendrogram (Fig. 2) visualizes the complete hierarchical amalgamation process with the optimal $K = 6$ cut highlighted. Radar (spider) plots (Fig. 3) display normalized cluster centroid Z-scores across all continuous features simultaneously. Choropleth maps (Fig. 4) encode cluster membership geographically across Japanese prefectures, enabling location-aware decision-making by students with geographic preferences or constraints.

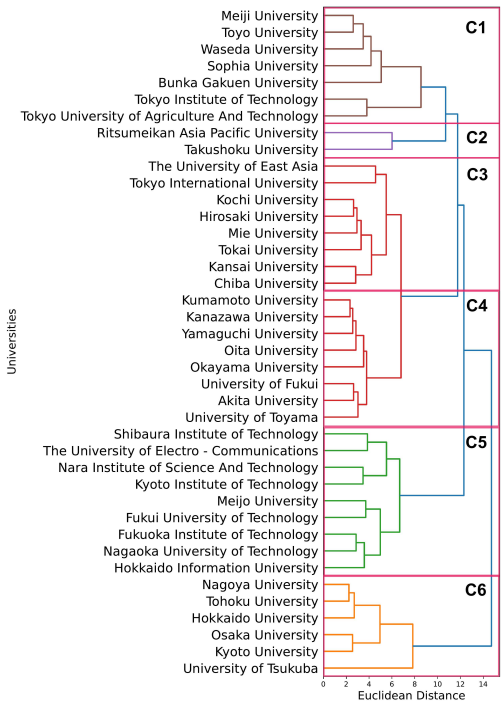

Fig. 2. Dendrogram of hierarchical clustering with Ward's Linkage across 40 Japanese universities. The optimal cut at $K = 6$ is indicated; each resulting cluster corresponds to one institutional archetype.

G. Decision Support Operationalization

The institutional knowledge taxonomy is operationalized as a student-facing decision support tool through a two-artifact delivery mechanism: a Comparative Assessment Matrix that provides a structured reference comparing the six archetypes across five strategic dimensions, and a Strategic Decision Matrix that implements the decision support tool as a binary checklist, where rows represent institutional archetypes, columns represent student decision criteria, and checkmarks indicate primary strengths. These artifacts are presented in detail in Section IV.E.

IV. Experiments and Results

This section presents the results of the hierarchical clustering analysis, internal validation metric evaluation, statistical characterization of cluster distinctiveness, geographic visualization, and synthesis of the institutional knowledge taxonomy into the decision support artifacts.

A. Clustering Analysis and Optimal K Selection

The dendrogram produced by Ward's Linkage hierarchical clustering is presented in Fig. 2. To identify the optimal number of clusters, we evaluated solutions from $K = 2$ to $K = 10$ across four internal validation metrics (Table 3).

Table 3. Internal validation metrics for $K = 2$ to $K = 10$ clustering solutions

K	Inertia	SC	CHI	DBI	Gap	sk
2	533.89	0.21	7.70	1.38	0.51	0.03
3	458.26	0.16	7.42	1.73	0.57	0.03
4	389.24	0.18	7.80	1.64	0.65	0.04
5	331.96	0.20	8.18	1.37	0.72	0.04
6*	**295.44**	**0.22**	**7.98**	**1.24**	**0.76**	**0.04**
7	264.85	0.23	7.83	1.09	0.80	0.04
8	241.75	0.16	7.57	1.22	0.84	0.04
9	219.45	0.17	7.46	1.28	0.86	0.04
10	201.37	0.17	7.30	1.23	0.89	0.04

Inertia = Within-Cluster Sum of Squares (WCSS); SC = Silhouette Coefficient; CHI = Calinski-Harabasz Index; DBI = Davies-Bouldin Index (lower is better); Gap = Gap Statistic; sk = standard error of Gap estimate ($B = 100$ Monte Carlo reference datasets). *$K = 6$ is retained for interpretability; no single internal index uniquely identifies an optimal K (see Section IV.A).

Internal validation metrics alone did not converge on a single optimal solution: the Silhouette Coefficient (SC) and Davies-Bouldin Index (DBI) favored $K = 7$, while the Calinski-Harabasz Index (CHI) peaked at $K = 5$. To resolve this ambiguity, the gap statistic [32] was computed using $B = 100$ Monte Carlo reference datasets generated uniformly over the range of each observed feature. No internal index uniquely identifies a single optimal K. The Silhouette Coefficient and Davies-Bouldin index favor $K = 7$, the Calinski-Harabasz index favors $K = 5$, and the gap-statistic rule, the smallest K such that $\text{Gap}(K) \geq \text{Gap}(K+1) - \text{sk}(K+1)$, is satisfied across the $K = 5$ to $K = 9$ range at the precision reported in Table 3. $K = 6$ is retained because it yields the most interpretable set of institutional archetypes for the decision-support application, capturing distinct combinations of disciplinary focus, research intensity, financial accessibility, and geographic environment that align with the six-cluster structure emerging from the dendrogram. The moderate absolute SC of 0.22 is consistent with the inherent continuity of institutional characteristics; universities span a spectrum rather than forming perfectly discrete categories.

B. Statistical Characterization of Cluster Profiles

Table 4 presents the cluster-level means and standard deviations for all continuous features, along with ANOVA F-statistics and p-values, and Chi-squared test results for categorical features. The $K = 6$ solution demonstrates robust statistical distinctiveness, with 15 of 18 features showing significant between-cluster differences ($p < 0.05$).

Table 4. Cluster profiles and statistical characterization of the $K = 6$ solution via ANOVA and Chi-squared tests

Category	Feature	C1 ($n = 7$)	C2 ($n = 2$)	C3 ($n = 8$)	C4 ($n = 8$)	C5 ($n = 9$)	C6 ($n = 6$)	ANOVA F (p)	p_BH
Institutional Metrics (cluster mean with SD)	Admission Score (/100)	57.57 (8.79)	45.90 (6.65)	47.25 (8.25)	50.90 (2.05)	49.07 (7.34)	61.27 (2.19)	4.97 (<0.002**)	<0.003**
	Research Student Counts	8 (8)	1 (1)	3 (5)	22 (38)	4 (6)	161 (182)	4.68 (<0.002**)	<0.003**
	Publication Counts	166,233 (136,780)	20,354 (7975)	145,394 (128,116)	242,627 (159,180)	53,349 (36,332)	951,993 (429,486)	18.00 (<0.001***)	<0.002**
	Researcher Counts	1418 (1417)	313 (150)	1134 (867)	1602 (865)	316 (193)	6705 (3386)	14.69 (<0.001***)	<0.002**
	Review Score (/5)	4.07 (0.14)	3.79 (0.21)	3.93 (0.09)	3.86 (0.08)	3.90 (0.15)	4.18 (0.03)	8.25 (<0.001***)	<0.002**
	Living Cost (USD/mo.)	1698 (0)	1295 (570)	1132 (89)	1003 (132)	1190 (297)	1089 (114)	11.56 (<0.001***)	<0.002**
Academic Discipline Distribution (% of researchers; cluster mean with SD)	Engineering & Technology	21.13 (10.57)	7.05 (9.47)	11.94 (3.85)	15.59 (4.32)	49.46 (11.89)	16.04 (3.06)	26.73 (<0.001***)	<0.002**
	Natural Sciences	23.38 (5.88)	2.12 (3.00)	16.86 (4.30)	26.87 (6.25)	29.48 (10.97)	24.31 (2.58)	7.08 (<0.001***)	<0.002**
	Health & Medicine	6.55 (4.78)	0.18 (0.25)	22.56 (9.30)	29.52 (6.95)	2.09 (1.63)	24.57 (6.41)	26.57 (<0.001***)	<0.002**
	Social Sciences	16.23 (9.50)	44.48 (14.88)	22.92 (8.82)	14.18 (3.16)	6.62 (5.64)	13.92 (2.07)	11.73 (<0.001***)	<0.002**
	Humanities & Law	14.10 (6.70)	19.79 (11.91)	12.39 (4.90)	4.92 (2.78)	1.93 (4.27)	8.90 (2.19)	9.20 (<0.001***)	<0.002**
	Natural Resources	5.05 (7.35)	0.17 (0.24)	5.16 (3.61)	1.74 (2.08)	1.64 (3.08)	3.01 (2.12)	1.39 (0.25)	0.25
	Business & Management	6.33 (5.24)	24.67 (12.17)	5.54 (3.17)	4.47 (2.36)	4.63 (3.86)	4.54 (1.66)	9.17 (<0.001***)	<0.002**
	Arts & Design	7.21 (5.15)	1.53 (2.16)	2.64 (2.59)	2.72 (2.46)	4.15 (4.70)	4.71 (1.77)	1.72 (0.15)	0.16
Univ. Type (Chi-sq. $p = 0.005$**)	National [n, %]	2, 29%	0, 0%	4, 50%	8, 100%	4, 44%	6, 100%		**0.006****
	Private [n, %]	5, 71%	2, 100%	4, 50%	0, 0%	5, 56%	0, 0%		
IDP Avail. (Chi-sq. $p < 0.001$***)	Availability [n,%]	1, 9%	0	0	2, 18%	2, 18%	6, 55%		**0.002****
ETP Avail. (Chi-sq. $p = 0.01$**)	Availability [n,%]	4, 24%	1, 6%	3, 18%	1, 6%	2, 12%	6, 35%		**0.01***
Tuition Discount (Chi-sq. $p = 0.06$)	Availability [n,%]	4, 31%	2, 15%	2, 15%	1, 8%	1, 8%	3, 23%	n.s.	**0.07**

Values for continuous features are cluster mean with Standard Deviation (SD) in parentheses. For University Type, [n, %] denotes within-cluster count and proportion. For IDP, ETP, and Tuition Discount availability, [n, %] denotes the count and percentage of all program-offering institutions in the cohort contributed by that cluster; the denominator is the total number of institutions offering that specific program across the cohort (11 for IDP, 17 for ETP, 13 for Tuition Discount), not the full cohort of 40, and therefore varies by program. n.s. = not significant. p_BH = Benjamini-Hochberg corrected p-value. **$p < 0.01$; ***$p < 0.001$.

It should be noted that Cluster 2 (Global Gateways) contains only $n = 2$ institutions; its cluster-level statistics are therefore indicative rather than definitive, and C2 is best interpreted as a case study of two highly internationalized private universities rather than a statistically robust population estimate. With only two observations, the reported mean is the arithmetic midpoint of the two institutions and the standard deviation reflects their pairwise difference rather than a population-level property. The potential inflation of between-cluster variance attributable to C2's extreme means is addressed by the Kruskal-Wallis robustness check; the unchanged set of 15 significant features under both parametric and non-parametric testing confirms that cluster distinctiveness is robust to C2's small size.

Among continuous features, Publication Counts ($F = 18.00$, $p < 0.001$), Researcher Counts ($F = 14.69$, $p < 0.001$), and Living Cost ($F = 11.56$, $p < 0.001$) exhibit the largest absolute F-statistics. Admission Score ($F = 4.97$, $p < 0.002$) and Review Score ($F = 8.25$, $p < 0.001$) confirm that academic selectivity and student-perceived campus quality are also structurally differentiated across the six archetypes.

Academic discipline distribution provides particularly strong discrimination. Engineering & Technology ($F = 26.73$, $p < 0.001$) and Health & Medicine ($F = 26.57$, $p < 0.001$) achieve the two highest F-statistics. Social Sciences ($F = 11.73$, $p < 0.001$), Humanities & Law ($F = 9.20$, $p < 0.001$), and Business & Management ($F = 9.17$, $p < 0.001$) also contribute substantially. Three features did not achieve significance: Natural Resources ($F = 1.39$, $p = 0.25$), Arts & Design ($F = 1.72$, $p = 0.15$), and Tuition Discount availability (Chi-sq., $p = 0.06$). The non-significance of Tuition Discount availability (Chi-sq., $p = 0.06$) is substantively informative rather than merely a null result. Unlike disciplinary composition, research capacity, and geographic environment which cluster along predictable institutional dimensions, tuition discount programs show no systematic alignment with cluster profile, as evidenced by the availability counts across clusters (C1: 4, C2: 2, C3: 2, C4: 1, C5: 1, C6: 3). This suggests that tuition concession policies

reflect institution-specific administrative decisions rather than systematic alignment with any particular institutional profile. For international students, this finding implies that tuition discount availability cannot be reliably inferred from cluster membership and should be verified directly with each institution during the application process. Prior to ANOVA, distributional assumptions were assessed via Shapiro-Wilk normality tests [34] and Levene tests for homogeneity of variance [35]. 12 of 14 continuous features violated one or both assumptions; for these features, Kruskal-Wallis non-parametric tests [36] were conducted as a robustness check and confirmed significant between-cluster differences for 10 of 12 violated features, with Natural Resources and Arts & Design remaining non-significant under both ANOVA and Kruskal-Wallis. To address multiple comparison concerns across 18 simultaneous tests, Benjamini-Hochberg correction was applied; the set of significant features remained unchanged at 15 of 18 (p_BH < 0.05), confirming that conclusions are robust to Type I error inflation [37]. For external validation, the Ward's linkage $K = 6$ partition was compared against k-means clustering ($K = 6$) using the Adjusted Rand Index (ARI) [38]. Across 100 k-means runs with independent random initializations, the mean ARI was 0.784 (SD = 0.075, range = 0.601 − 0.964), with all 100 runs achieving ARI ≥0.6, indicating substantial agreement [39] and confirming that the cluster structure is algorithm-independent.

C. Cluster Visualization and Geographic Distribution

The radar plots in Fig. 3 display normalized cluster centroid Z-score profiles across all continuous features, enabling rapid visual comparison of the six institutional archetypes. The geographic distribution in Fig. 4 reveals that cluster membership reflects meaningful spatial patterns consistent with Japan's regional higher education ecosystem.

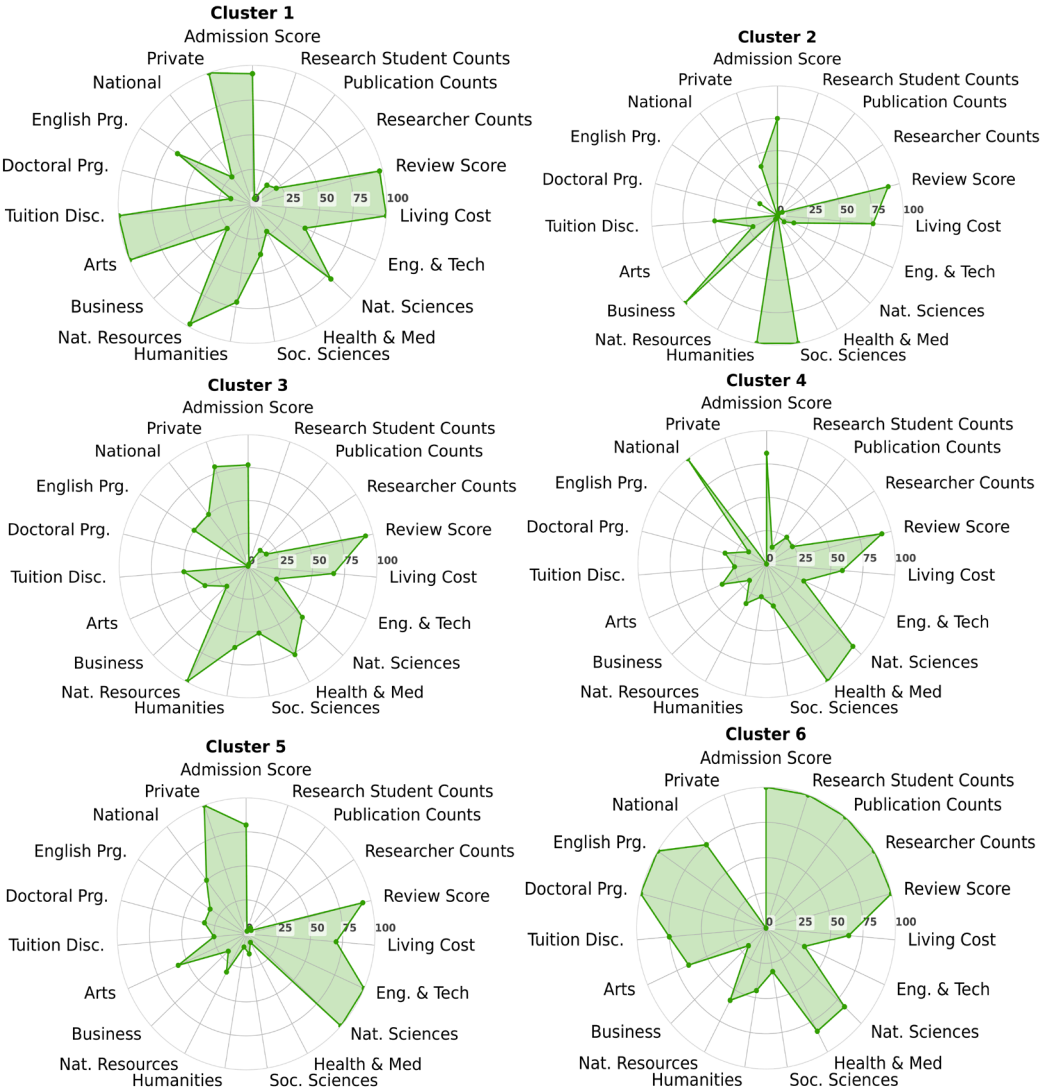

Fig. 3. Radar (spider) plots of normalized cluster centroid Z-score profiles across continuous features for each of the six institutional archetypes.

C1 (Hyper-Urban Core) is entirely concentrated in Tokyo at the highest living cost ($1698/month). C2 (Global Gateways) presents a geographically diverse pairing united by their exceptional internationalization focus (Social Sciences 44.48%, Business & Management 24.67%). C3 (Commuter Belts & Coastal Anchors) spans the Greater Tokyo commuter zone and Kansai suburbs, with balanced disciplinary profiles strong in Health & Medicine (22.56%) and Social Sciences (22.92%). C4 (Traditional Heartland & Snow Country) clusters exclusively National institutions along Japan's Sea of Japan coast offering the lowest living costs ($1003/month). C5 (Golden Route & Cultural Capitals) spans Japan's industrial corridor with the defining characteristic of the highest Engineering & Technology

concentration (49.46%). C6 (Innovation Arch & Regional Powerhouses) groups elite National universities distributed across major regional hubs outside Tokyo.

D. Institutional Archetype Profiles

Drawing on the statistical characterization (Table 4) and visualizations (Figs. 3 and 4), each cluster is assigned a descriptive label reflecting its defining statistically significant characteristics ($p < 0.05$):

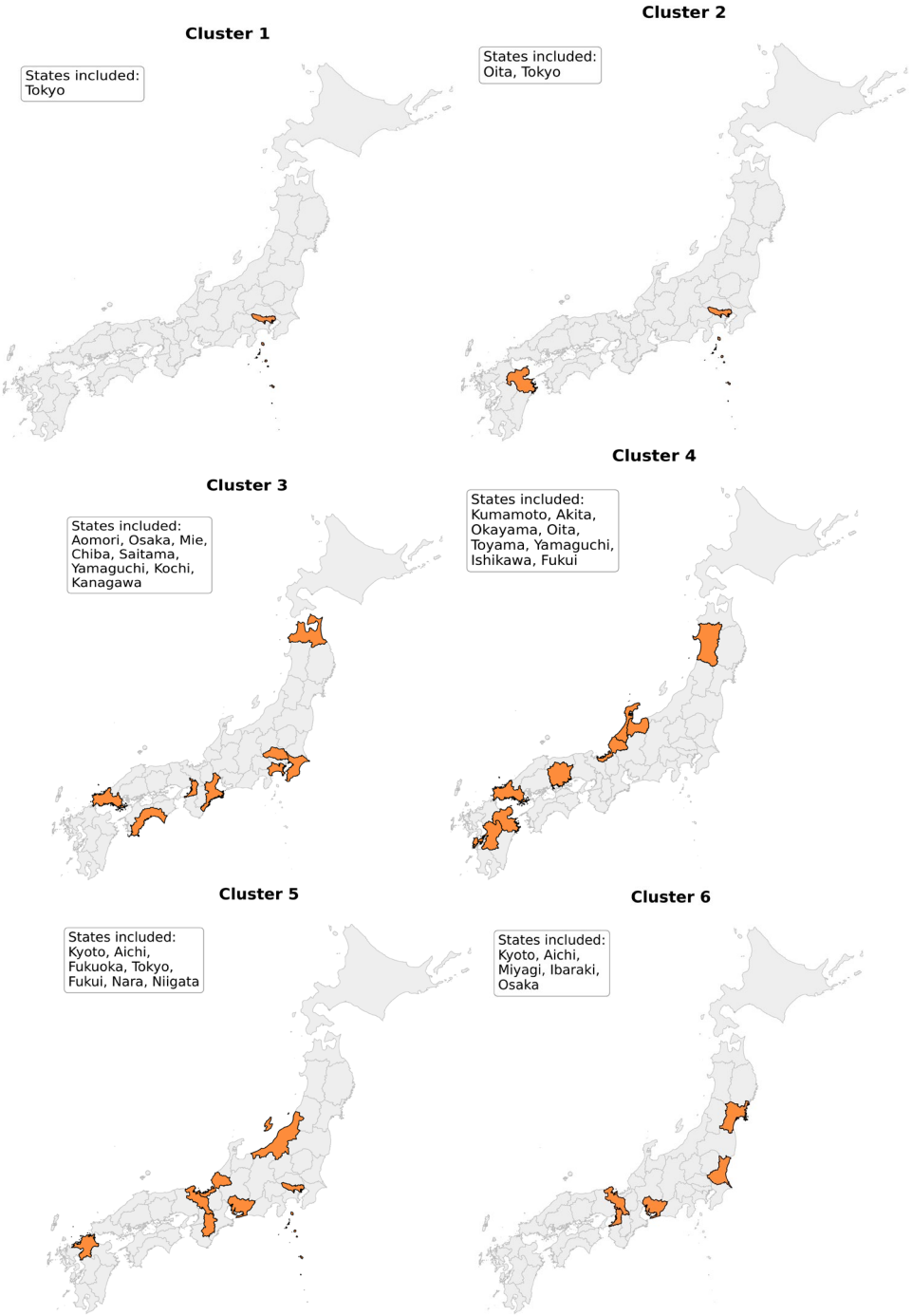

Fig. 4. Choropleth maps of Japan showing the geographic distribution of cluster membership by prefecture.

Cluster 1—Hyper-Urban Core (Assessed Difficulty: Hard-Financial). Seven private and national universities concentrated in Tokyo. Highest living cost ($1698/month). Above-average admission selectivity (57.57/100) and research output. Strong in Natural Sciences (23.38%) and Social Sciences (16.23%). Optimal for students seeking metropolitan professional networks who can manage high living costs.

Cluster 2—Global Gateways (Assessed Difficulty: Easy-Moderate, Niche). Two private universities with the highest internationalization focus: Social Sciences (44.48%), Business & Management (24.67%), Humanities & Law (19.79%). Lowest admission selectivity (45.90/100). Optimal for students prioritizing international academic environments and social sciences.

Cluster 3—Commuter Belts & Coastal Anchors (Assessed Difficulty: Moderate, Balanced). Eight institutions across suburban Tokyo, Kansai, and coastal prefectures. Moderate cost ($1132/month). Balanced disciplinary profile with notable concentrations in Health &

Medicine (22.56%) and Social Sciences (22.92%). Optimal for students seeking affordable suburban environments with broad academic programs.

Cluster 4—Traditional Heartland & Snow Country (Assessed Difficulty: Moderate, Value). Eight exclusively National universities in regional prefectures. Lowest living cost ($1003/month). Strongest concentration in Health & Medicine (29.52%) and Natural Sciences (26.87%). Optimal for students prioritizing cost efficiency and biomedical or natural science disciplines.

Cluster 5—Golden Route & Cultural Capitals (Assessed Difficulty: Moderate, Tech). Nine institutions spanning Japan's industrial and cultural corridor. Defining characteristic: Engineering & Technology concentration of 49.46%, the highest in the cohort by a substantial margin. Moderate cost ($1190/month). Optimal for students targeting STEM careers in Japan's industrial and technology sector.

Cluster 6—Innovation Arch & Regional Powerhouses (Assessed Difficulty: Very Hard, Elite). Six exclusively National universities forming Japan's elite research network outside Tokyo (Tsukuba, Kyoto, Osaka, Hokkaido, Tohoku, Nagoya). Highest admission scores (61.27/100), researcher counts (6705), and publication volumes (951,993). Highest share of IDP-offering institutions in the cohort (55%) and ETP-offering institutions (35%). Optimal for students with strong academic profiles targeting doctoral programs and research careers.

E. Decision Support Artifacts

The six institutional archetypes are synthesized into two complementary decision support artifacts. The Comparative Assessment Matrix (Table 5) enables rapid cross-archetype comparison across five strategic dimensions. The Strategic Decision Matrix (Table 6), the core decision support artifact, operationalizes each archetype's primary strengths as a binary checklist that students self-apply to identify their optimal institutional target cluster.

Table 5. Comparative institutional archetype assessment matrix

Dimension	C1: Hyper-Urban	C2: Global Gateways†	C3: Commuter	C4: Heartland	C5: Golden Route	C6: Innovation
Academic Core	Generalist / Research	Social Sci. / Humanities	Balanced / Health	Health / Natural Sci.	Engineering-Dominant	Elite Research
Financial Tier	High ($1698)	High ($1295)	Moderate ($1132)	Value ($1003)	Moderate ($1190)	Moderate-High ($1089)
Environment	Hyper-Urban (Tokyo)	Urban / Resort Duality	Suburban / Coastal	Nature / Snow Country	Industrial / Historical	Major Regional Hub
Intl. Programs	ETP: moderate	ETP: high; IDP: low	ETP: moderate	IDP: moderate	IDP: moderate	IDP & ETP: highest
Best For	Prestige & Networking	Intl. Experience	Balanced Lifestyle	Budget & Nature	STEM/Tech Careers	Top-tier Research

† C2 comprises two institutions only (Ritsumeikan Asia Pacific University and Takushoku University) and should be interpreted as a two-university case study rather than a generalizable archetype.

Table 6. Strategic decision matrix: Checklist for international applicant self-assessment

Cluster	Academic Profile			Financial		Environment		Req. Traits		Strategic Fit		
	Research	STEM	Soc/Arts	High $	Value $	Urban	Nature	Adaptability	Agility	Elite	Career	Niche
C1: Hyper-Urban	✓		✓	✓		✓		✓	✓	✓	✓	
C2: Global Gateways†			✓	✓		✓		✓				✓
C3: Commuter			✓		✓	✓						
C4: Heartland		✓			✓		✓	✓				✓
C5: Golden Route		✓			✓				✓		✓	
C6: Innovation	✓	✓			✓	✓			✓	✓		

† Selecting C2 is equivalent to directly considering Ritsumeikan Asia Pacific University and Takushoku University specifically. (✓) indicates a primary strength of the archetype on that decision criterion. Required Traits and Strategic Fit columns are defined as follows. Required Traits: Adaptability = comfort with unfamiliar linguistic and cultural environments; Agility = preference for broad multi-disciplinary programs. Strategic Fit: Niche = clearly defined interest in social sciences, humanities, or international studies; Elite = strong academic profile suited to selective admission and doctoral pathways; Career = clear industry orientation, particularly in STEM. To use this matrix: (1) identify only the columns that match your personal academic profile, financial situation, environment preference, required traits, and strategic goals; (2) select the archetype row that covers the highest proportion of your identified columns. Do not simply count total ✓ marks across all columns.

V. Discussion

A. The EDM Framework for International Student Decision Support

The central contribution of this study is the integration of an EDM analytical pipeline with a decision support mechanism for international university selection. The $K = 6$ institutional knowledge taxonomy reveals a structural landscape that transcends the single prestige axis of conventional ranking systems. The six archetypes represent empirically distinct combinations of research intensity, disciplinary specialization, financial accessibility, and geographic environment that are directly relevant to student decision-making but invisible in standard ranking outputs.

The Strategic Decision Matrix (Table 6) operationalizes this insight as a student-facing decision support tool that delivers structured, actionable guidance at the point of decision without requiring professional counseling expertise. Consider a Thai engineering student with a $1200/month

budget seeking English-medium instruction. The matrix immediately identifies C5 (Golden Route) as the optimal archetype: moderate cost, the highest Engineering and Technology concentration (49.46%), and a direct STEM-career pathway. This rapid filtering reduces the decision space from 40 universities to 9 specifically matched institutions, transforming information overload into a manageable, reasoning-supported shortlist.

The question of whether the matrix demonstrably improves student decision quality warrants consideration. In the absence of a prospective user study, decision quality can be assessed conceptually along two dimensions. First, process quality: the matrix replaces informal, unstructured peer advice with a transparent, criteria-based filtering mechanism that externalizes implicit preferences and reduces susceptibility to prestige bias. Second, outcome alignment: students whose selected cluster matches their stated academic, financial, and lifestyle profile are more likely to shortlist institutions that genuinely fit their needs rather than institutions that are merely visible. Empirical validation of these claims through controlled user studies comparing matrix-assisted versus unassisted selection decisions is identified as a primary direction for future research.

B. Statistical Distinctiveness and Methodological Validity

The statistical characterization confirms that the six clusters represent structurally distinct institutional types. The significance of 15 of 18 features ($p < 0.05$) across complementary test types (ANOVA for continuous, Chi-squared for categorical) provides multi-dimensional evidence of cluster validity, with conclusions remaining stable under non-parametric Kruskal-Wallis testing and Benjamini-Hochberg correction across all 18 simultaneous tests. It is important to note that the same feature set informs both cluster formation and subsequent statistical characterization, meaning the tests measure cluster structural properties rather than providing independent external validation. This limitation is directly addressed by the algorithm-independent external validation reported in Section IV.B: across 100 *k*-means runs with independent random initializations, the mean ARI was 0.784 (SD = 0.075), with all 100 runs achieving ARI $\geq$ 0.6, confirming that the cluster structure is not an artifact of Ward's linkage. The consistency of significance across the full multi-dimensional feature space provides substantial confidence that the cluster structure reflects genuine heterogeneity in the Japanese university landscape.

C. Robustness, Generalizability, and Scalability

The framework's reproducibility is grounded in its exclusive reliance on publicly accessible institutional data [24–27], though data sources vary in their degree of standardization: while researchmap and MEXT provide structured, machine-readable outputs, Daigaku-hensati.com is a Japanese-language commercial website and Numbeo is a user-contributed crowdsourced database. The robustness of the analytical pipeline is further supported by three complementary checks. First, the $K = 6$ partition demonstrates stability under dimensionality reduction: PCA robustness checks across three standard component-retention criteria yielded ARI = 0.916 under the Kaiser criterion and ARI = 1.000 under both the 80% and 95% cumulative variance thresholds, confirming that the cluster structure is not an artifact of high-dimensional Euclidean geometry. Second, the partition is algorithm-independent, as confirmed by external validation against *k*-means clustering (mean ARI = 0.784 across 100 runs). Third, cluster distinctiveness conclusions are robust to multiple comparison inflation, as the set of 15 significant features remained unchanged after Benjamini-Hochberg correction. The data-agnostic methodology is designed to generalize beyond the Japan-Thailand context: the clustering pipeline can be applied to any higher education system for which comparable multi-dimensional open institutional data exists, substituting Japan-specific sources with equivalent sources from other national higher education systems.

D. Social Impact and Equity Implications

A design goal of this framework is the democratization of high-quality decision support previously accessible only through expensive international education consulting services. By surfacing high-performing regional National universities in C3 and C4 as competitive alternatives to metropolitan institutions, with lower costs, strong academic profiles, and authentic Japanese cultural experiences, the decision matrix shifts the student knowledge base away from the metropolitan prestige bias that currently dominates informal Thai student guidance networks. This has equity implications: students from lower-income backgrounds who cannot afford C1 ($1698/month) are guided toward C4 (the lowest-cost National university profile at $1003/month) with full evidence of the cluster's academic quality profile.

VI. Limitations

Several boundary conditions of this study should be acknowledged. First, several universities (notably Tokai University) operate dispersed multi-campus networks across different prefectures. Living cost estimates reflect the institution's primary campus location; students enrolling at satellite campuses may encounter substantially different financial and environmental conditions. Campus-level verification is recommended before enrollment decisions are finalized.

Second, researchmap data coverage may introduce systematic institutional bias. Participation in researchmap is mandatory at some institutions and voluntary at others; institutions with lower faculty entry rates will have systematically underestimated research capacity metrics.

Third, Nara Institute of Science and Technology (NAIST), assigned to C5 (Golden Route), is a graduate-only institution offering no undergraduate degree programs. The clustering algorithm correctly identifies its structural similarity to engineering-specialized institutions, but students seeking undergraduate programs should treat NAIST's C5 membership as applicable to postgraduate applicants only.

Fourth, the cohort is bounded by the JSPS Bangkok Office registry [4], which captures 40 of the hundreds of Japanese universities Thai students might consider. While this boundary ensures verified institutional relevance to the Thai-Japan context, findings should not be extrapolated to the broader Japanese higher education system without reconducting the analysis on an expanded cohort.

Fifth, living cost estimates are derived from city-level

Numbeo figures and therefore do not capture intra-urban variation. This limitation is most visible in Cluster 1, where all seven Tokyo-based institutions share an identical value of $1698/month, producing the SD = 0 reported in Table 4. In practice, costs near Waseda University in Shinjuku differ from those near Tokyo University of Agriculture and Technology in Koganei. While this granularity limitation does not materially affect the clustering outcome, students should treat living cost figures as prefecture-level benchmarks and verify campus-specific costs before making enrollment decisions. Additionally, Numbeo is a user-contributed crowdsourced database and does not undergo the same verification processes as official government or institutional data sources. Living cost estimates should therefore be interpreted as indicative community-reported figures rather than authoritative measurements.

Finally, while Z-score normalization reduces the geometric impact of mixed feature types, Gower distance, which applies type-appropriate sub-metrics per feature, represents a theoretically preferable alternative for datasets combining continuous and binary variables and is recommended as a direction for future work.

VII. Conclusions

This study addresses a recognized gap at the intersection of Educational Data Mining, learning analytics, and ICT in education: the absence of a data-driven, student-facing decision support tool for international university selection. By applying hierarchical clustering to a multi-dimensional dataset of 40 Japanese universities with Thai institutional ties, we derive six empirically grounded institutional archetypes that capture the structural heterogeneity of the Japanese higher education landscape. These archetypes are operationalized as a student-facing decision support tool, specifically the Strategic Decision Matrix, that enables students to systematically align their individual profiles with the appropriate institutional cluster, reducing decision complexity from 40 undifferentiated options to a targeted, reasoning-supported shortlist without requiring expert advisory access.

Cluster validity is established through a multi-layered evidence framework: 15 of 18 features achieve statistical significance under both parametric and non-parametric testing, with conclusions stable after Benjamini-Hochberg correction; algorithm-independent external validation against *k*-means clustering yields a mean ARI of 0.784 across 100 runs; and PCA robustness checks confirm cluster stability across three dimensionality reduction criteria (ARI $\geq$ 0.916).

The framework demonstrates that EDM clustering methods, established in the IJIET literature by works such as Nuankaew *et al.* [11] for student-level profiling, can be productively applied to institutional-level data to generate actionable knowledge representations for educational decision support. The methodology is reproducible, scalable, and data-agnostic, providing a blueprint for deployment across other national higher education systems. By delivering structured, data-backed guidance at the point of selection, the matrix exemplifies a knowledge-centered, human-oriented approach to electronic performance support in international higher education.

Future research will pursue two primary directions. First, the integration of Principal Component Analysis (PCA) as a pre-clustering dimensionality reduction step was evaluated as a robustness check and confirmed cluster stability across all retention criteria; future work will explore PCA as a full pre-clustering pipeline to further improve cluster boundary precision. Second, the development of an interactive web-based implementation of the decision matrix will deploy the institutional knowledge framework as a live digital tool within higher education guidance systems, enabling prospective validation of the framework's impact on student decision quality and satisfaction.

Data Availability

The original data presented in the study are openly available in Zenodo at https://doi.org/10.5281/zenodo.18073129.

Appendix

Table A1. University cluster membership details

Cluster Name	University Name	Type	Prefecture
Cluster 1: Hyper-Urban Core	Meiji University	Private	Tokyo
	Toyo University	Private	Tokyo
	Waseda University	Private	Tokyo
	Sophia University	Private	Tokyo
	Bunka Gakuen University	Private	Tokyo
	Tokyo Institute of Technology	National	Tokyo
	Tokyo Univ. of Agriculture and Technology	National	Tokyo
Cluster 2: Global Gateways	Ritsumeikan Asia Pacific University	Private	Oita
	Takushoku University	Private	Tokyo
Cluster 3: Commuter Belts & Coastal Anchors	The University of East Asia	Private	Yamaguchi
	Tokyo International University	Private	Saitama
	Kochi University	National	Kochi
	Hirosaki University	National	Aomori
	Mie University	National	Mie
	Tokai University	Private	Kanagawa
	Kansai University	Private	Osaka
	Chiba University	National	Chiba
Cluster 4: Traditional Heartland & Snow Country	Kumamoto University	National	Kumamoto
	Kanazawa University	National	Ishikawa
	Yamaguchi University	National	Yamaguchi
	Oita University	National	Oita
	Okayama University	National	Okayama
	University of Fukui	National	Fukui

	Akita University	National	Akita
	University of Toyama	National	Toyama
Cluster 5: Golden Route & Cultural Capitals	Shibaura Institute of Technology	Private	Tokyo
	The Univ. of Electro-Communications	National	Tokyo
	Nara Inst. of Science and Technology†	National	Nara
	Kyoto Institute of Technology	National	Kyoto
	Meijo University	Private	Aichi
	Fukui University of Technology	Private	Fukui
	Fukuoka Institute of Technology	Private	Fukuoka
	Nagaoka Univ. of Technology	National	Niigata
	Hokkaido Information University	Private	Hokkaido
Cluster 6: Innovation Arch & Regional Powerhouses	University of Tsukuba	National	Ibaraki
	Kyoto University	National	Kyoto
	Osaka University	National	Osaka
	Hokkaido University	National	Hokkaido
	Tohoku University	National	Miyagi
	Nagoya University	National	Aichi

† Graduate-only institution. NAIST offers no undergraduate degree programs; its C5 membership applies to postgraduate applicants only.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

T.C. and N.I.N.A. conducted data collection, pre-processing, and clustering analysis; S.T. designed the research framework and supervised the study; T.C. wrote the initial manuscript; N.I.N.A. and S.T. reviewed and revised the manuscript; all authors have read and approved the final version.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support received from International SciKU Branding (ISB), Faculty of Science, Kasetsart University. We thank Phatthakarn Thanaphongphirom from the Department of Statistics, Faculty of Science, Kasetsart University, for providing insightful statistics knowledge and guidance. Finally, we sincerely appreciate Suparat Patcharaworakul for their support to the first author throughout this project.

REFERENCES

[1] Ministry of Education, Culture, Sports, Science and Technology (MEXT). Ministry of Education, Culture, Sports, Science and Technology. [Online]. Available: https://www.mext.go.jp/

[2] Japan Society for the Promotion of Science (JSPS). Japan society for the promotion of science. [Online]. Available: https://www.jsps.go.jp/english/

[3] Japan Student Services Organization (JASSO). Japan student services organization. [Online]. Available: https://www.jasso.go.jp/en/

[4] Japan Society for the Promotion of Science Thai Office. JSPS Bangkok Office: Japanese university in Thailand. [Online]. Available: https://jsps-th.org/japanese_univ/

[5] R. Kler, S. Goel, I. S. Ray, A. B. B. A. Hamid, and N. I. B. Ya'akub, "Inimitable elements impacting student mobility among developing countries and developed countries," *International Journal of Professional Business Review*, vol. 8, no. 6, e02223, 2023.

[6] R. Sepehr, M. Shekofteh, A. Shahbodaghi, and S. Jambarsang, "Examining the correlation between presence in university ranking systems with countries' income and R&D expenditure," *DESIDOC Journal of Library & Information Technology*, vol. 45, no. 3, pp. 217–223, 2025.

[7] Y. Sato, "A case of policy evaluation utilizing a logical framework: Evaluation of Japan's foreign student policy towards Thailand," *Evaluation*, vol. 11, no. 3, pp. 351–378, 2005.

[8] Y. Sato, "A trial of impact evaluation utilizing effect size statistics: Its application to the evaluation of Japan's foreign student policy," *Evaluation*, vol. 18, no. 2, pp. 230–245, 2012.

[9] D. Gasper, "Evaluating the 'Logical Framework Approach' towards learning-oriented development evaluation," *Public Administration and Development*, vol. 20, no. 1, pp. 17–28, 2000.

[10] H. Tomita and T. Mochizuki, "Thai students' preferences in studying abroad in Japan and the Japanese government's policy on accepting foreign students in universities," *JSN Journal*, vol. 6, no. 1, pp. 132–146, 2016.

[11] P. Nuankaew, T. Sittiwong, and W. S. Nuankaew, "Characterization clustering of educational technologists achievement in higher education using machine learning analysis," *International Journal of Information and Education Technology*, vol. 12, no. 9, pp. 881–888, 2022. doi: 10.18178/ijiet.2022.12.9.1697

[12] G. Matlis, N. Dimokas, and P. Karvelis, "Unveiling university groupings: A clustering analysis for academic rankings," *Data*, vol. 9, no. 5, 67, 2024.

[13] M. Maral, "Evaluation model for research tendencies and performance of universities," *Scientometrics*, vol. 131, no. 1, pp. 501–532, 2026. doi: 10.1007/s11192-026-05543-y

[14] STUDY in JAPAN. STUDY in JAPAN: Comprehensive guide for studying in Japan. [Online]. Available: https://www.studyinjapan.go.jp/en/

[15] Kawaijuku Educational Information Network. Kei-Net Plus: University search system. [Online]. Available: https://search.keinet.ne.jp/search/option/

[16] A. Abbott and M. Silles, "Determinants of international student migration," *The World Economy*, vol. 39, no. 5, pp. 621–635, 2016.

[17] M. Radovanović, S. Jovčić, A. Petrovski, and E. Cirkin, "Evaluation of University Professors Using the Spherical Fuzzy AHP and Grey MARCOS Multi-criteria Decision-Making Model: A Case Study," *Spectrum of Decision Making and Applications*, vol. 2, no. 1, pp. 197–217, 2025. doi: 10.31181/sdmap21202518

[18] M. B. Bouraima, I. Badi, O. O. Makinde, E. A. Kassembe, and F. Muya, "Prioritizing Strategies for the Development and Sustainability of Higher Education in Developing Countries: A FullEX MCDM Approach," *Spectrum of Decision Making and Applications*, vol. 2, no. 1, pp. 166–176, 2025. doi: 10.31181/sdmap21202516

[19] T. Basuri, K. H. Gazi, S. G. Das, and S. P. Mondal, "Ranking Higher Education Institutions Using Entropy-VIKOR with Generalized Pentagonal Intuitionistic Fuzzy Numbers," *Journal of Contemporary Decision Science*, vol. 2, no. 1, pp. 64–83, 2026.

[20] M. M. Sangaré-Oumar, S. Qian, Y. Qiu, M. B. Bouraima, I. Badi, and N. A. Tengecha, "Evaluation of Factors Influencing University Graduates' Unemployment in Developing Country: A Multi-criteria Decision-making Perspective," *Spectrum of Operational Research*, vol. 2, no. 1, pp. 192–201, 2025. doi: 10.31181/sor21202516

[21] M. Subotic, M. Maric, B. Dudic, S. M. Veljkovic, D. Celic, and A. Peric, "Influence of Educational Factors on the Entrepreneurial Potential of Students as a Prerequisite for Economic Development," *International Journal of Economic Sciences*, vol. 15, no. 1, pp. 231–249, 2026. doi: 10.31181/ijes1512026238

[22] J. Kannan, V. Jayakumar, and M. Pethaperumal, "Advanced Fuzzy-based Decision-making: The Linear Diophantine Fuzzy CODAS Method for Logistic Specialist Selection," *Spectrum of Operational Research*, vol. 2, no. 1, pp. 41–60, 2025. doi: 10.31181/sor2120259

[23] S. K. Sahoo, S. Sahu, B. B. Choudhury, P. R. Dhal, and I. Dhar, "Transforming Decision-making in Generation Z Education: A Systematic and Bibliometric Review of Artificial Intelligence," *Journal of Contemporary Decision Science*, vol. 2, no. 1, pp. 114–130, 2026.

[24] Japan Science and Technology Agency (JST). Japan Science and Technology Agency. [Online]. Available: https://www.jst.go.jp/

[25] Daigaku Hensati. [Online]. Available: https://daigaku-hensati.com/
[26] Minkou. [Online]. Available: https://www.minkou.jp/
[27] Numbeo: The world's largest database of user contributed data about cities and countries worldwide. [Online]. Available: https://www.numbeo.com/
[28] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.
[29] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
[30] T. Caliński and J. A. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics—Theory and Methods*, vol. 3, pp. 1–27, 1974.
[31] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.
[32] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001. doi: 10.1111/1467-9868.00293
[33] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When Is 'Nearest Neighbor' Meaningful?" in *Database Theory—ICDT'99*, C. Beeri and P. Buneman, Eds. Berlin, Heidelberg: Springer, 1999, pp. 217–235. doi: 10.1007/3-540-49257-7_15
[34] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3–4, pp. 591–611, 1965. doi: 10.1093/biomet/52.3-4.591
[35] H. Levene, "Robust tests for equality of variances," in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin *et al.*, Eds. Stanford, CA: Stanford University Press, 1960, pp. 278–292.
[36] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-criterion Variance Analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. doi: 10.2307/2280779
[37] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x
[38] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985. doi: 10.1007/BF01908075
[39] D. Steinley, "Properties of the Hubert-Arabie Adjusted Rand Index," *Psychological Methods*, vol. 9, no. 3, pp. 386–396, 2004. doi: 10.1037/1082-989X.9.3.386

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).