SINERGI Vol. 30, No. 3, October 2026: 687-706
http://publikasi.mercubuana.ac.id/index.php/sinergi
http://doi.org/10.22441/sinergi.2026.3.002

Hybrid KMeans–HDBSCAN clustering of Construction Cost Index dynamics for adaptive infrastructure planning in Indonesia

Eddy Suprihadi[1], Sri Hastutik[2]*
[1]Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia
[2]Economic and Business, Universitas Wisnuwardhana, Indonesia

Abstract
Previous studies on regional construction cost estimation in Indonesia have predominantly relied on uniform national indices or conventional econometric models, which often fail to capture spatial heterogeneity and evolving temporal trends. This study introduces a data-driven hybrid clustering framework that integrates KMeans and HDBSCAN to classify Indonesian districts and cities using Construction Cost Index (CCI) dynamics from 2014 to 2024, with economic plausibility assessed using Gross Regional Domestic Product (GRDP). Empirically, KMeans yields three interpretable macro-clusters, but produces a highly imbalanced partition in which the dominant cluster contains 474 districts (over 90% of observations), indicating limited sensitivity to localized irregularities. In contrast, HDBSCAN identifies 35 micro-clusters and flags 103 districts as noise, explicitly isolating anomalous trajectories that centroid-based clustering may absorb. Cluster validity is supported by statistically significant between-cluster differences in CCI (one-way ANOVA on 2020 CCI: KMeans F = 1696.49, p = 2.35 × 10 $^{-226}$*; HDBSCAN F = 370.91, p = 9.22 × 10* $^{-267}$*, excluding noise). The temporal profiles reveal heterogeneous cost regimes, including persistently high-cost zones, transitional regions exhibiting structural correction, and stable low-cost areas. By exposing these hidden disparities and providing an anomaly-sensitive zoning structure, the proposed framework supports differentiated budgeting, targeted subsidies, and adaptive infrastructure planning for geographically diverse settings such as Indonesia.*

Keywords:
CCI;
Clustering;
Disparity;
HDBSCAN;
KMeans;

Article History:
Received: June 4, 2025
Revised: March 4, 2026
Accepted: March 12, 2026
Published: October 2, 2026

Corresponding Author:
Sri Hastutik
Economic and Business, Universitas Wisnuwardhana, Indonesia
Email:
srihastutik@wisnuwardhana.ac.id

INTRODUCTION

Infrastructure development plays a crucial role in advancing national economic growth by facilitating the movement of goods, services, and people, while fostering social welfare and regional integration [1][2]. In Indonesia, however, equitable distribution of infrastructure remains a persistent challenge due to the country's archipelagic geography and socioeconomic heterogeneity [3, 4, 5]. National budgeting frameworks often rely on uniform construction cost assumptions, which fail to accommodate variations in logistical complexity, economic structure, and local price dynamics [5][6]. This generalized approach can lead to inefficient allocation of resources and exacerbate inequalities in infrastructure provision [7].

In empirical terms, the scale of regional construction cost disparity is substantial [8]. Using the CCI trajectories analyzed in this study, high-cost zones exhibit average CCI values above 340 and peak at 365.79 (2018), whereas stable low-cost zones remain largely within the 85-115 range across years. This implies a multi-fold cost differential between districts operating under extreme logistical or market constraints and those with relatively predictable cost environments. When national budgeting relies on uniform baselines, such disparities can translate into systematic under-budgeting in high-cost regions, potentially increasing the risk of procurement delays and scope adjustments, while simultaneously over-allocating resources in lower-cost regions where comparable outputs could be achieved with smaller budget envelopes. These empirical contrasts highlight why an adaptive, region-sensitive zoning framework is required for equitable and efficient infrastructure planning. Expressed as a relative gap, a high-cost cluster average above 340 compared with a low-cost baseline around 110 corresponds to an increase of approximately 209%, with peak values (e.g., 365.79 in 2018) reaching roughly 233% above the low-cost baseline.

In practical terms, reliance on uniform construction cost assumptions can generate tangible implementation and equity risks across Indonesia's diverse regions. When regional cost conditions are under-estimated in remote or logistically constrained districts, infrastructure programs may face procurement delays, scope reductions, and budget overruns as projects compete with local supply constraints and higher transportation burdens [9][10]. Conversely, over-estimation in relatively lower-cost districts can reduce allocative efficiency by diverting limited public funds away from areas with greater infrastructure needs [11]. Over time, these distortions may weaken both the effectiveness of public investment and the equity of service provision, particularly in an archipelagic context where market access, geography, and local price dynamics vary substantially across districts [11][12]. This motivates the need for an adaptive, data-driven approach to cost zoning that can represent broad regional structures while also flagging localized anomalies relevant for differentiated planning and budgeting.

The availability of longitudinal Construction Cost Index (CCI) data provides a valuable empirical foundation for addressing these inefficiencies [13]. Despite this, infrastructure planning in Indonesia continues to depend on static cost assumptions, overlooking both spatial and temporal variability in cost behaviour [14]. This disconnect between available data and its application in policy underscores the need for a more adaptive, data-driven approach to infrastructure cost zoning.

Various methodological approaches have been employed in prior research to address construction cost estimation and regional cost disparities. AlTalhoni et al. [13] applied econometric and time-series forecasting to predict CCI values, producing accurate short-term projections but offering limited insight into spatial heterogeneity. Aslam et al. [14] explored artificial intelligence-based forecasting for national-level construction costs but did not account for localized variations critical to regional planning. More broadly, recent evidence suggests that data-driven analytics can strengthen policy effectiveness by improving decision-making quality and optimizing resource allocation, reinforcing the relevance of integrating computational approaches into infrastructure planning [15]. Saeidlou and Ghadiminia [16] proposed a deep neural network (DNN) framework that achieved high predictive accuracy, yet its lack of transparency limited its applicability for policy decision-making. In the clustering domain, Balaram et al. [17] introduced a scalable two-level clustering framework to extract electric vehicle (EV) usage patterns from long time-series profiles, demonstrating how temporal segmentation and multi-stage grouping can reveal behaviourally distinct cohorts for operational decision support. Enes et al. [18] similarly proposed a configurable pipeline for clustering multivariate time-series data through feature extraction, dimensionality reduction, and unsupervised clustering (including KMeans), with an optional outlier-detection step to isolate anomalous patterns. Together, these studies highlight the practical value of centroid-based and feature-driven clustering for structuring complex temporal signals; however, they also suggest that heterogeneous real-world data often require complementary methods that can better accommodate irregular cluster geometry and explicit noise handling. Saklani et al. [19] applied clustering to electric vehicle infrastructure siting, highlighting the benefits of spatial data segmentation, but without integrating temporal dynamics or economic validation.

More advanced density-based methods, such as HDBSCAN [10], have proven effective in identifying non-spherical clusters and explicitly handling noise across diverse applied settings, while recent implementations have further strengthened their practicality through improved computational efficiency and fast prediction mechanisms for unseen data points [20]. Nevertheless, prior methodological work on KMeans has shown that, despite improvements in initialization, cluster-number selection, and computational efficiency, centroid-based partitioning still tends to perform best under relatively convex and well-separated cluster structures and remains less robust when data exhibit irregular geometries, local density variation, or noise [21], reinforcing the need for density-sensitive alternatives and hybrid designs.

Beyond forecasting-oriented construction cost studies, a related stream of research concerns construction cost zoning and regionalization, where the goal is to delineate geographically meaningful cost zones to support differentiated budgeting, benchmarking, and policy implementation. In this domain, zoning is typically motivated by persistent spatial frictions such as market access, transportation and logistics constraints, and regional differences in construction supply chains, which imply that national averages may systematically misrepresent local cost realities. Existing zoning efforts often rely on administrative grouping, descriptive cost indices, or single-method clustering to produce regional typologies; however, many approaches remain limited in their ability to represent temporal cost trajectories and to explicitly isolate irregular patterns and noise that are often most policy-relevant.

Table 1. Summary of related studies and critical appraisal

Study	Methodology	Main Contribution	Limitation / Critique	SOTA Positioning (Relevance to this study)
AlTalhoni et al. [13]	Econometric and time-series forecasting	Accurate short-term prediction of construction cost indices	Limited insight into spatial heterogeneity	Strong SOTA for CCI forecasting literature; indirectly relevant, but not designed for spatial cost zoning or hybrid clustering
Aslam et al. [14]	Artificial intelligence models for cost forecasting	National-level predictive framework	Neglected localized variations essential for regional planning	SOTA for AI-driven CCI prediction; complementary background, but not addressing spatial clustering or policy zoning
Saeidlou & Ghadiminia [16]	Deep Neural Network (DNN) framework	High predictive accuracy	Lack of interpretability, limiting policy applicability	SOTA for deep learning cost estimation; relevant methodologically, but not aligned with interpretability and zoning objectives
Balaram et al. [17]	Two-level scalable clustering of long EV time-series profiles (temporal segmentation + graph-based clustering)	Revealed distinct EV usage/charging behaviour patterns from long time-series data; scalable extraction of behavioural cohorts	Domain-specific (EV usage); not focused on spatial zoning or construction costs	SOTA example of multi-stage clustering for long time-series, supporting the need for temporal-aware grouping beyond static clustering
Enes et al. [18]	Configurable multivariate time-series clustering pipeline (feature extraction + dimensionality reduction + unsupervised clustering incl. KMeans) with optional outlier detection; scalable with Spark	Demonstrated an end-to-end unsupervised pipeline for clustering multivariate time series and isolating anomalies at scale	Application in HPC monitoring; does not address spatial interpretability or policy-facing zoning	SOTA methodological reference for feature-based time-series clustering and outlier handling, motivating anomaly-sensitive clustering components (e.g., HDBSCAN) in heterogeneous settings
Saklani et al. [19]	Cluster analysis for EV infrastructure siting	Showcased clustering for spatial decision-making	Did not incorporate temporal dynamics or economic validation	Applied SOTA for spatial clustering in planning problems; useful comparator, but lacks temporal trajectories + economic plausibility checks central to your study
Abdulnassar et al. [21]	Enhanced KMeans (seed selection + silhouette-based k + Kmeans9+)	Improved efficiency and stability of KMeans clustering	Retains centroid-based limitations for irregular clusters and noise	SOTA reference for practical KMeans improvements; supports need for density-based complements in heterogeneous settings
Carrasco & Hernandez-del-Valle [22]	HDBSCAN for anomaly detection in economic analysis	Effective at identifying non-spherical clusters and outliers	Rarely applied to regional construction cost zoning	Strong SOTA reference for HDBSCAN in socio-economic data; directly supports your use of density-based clustering for irregular clusters and noise interpretation

Despite the demonstrated effectiveness of density-based clustering in other domains [22][23], its systematic adoption for construction cost zoning remains rare. A concise summary of these related studies, including their methodological approaches, key contributions, and limitations, is presented in Table 1 to provide a clear comparative overview of the existing literature and highlight the research gap addressed in this study.

Despite the growing use of data-driven techniques for construction cost analysis, the existing literature leaves several persistent gaps for policy-facing cost zoning in Indonesia. Most studies remain forecasting-oriented or cross-sectional, and therefore do not cluster regions based on multi-year CCI trajectories that capture dynamic cost regimes over time. Where clustering is employed, zoning efforts often rely on a single algorithmic family, typically centroid-based partitioning, without an explicit mechanism to accommodate irregular cluster geometry or to flag anomalous districts as noise, even though such exceptions are critical for differentiated budgeting and feasibility assessment [24]. Although density-based methods such as HDBSCAN are well established for identifying non-spherical clusters and explicitly handling noise points, their systematic adoption in construction cost zoning remains limited, particularly in frameworks that aim to produce policy-facing, interpretable zones under heterogeneous regional conditions [25]. Motivated by the availability of official Construction Cost Index (CCI) statistics for Indonesia [26] and their demonstrated use in applied cost benchmarking and adjustment studies [27], this study develops a reproducible hybrid KMeans–HDBSCAN framework to identify statistically distinct and spatially coherent cost zones, examine their temporal patterns in relation to economic and geographic contexts, and evaluate their policy relevance for adaptive infrastructure planning and differentiated budgeting in Indonesia.

METHOD

This study adopts a systematic, data-driven methodology aimed at classifying Indonesian districts and cities according to the temporal behaviour of their construction cost indices. The overarching goal is to inform regionally adaptive infrastructure planning strategies. The analytical process comprises multiple sequential stages: data collection and integration, temporal feature engineering, dimensionality reduction, unsupervised clustering, and cluster validation.

Research Framework

Figure 1 outlines the sequential stages of this study, beginning with the acquisition of CCI and GRDP data, followed by preprocessing through formatting, interpolation of missing values, and z-score normalization. Temporal feature engineering is applied to extract multi-year CCI patterns, and Principal Component Analysis (PCA) reduces dimensionality while preserving at least 80% of variance. The clustering stage integrates KMeans for macro-level segmentation and HDBSCAN for micro-level, irregular pattern detection. Trend profiling then examines cluster trajectories through regression analysis, while validation combines statistical (ANOVA), spatial mapping, and economic (GRDP correlation) checks. Finally, results are interpreted to propose differentiated cost zoning strategies.

The research model integrates centroid-based and density-sensitive clustering with temporal, spatial, and economic validation, offering a novel methodological framework for adaptive infrastructure planning. Figure 1 summarizes the end-to-end workflow, linking data acquisition, preprocessing, hybrid clustering, and multi-perspective validation.

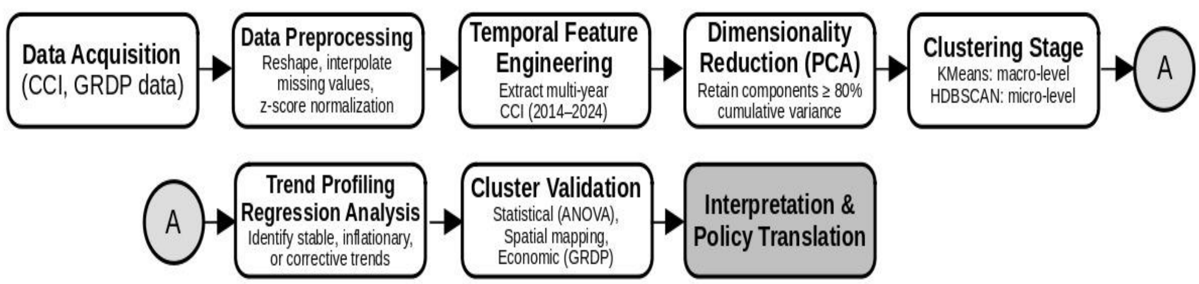

Figure 1. Flowchart and Research Model

Data Sources

The primary dataset used in this study is the construction cost index (CCI) for Indonesian districts and cities. The CCI data span the years 2014 to 2024 and serve as a proxy for the localized cost of building materials and services. The 2014–2024 window was selected to provide a decade-long horizon that captures persistent regional cost structures as well as medium-term fluctuations in construction prices. This period also aligns with the consistent availability and reporting of CCI across districts, enabling cross-regional comparability under a unified measurement framework. Moreover, the decade coincides with major phases of infrastructure expansion in Indonesia, making regional cost differentiation particularly policy-relevant for adaptive budgeting and planning. This dataset was selected due to its official status and wide coverage across the archipelago. To support validation and economic interpretation, provincial-level Gross Regional Domestic Product (GRDP) data were incorporated for the 2016–2024 period. Although GRDP lacks district-level granularity, it remains a robust macroeconomic indicator suitable for comparative assessment of cluster validity.

Data Preprocessing

Initial preprocessing involved reshaping the CCI data into a long panel format, structured by district, province, year, and CCI value. Redundant entries and inconsistencies in district naming were removed to ensure data quality. The time series exhibited minimal missing values, which were resolved through linear interpolation along the temporal axis. Standardization was then performed using z-score normalization to eliminate variance in scale, which is essential for distance-based clustering algorithms. The normalized 11-year CCI time series was subsequently reduced in dimensionality via PCA. Principal components were retained until at least 80% of cumulative variance was explained, thereby preserving essential temporal features while reducing redundancy.

Clustering Procedure

Two clustering techniques were employed: KMeans and HDBSCAN. KMeans was selected for its efficiency and ease of interpretation in partitioning data into a predefined number of clusters. The optimal number of clusters was determined using the Elbow Method and Silhouette Score. However, as KMeans assumes isotropic, equally sized clusters, its interpretive power is constrained in heterogeneous or irregular datasets.

To address these limitations, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) was used as a complementary method. Unlike KMeans, HDBSCAN does not require prior specification of the number of clusters and can detect arbitrary-shaped clusters and noise. This is particularly advantageous in modelling the complex, uneven economic terrain of Indonesia. Together, KMeans and HDBSCAN offer both macro and micro-level insights into regional construction cost dynamics.

HDBSCAN Configuration and Stability

HDBSCAN was employed to capture fine-grained heterogeneity and explicitly handle noise points that may otherwise be absorbed by centroid-based partitioning. The model was implemented in Python using the PCA-reduced feature space derived from standardized multi-year CCI trajectories. Euclidean distance was adopted as the similarity metric in this latent space because PCA produces orthogonal components with comparable scale after z-score standardization, making Euclidean proximity an appropriate measure of similarity between temporal cost profiles. The key hyperparameter, min_cluster_size, was set to 10, which defines the minimum allowable cluster size and controls the granularity of the detected micro-structures. The min_samples parameter was set to 5 to regulate the strictness of density connectivity and sensitivity to noise; larger values would produce more conservative clustering with a greater number of noise-labeled points, whereas smaller values would generate more permissive groupings. The cluster_selection_method was set to "eom" (excess of mass) to obtain the most stable partition from the condensed cluster hierarchy.

To assess clustering stability, a sensitivity analysis was conducted by varying min_cluster_size over a small grid of 8, 10, and 15, while keeping the remaining parameters fixed. Stability was evaluated using two criteria: (i) the persistence scores produced by HDBSCAN and (ii) pairwise agreement between cluster labelling measured by the adjusted Rand index (ARI), calculated after excluding noise points. Across these alternative settings, the principal high-cost and low-cost structures remained qualitatively consistent, while variations mainly affected the granularity of intermediate clusters and the proportion of observations labelled as noise. This indicates that the core zoning patterns identified in the study are robust and not merely artifacts of a single parameter choice.

Trend Profiling and Regression Analysis

Once clustering was complete, average CCI trajectories for each cluster were computed and visualized to assess stability, growth, or volatility. Linear trend lines and first-difference calculations were applied to identify slope direction and rate of change. These temporal patterns were examined in the context of known policy shifts or macroeconomic events. Where applicable, correlation analysis was performed between the CCI trends and provincial GRDP figures to explore the extent to which macroeconomic performance aligns with construction cost evolution at the regional level.

Cluster Validation

To ensure that the derived clusters were statistically meaningful and economically coherent, multiple validation techniques were applied. One-way ANOVA was conducted to test for significant differences in CCI means across clusters. Spatial coherence was then evaluated by mapping the clusters geographically, assessing whether similar CCI patterns emerged in contiguous or demographically similar areas. Furthermore, aggregated GRDP values were used to examine economic distinctions between clusters, supporting the hypothesis that cost groupings reflect real structural disparities rather than algorithmic noise.

Visualization Strategy

To support interpretation and policy translation, the study employed a multi-pronged visualization approach. Line charts were generated to illustrate temporal CCI patterns within each cluster, while boxplots provided information on intra-cluster variation and outliers. Heatmaps were used to show year-over-year shifts in cluster-level CCI averages. Additionally, scatter plots with regression overlays visualized the relationship between CCI and GRDP within key clusters exhibiting strong correlations. These visualizations not only strengthened the interpretability of the results but also served as tools for communicating insights to policymakers and stakeholders.

RESULTS AND DISCUSSION

Overview of Clustering Outcomes

This study implemented two unsupervised learning techniques, KMeans and HDBSCAN, to classify Indonesian districts and cities based on the temporal evolution of their construction cost indices (CCI) from 2014 to 2024. The clustering outputs differ significantly in resolution and interpretability. KMeans identified three broad clusters, while HDBSCAN generated 35 fine-grained clusters, with an additional group of 103 districts labeled as noise due to their inability to form cohesive patterns.

This dual approach serves complementary purposes: KMeans offers a macro-level typology useful for national budgeting strategies, while HDBSCAN captures local heterogeneity and anomalies that may warrant focused intervention.

KMeans Clustering Analysis

The KMeans algorithm partitioned 514 districts into three clusters, revealing distinct patterns in regional construction cost behaviour across time. Each cluster encapsulates a specific structural profile, reflecting variations in economic conditions, logistical challenges, and infrastructure investment dynamics. This segmentation offers a valuable lens for informing differentiated infrastructure policies and cost zoning strategies at the national level.

The clustering results are as follows:

1. Cluster 0 (n = 474): This is the dominant group, exhibiting moderate CCI values with relatively stable behaviour over time.
2. Cluster 1 (n = 10): Characterized by exceptionally high CCI values, with average CCI exceeding 368 in 2016 and climbing to over 392 in 2018, suggesting intense cost pressure in a limited number of high-density urban areas.
3. Cluster 2 (n = 30): Represents transitional regions with initially elevated CCIs (around 168 in 2016) that declined steadily to approximately 150 by 2024.

These results indicate that while most regions maintain moderate construction cost dynamics, a small subset of districts face disproportionately high or structurally declining costs. This distinction supports the need for differentiated subsidy formulas and infrastructure cost benchmarks.

The temporal patterns across these clusters are statistically distinct, with Cluster 1 clearly emerging as an outlier with persistently high costs, while Cluster 2 shows signs of correction or economic deflation in infrastructure pricing.

Table 2 presents the average Construction Cost Index (CCI) for each KMeans cluster over the period 2014–2024. The data highlight clear temporal and structural distinctions among the three clusters. Cluster 0, which contains the majority of districts, maintains a stable cost profile, with CCI values consistently around 100, indicating a relatively predictable cost environment.

Table 2. Average Construction Cost Index (CCI) from 2014 to 2024 by KMeans Cluster

Year	Cluster		
	0	1	2
2014	97.98	380.06	182.16
2015	96.40	371.45	174.31
2016	100.59	368.54	168.23
2017	99.43	375.14	168.25
2018	103.15	392.46	165.29
2019	102.03	384.75	158.17
2020	100.57	372.37	150.88
2021	100.38	368.32	150.27
2022	100.97	322.03	143.76
2023	99.66	299.98	138.28
2024	99.51	290.86	138.81

Cluster 1 stands out as an extremely high-cost group, with values exceeding 368 in 2016 and peaking at 392.46 in 2018 before experiencing a gradual decline after 2020. This suggests sustained cost pressures in a small number of urban or logistically complex regions, followed by potential market corrections. Meanwhile, Cluster 2 reflects transitional regions that started with elevated costs (around 182 in 2014) but exhibited a steady downward trajectory, reaching approximately 138 by 2024, indicating possible economic adjustment or improved infrastructure efficiency over time.

The temporal evolution shown in Table 2 reinforces the earlier observation that Cluster 1 functions as a persistent outlier, while Cluster 2 demonstrates structural correction or deflation in cost levels, contrasting with the stability of Cluster 0. This pattern underlines the necessity for differentiated policy interventions, where high-cost clusters may require targeted subsidies or centralized procurement strategies, while transitional clusters could benefit from investment incentives to sustain their downward cost trends.

Figure 2 compares the cluster-size distributions produced by KMeans and HDBSCAN. Under KMeans, the partition is highly imbalanced: Cluster 0 contains more than 90% of districts, indicating that the algorithm is largely driven by the global centroid structure of the data and tends to aggregate observations into a dominant "mainstream" group. While this coarse segmentation can be useful for macro-level simplification, it also implies limited sensitivity to sparse regions, atypical logistical settings, and localized deviations in construction cost behaviour. This is precisely where HDBSCAN becomes complementary, as its density-based mechanism can resolve finer-grained micro-structures and explicitly separate irregular patterns and anomalies that centroid-based clustering may absorb, thereby supporting more differentiated planning and budgeting decisions.

This dominance reflects a well-known limitation of centroid-based clustering in highly heterogeneous settings, where a single global centroid can absorb diverse local regimes into a broad cluster, leaving smaller or atypical regions under-resolved. In an archipelagic context such as Indonesia, where logistical constraints and market access vary sharply across districts, this aggregation risks masking structurally distinct cost behaviours that are critical for differentiated budgeting. Accordingly, HDBSCAN is introduced as a complementary method to recover fine-grained structures and explicitly isolate anomalies that KMeans may subsume under the dominant cluster.

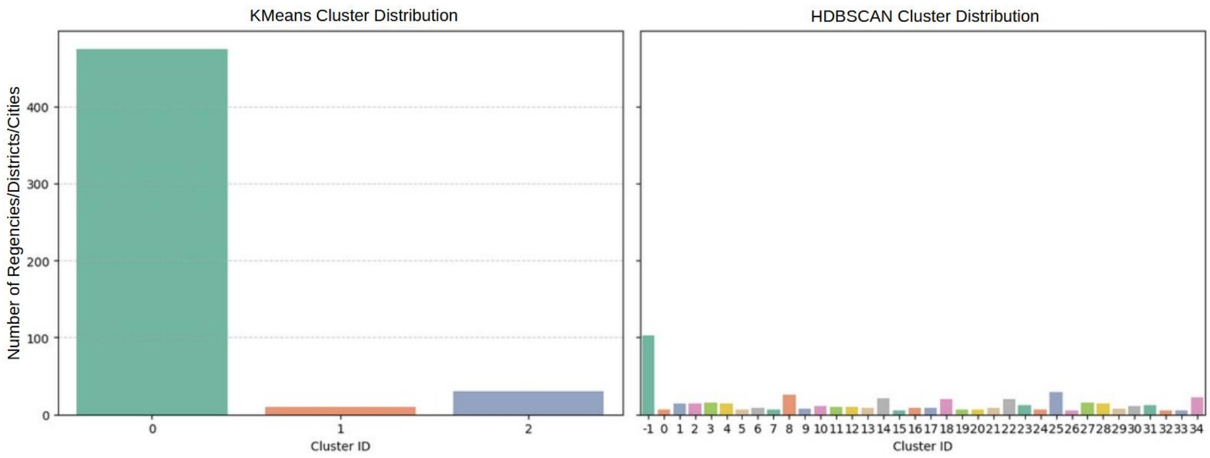

Figure 2. Comparison of Cluster Size Distribution between KMeans and HDBSCAN Methods

Conversely, HDBSCAN demonstrates a radically different clustering behaviour. It yielded 35 clusters with a more balanced spread of members and over 100 districts marked as noise (Cluster -1). This reflects the model's density-sensitive nature, enabling it to identify both tightly-bound regional clusters and isolated outliers that KMeans may absorb into dominant groups.

The contrast between the two methods highlights their complementary strengths: KMeans provides tractable general segmentation, while HDBSCAN enables granular differentiation and anomaly detection. This divergence is particularly relevant for infrastructure planning, where uniform cost assumptions (KMeans-like grouping) must be balanced against territorial exceptions and transition zones (HDBSCAN granularity).

To better understand the size distribution of clusters produced by the KMeans algorithm, Figure 2 (left panel) shows that Cluster 0 accounts for the majority of districts (n = 474), while Clusters 1 and 2 contain only a small number of regions. This imbalance illustrates KMeans' tendency to form centroid-dominant partitions, often absorbing subtle regional variations into a large general category.

HDBSCAN Clustering and Micro-Level Insights

In contrast, HDBSCAN revealed 35 micro-clusters with distinct cost behaviours, as well as 103 noise-labeled districts. These noise points represent cases with inconsistent or highly irregular CCI patterns that could not be grouped reliably. For instance:

1. HDBSCAN Cluster 0 (n = 6) exhibits extremely high CCIs similar to KMeans Cluster 1, with averages reaching 347 in 2016 and peaking at 365.79 in 2018.
2. Cluster 1 (n = 14) captures moderately high-cost regions, averaging 144.95 in 2016 but gradually decreasing to 131.56 by 2020.
3. Clusters 3, 4, 5, and 6 represent stable low-cost districts with average CCIs around 85–115 and little annual variance.

This higher-resolution clustering enables nuanced zoning of development priorities. For instance, clusters with gradual increases (e.g., Cluster 6) might be targeted for early-stage support, while persistently high-cost areas (Clusters 0 and 1) may require targeted interventions such as centralized procurement or price stabilization programs.

Table 3 summarizes the average Construction Cost Index (CCI) from 2014 to 2024 for selected HDBSCAN clusters.

Table 3. Excerpt*:* Average Construction Cost Index (CCI) from 2014 to 2024 by HDBSCAN Cluster

Year	**Cluster**			
	0	**1**	**4**	**6**
2014	361.97	160.59	98.41	85.11
2015	348.17	150.17	94.13	87.18
2016	347.27	144.95	113.28	112.85
2017	352.27	144.29	114.07	110.58
2018	365.79	140.93	117.59	113.57
2019	357.96	137.32	117.62	114.16
2020	340.96	131.56	115.68	112.97
2021	335.97	132.59	113.5	111.59
2022	293.83	129.28	112.76	112.55
2023	274.66	126.13	110.73	110.83
2024	270.17	126.05	110.55	112.07

The data highlight a more granular cost behaviour compared to KMeans, revealing distinct micro-clusters. Cluster 0, for example, consistently shows exceptionally high CCI values, averaging above 340 throughout the period and peaking at 365.79 in 2018. This mirrors the cost profile of KMeans Cluster 1, reinforcing its identification as a high-cost outlier group concentrated in urban or logistically constrained areas.

Meanwhile, Cluster 1 represents moderately high-cost regions where the CCI declined steadily from 160.59 in 2014 to 126.05 in 2024, suggesting structural adjustment or policy-induced cost moderation. By contrast, Clusters 4 and 6 exhibit stable, low-cost profiles, with average CCI values consistently in the range of 85–115, showing minimal annual fluctuation. These clusters can be interpreted as baseline zones with predictable cost environments, suitable for standardized budgeting frameworks.

The temporal trends displayed in Table 3 illustrate the value of HDBSCAN's finer resolution in identifying transitional, stable, and high-cost anomalies. Such distinctions are critical for micro-level infrastructure zoning, where high-cost clusters may require price stabilization policies, while low-cost clusters can be maintained through standard funding protocols.

Figure 2 (right panel) displays the HDBSCAN cluster distribution, which is considerably more fragmented. The algorithm discovered 35 distinct micro-clusters, with a significant portion of districts (n = 103) categorized as noise. This reflects HDBSCAN's sensitivity to local density variations and its ability to isolate irregular or volatile cost behaviours that KMeans fails to detect.

Trend Interpretation and Policy Implications

Across both clustering methods, a clear trend emerges: construction costs in certain clusters are consistently rising, while others exhibit relative stability or mild decline. This divergence underscores the inadequacy of uniform budget assumptions in national infrastructure planning.

The results suggest that:

1. Cluster 1 (KMeans and HDBSCAN) demands urgent cost-control strategies.
2. Cluster 2 (KMeans) may indicate regions in transition or economic correction.
3. HDBSCAN Clusters 3–10 could benefit from localized pricing models to stimulate infrastructure growth efficiently.

Furthermore, noise districts in HDBSCAN (n = 103)—often remote or socio-politically unique—highlight the limitations of conventional economic assumptions. These areas may require special attention under resilience and adaptation frameworks.

Figure 3 presents the average CCI trajectories from 2016 to 2024 for the clusters identified by KMeans (left panel) and HDBSCAN (right panel), allowing a direct comparison of how regional construction costs evolve over time under macro- versus micro-level segmentation. The KMeans curves summarize broad, dominant patterns in the data, whereas the HDBSCAN curves reveal greater heterogeneity by separating districts into finer-grained groups with more distinctive temporal profiles. The divergence across cluster trend lines indicates that regional differences are not limited to cost levels but also extend to the direction and persistence of cost movements over time, which is central to multi-year budgeting and phased infrastructure delivery. Taken together, the figure reinforces the rationale for the hybrid design: KMeans offers an interpretable high-level structure for national planning, while HDBSCAN complements it by uncovering localized dynamics and atypical trajectories that would otherwise be absorbed into broader centroid-based clusters.

The KMeans segmentation reveals a clear stratification of cost behaviour:

1. Cluster 1, comprising only a few districts, consistently maintains the highest CCI values, peaking above 390 in 2018 and remaining elevated through 2024. This reflects sustained cost pressures in specialized high-cost zones, likely regions facing persistent structural cost pressures, including logistical constraints and limited market accessibility.
2. Cluster 2 exhibits a downward trajectory, declining from an average of 168 in 2016 to approximately 144 by 2024, suggesting either policy interventions or structural economic shifts.
3. Cluster 0, the largest by membership, maintains a relatively stable cost profile, serving as a benchmark for standard infrastructure budgeting assumptions.

In contrast, the HDBSCAN clusters offer greater resolution. Cluster 0 stands out with consistently extreme CCI values, reinforcing its identification as a cost outlier. Clusters 1 through 6 exhibit differentiated behaviours: some maintain plateau-like trends (e.g., C5 and C6), while others show gradual cost moderation over time (e.g., C1 and C2).

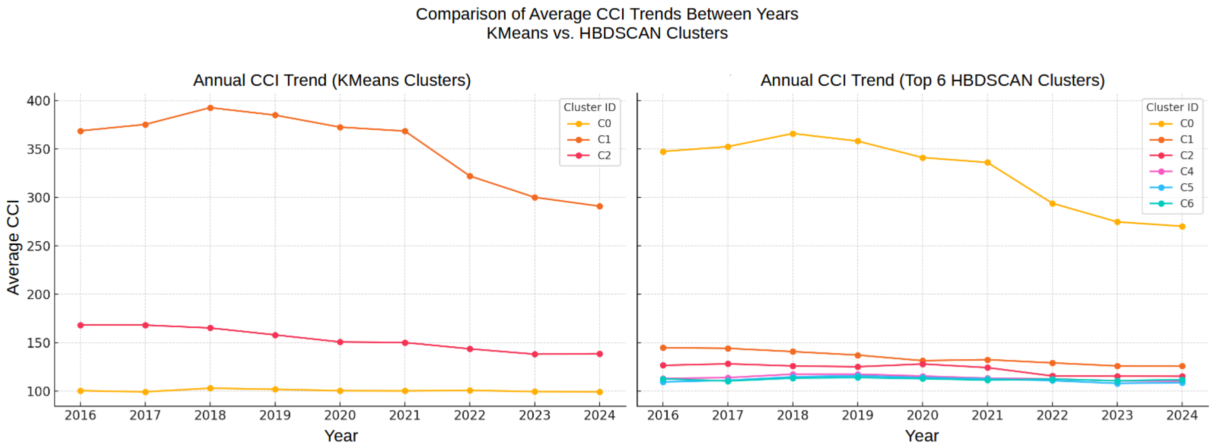

Figure 3. Average Construction Cost Index (CCI) Trends from 2016 to 2024 by Cluster

The diversity of trends across HDBSCAN clusters demonstrates the algorithm's strength in capturing localized variations and invisible cost dynamics under coarser segmentation like KMeans.

Together, these visual trends reinforce the notion that infrastructure planning should not rely on static or nationally uniform cost baselines. Instead, policy frameworks must account for both enduring structural disparities (as seen in KMeans) and sub-regional variability and volatility (as revealed by HDBSCAN), enabling more adaptive and equitable allocation of resources.

Figure 4 illustrates the distributional dynamics of the Construction Cost Index (CCI) between 2020 and 2024 for selected HDBSCAN clusters (IDs: 0, 1, 4, and 6) using year-by-year boxplots. By summarizing the median, interquartile range, and outliers within each cluster-year, the figure shows that cost behaviour differs not only in central tendency but also in dispersion and tail behaviour across groups. Several clusters exhibit widening spreads and more pronounced outliers over time, suggesting that localized cost volatility and atypical districts persist even within structurally similar groups. This distributional evidence reinforces the need for an anomaly-sensitive component in regional cost zoning: while macro-level clustering can provide a useful simplification for national planning, density-based grouping offers additional resolution by isolating heterogeneous variability and highlighting districts that may require differentiated budgeting assumptions or deeper feasibility assessment.

The results highlight significant temporal dynamics:

1. Cluster 0, representing high-cost urban or logistically constrained areas, experienced a notable decline in CCI, from a median above 340 in 2020 to around 280 in 2024. This may indicate either a market correction post-COVID or the effects of targeted interventions in cost regulation and procurement practices.
2. Cluster 1, categorized by moderately elevated costs, shows a slight reduction in median CCI values and narrower dispersion in 2024, suggesting increasing price stability.
3. Clusters 4 and 6, which represent low-cost and stable regions, maintained consistent distributions over time, reinforcing their suitability for baseline budgeting models.

This comparative boxplot provides stronger empirical evidence than mean-based trends alone, revealing both intra-cluster volatility and inter-temporal structural shifts. These insights are critical for designing adaptive infrastructure policies, especially when forecasting future budget needs under regional heterogeneity.

Figure 5 compares the temporal structure of construction costs through a heatmap of average CCI values from 2016 to 2024. The left panel summarizes the three KMeans clusters, providing a compact macro-level view of cost evolution, whereas the right panel focuses on the six HDBSCAN clusters with the highest mean CCI in 2020, revealing more differentiated high-cost profiles. The contrast between panels indicates that while KMeans captures broad patterns, it can compress heterogeneous trajectories into a small number of groups, thereby limiting its ability to represent localized variability.

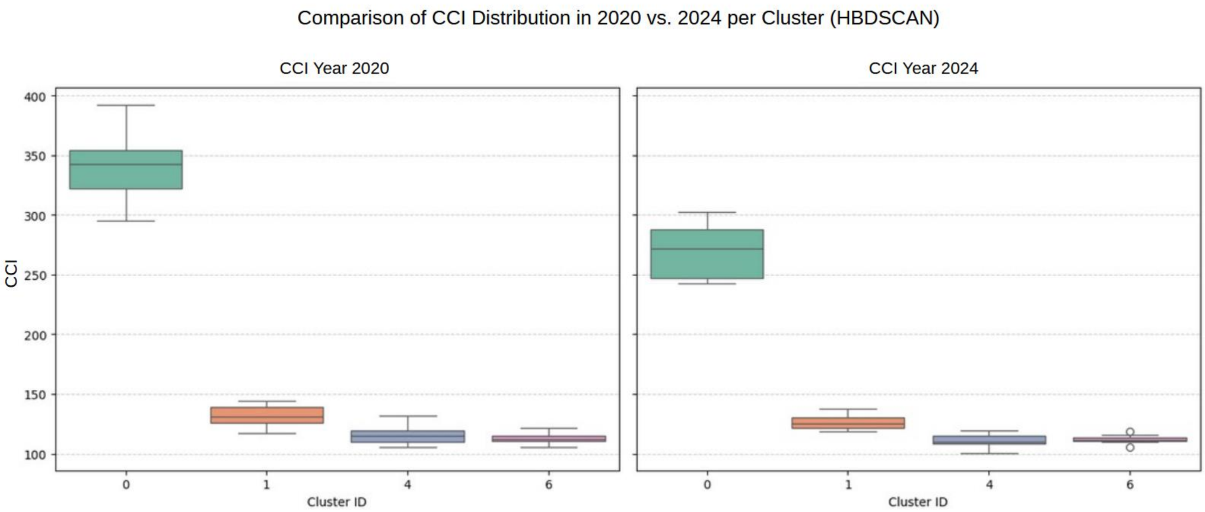

Figure 4. Distribution of CCI in 2020 and 2024 by Selected HDBSCAN Clusters

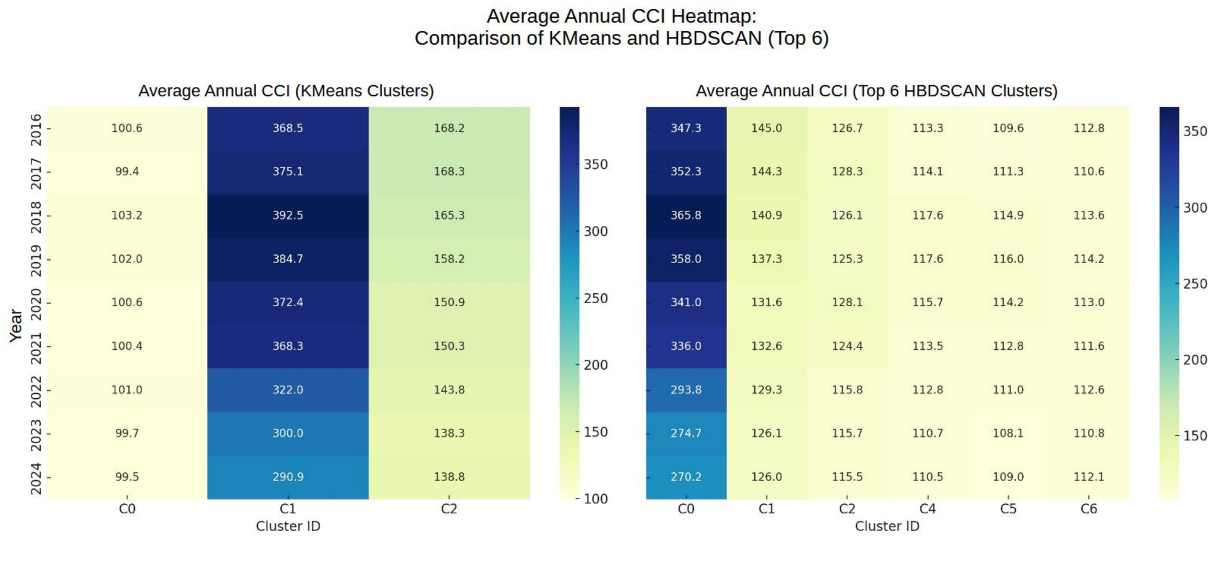

Figure 5. Temporal Heatmap of Average CCI (2016–2024) for KMeans and Top HDBSCAN Clusters

By isolating distinct high-cost clusters and, in the full model output, distinguishing additional micro-structures and anomalous districts, HDBSCAN provides the granularity needed to surface irregular cost dynamics that may be obscured under centroid-based partitioning. This comparative heatmap therefore supports the hybrid rationale: KMeans offers an interpretable national segmentation, while HDBSCAN refines the analysis by uncovering fine-grained temporal heterogeneity that is directly relevant for differentiated budgeting and targeted planning interventions.

The KMeans heatmap reveals a clear segmentation of macro-level cost behaviour. Cluster 1 consistently exhibits the highest average CCI values, peaking above 390 in 2018 and remaining well above 370 through 2020. This pattern corresponds to a small group of high-density or high-cost urban districts. In contrast, Cluster 2 shows a declining trend, indicative of possible post-peak correction or policy-induced deflation in localized infrastructure costs. Cluster 0 remains relatively stable, reflecting the majority of districts with moderate and consistent construction costs.

The HDBSCAN heatmap, on the other hand, uncovers greater granularity and temporal heterogeneity. Cluster 0, representing the extreme-cost outliers, maintains consistently high CCI values above 340 across all years, mirroring the behaviour of KMeans Cluster 1 but with tighter grouping. Clusters 1 through 6 exhibit varying degrees of decline, stabilization, or slow growth, offering a nuanced view of temporal shifts in medium-cost regions. These micro-clusters highlight regional pockets where cost dynamics diverge from the national average, which would have been obscured in a broader KMeans segmentation.

This comparative heatmap underscores the complementary strengths of the two clustering approaches. While KMeans offers interpretable and tractable clusters suitable for national budgeting frameworks, HDBSCAN excels at capturing sub-national variability and identifying localized anomalies. Such insights are particularly valuable for adaptive infrastructure planning, where temporal behaviour is as critical as spatial or economic grouping.

Integrating Clustering into Infrastructure Zoning

The findings from this study establish a solid analytical basis for the implementation of dynamic infrastructure zoning throughout Indonesia's varied regional landscape. By categorizing each district according to its specific construction cost behaviour, policymakers gain a flexible framework that goes beyond traditional administrative classifications. These cluster-based zones allow planners to customize infrastructure policies, funding strategies, and regulatory standards in line with the actual structural conditions indicated by the data, rather than relying on broad, uniform assumptions at the national level.

In practical terms, incorporating CCI cluster identifiers into national development planning systems creates several viable pathways. For example, procurement standards and budget limits can be adjusted based on the typical cost patterns of each cluster, thereby ensuring that funding levels are aligned with regional price variations. Subsidy allocations can also be differentiated according to long-term cost trends. This approach guarantees that financially challenged or high-cost regions receive the necessary support, without creating distortions in more stable areas. Additionally, districts identified as outliers by HDBSCAN—those with inconsistent or unpredictable cost patterns, can be singled out for closer scrutiny, specialized audits, or policy trials, especially in remote, unstable, or post-disaster contexts.

This zoning approach improves both fiscal efficiency and fairness in the delivery of public infrastructure. It facilitates a transition from reactive, uniform budgeting to a more proactive and evidence-based strategy that acknowledges geographic, economic, and logistical diversity. Given Indonesia's archipelagic geography and uneven development, such a data-driven framework holds significant potential. It enables the government to target investments more effectively, manage cost inflation, and ensure that infrastructure projects are both economically sustainable and socially responsive. By incorporating these clustering insights into long-term planning tools, national infrastructure policy can advance toward greater adaptability, resilience, and equity.

Validity Analysis

We conducted a statistical validation using one-way Analysis of Variance (ANOVA) on the 2020 CCI values to ensure that the identified clusters reflect meaningful economic segmentation rather than spurious groupings. This approach evaluates whether the differences in average construction costs across clusters are statistically significant.

For the KMeans clustering, the ANOVA test yielded an F-statistic of 1696.49 with a p-value of 2.35×10^{-226}, indicating that the mean CCI values among the three clusters differ significantly. The result confirms that the clustering captured distinct cost profiles, particularly the high-cost outlier group (Cluster 1), compared to the more moderate Clusters 0 and 2.

Similarly, for the HDBSCAN clustering, after excluding the noise-labeled districts (Cluster -1), the ANOVA returned an F-statistic of 370.91 with a p-value of 9.22×10^{-267}. Despite the larger number of clusters (n = 35), the statistically significant result underscores the robustness of HDBSCAN in identifying structurally distinct regions based on their CCI trajectories.

Table 4 reports the F-statistics and p-values obtained from one-way ANOVA tests assessing the statistical significance of mean differences in construction cost indices (CCI) across clusters. Noise points were excluded for HDBSCAN to ensure comparability.

These statistical outcomes validate the interpretability of the clustering results. The clusters are not arbitrary aggregations but represent economically coherent groupings of regions with shared cost characteristics.

In addition to statistical significance, the spatial coherence of clusters was also evident. For example, several HDBSCAN clusters, such as those representing eastern Indonesian provinces with similar infrastructural challenges and cost structures, are geographically concentrated. This spatial alignment reinforces the argument that the clusters reflect real economic and geographic constraints rather than noise or model overfitting.

Figure 6 maps the spatial distribution of cluster memberships across Indonesian provinces, contrasting the macro-level KMeans zoning (left panel) with the six highest-cost HDBSCAN clusters (right panel). The KMeans map conveys a broad, easily interpretable regional structure, whereas the HDBSCAN map reveals more selective high-cost pockets that are geographically concentrated in specific areas rather than evenly dispersed. This spatial contrast indicates that regional cost heterogeneity is not fully represented by coarse centroid-based partitions and that high-cost behaviour can be localized, reflecting differences in market access, logistics, and regional supply constraints. From a planning perspective, these geographically distinct high-cost clusters provide actionable signals for differentiated budgeting assumptions and targeted feasibility assessment, strengthening the case for the hybrid framework in which KMeans supports national-level prioritization while HDBSCAN adds anomaly-sensitive refinement for region-specific policy interventions.

Table 4. One-way ANOVA test results for KMeans and HDBSCAN cluster validation using CCI 2020 values

Clustering Method	F-statistic	p-value	Number of Clusters	Noise Excluded
KMeans	1696.49	2.35×10^{-226}	3	N/A
HDBSCAN	370.91	9.22×10^{-267}	35	Yes

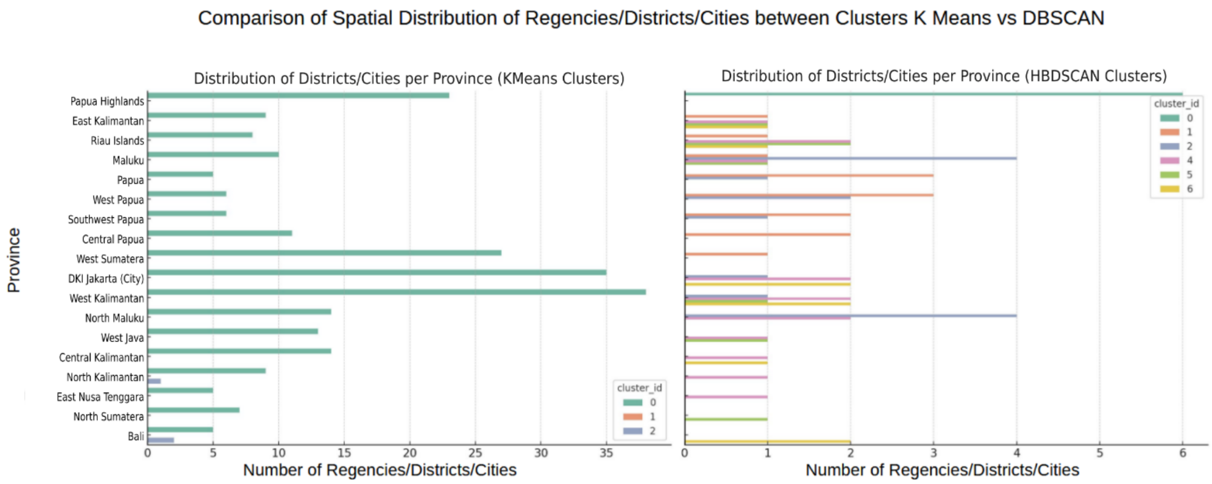

Figure 6. Comparative Spatial Distribution of Cluster Membership (KMeans vs HDBSCAN)

The KMeans segmentation shows that Cluster 0, the largest group, spans nearly all major provinces, reflecting its role as the national baseline. In contrast, Cluster 1, associated with extremely high construction costs, is concentrated in urbanized and high-density regions, including DKI Jakarta and Bali.

The HDBSCAN distribution reveals greater spatial coherence and fragmentation. Cluster 0, for example, comprises districts in Papua Highlands, signaling regional isolation and extreme cost profiles. Several provinces such as East Nusa Tenggara and South Sulawesi contribute to multiple clusters, indicating internal heterogeneity in infrastructure cost dynamics.

This spatial analysis reinforces the argument that the identified clusters are statistically valid and geographically interpretable, further justifying their use in adaptive infrastructure zoning.

To further validate the economic coherence of the clustering outcomes, we examined the correlation between regional GDP (GRDP) and the Construction Cost Index (CCI) across clusters. While the overall correlation was weakly positive, stratifying the analysis by cluster revealed important differences. In several HDBSCAN clusters, the correlation between GRDP and CCI exceeded 0.5 in absolute terms, indicating a stronger association between economic output and construction cost levels within those groupings.

For example, HDBSCAN Cluster 1 demonstrated a moderately strong positive correlation ($r = 0.57$), suggesting that in this cluster, higher regional economic output is closely tied to higher construction costs. This relationship may reflect economically active regions where demand-driven cost escalation and supply chain intensity drive up prices. Conversely, clusters such as Cluster 3 showed a moderately negative correlation ($r = –0.35$), highlighting areas where high costs may not be accompanied by high economic returns—possibly due to remoteness, inefficiency, or infrastructural constraints. These divergences reinforce the clusters' interpretability and utility in capturing not only statistical differences in CCI but also underlying economic dynamics.

Tabe 5 lists a selection of districts within high-correlation clusters, including Simeulue, South Aceh, and Southeast Aceh, where consistent GRDP and CCI relationships further support the robustness of the segmentation. The presence of coherent economic signals within certain clusters validates that the groupings are not arbitrary but grounded in measurable regional conditions, thereby enhancing their relevance for planning and budgeting decisions.

The relationship between regional GDP (GRDP) and the Construction Cost Index (CCI) within selected HDBSCAN clusters—particularly those showing notably strong positive or negative correlations is illustrated in Figure 7.

A clear positive association emerges in clusters such as HDBSCAN Cluster 32, indicating that regions with higher economic output also tend to experience elevated construction costs likely driven by greater demand and investment intensity. In contrast, clusters like HDBSCAN Cluster 8 demonstrate a strong negative correlation, suggesting that certain areas face high construction costs despite lower economic productivity, potentially due to geographic isolation or logistical inefficiencies.

Table 5. Districts within HDBSCAN clusters exhibiting high correlation (|r| > 0.5) between regional GDP (GRDP) and Construction Cost Index (CCI).

District/City	Province	Clustering Method	Cluster ID	GRDP 2020	CCI 2020	r (GRDP–CCI)
Simeulue	Aceh	HDBSCAN	8	31633	109.99	–0.538
South Aceh	Aceh	HDBSCAN	24	31633	98.07	–0.521
Southeast Aceh	Aceh	HDBSCAN	24	31633	100.77	–0.521
East Aceh	Aceh	HDBSCAN	32	31633	95.43	0.64
West Aceh	Aceh	HDBSCAN	30	31633	89.63	0.65

.

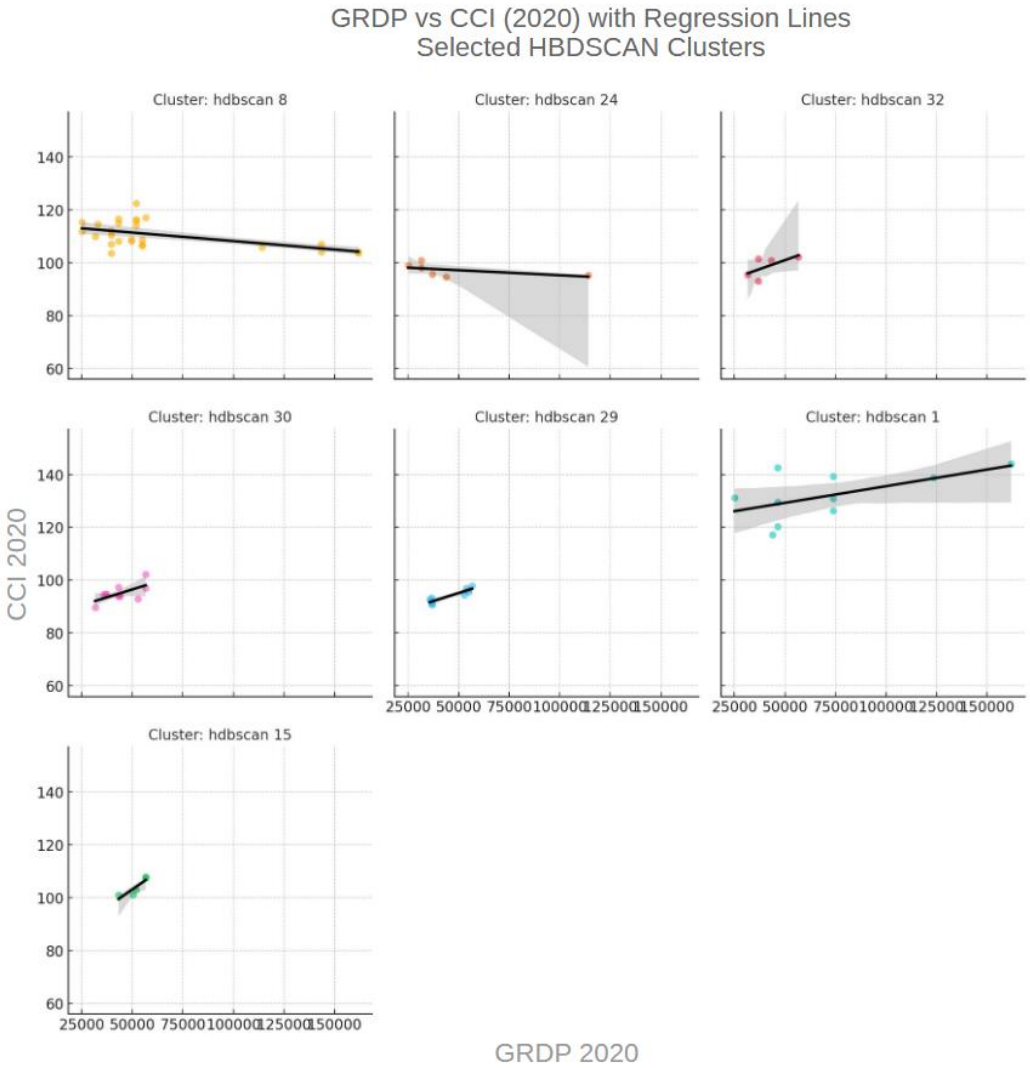

Figure 7. GRDP vs CCI (2020) with Regression Lines for Selected HDBSCAN Clusters

These divergent intra-cluster patterns enhance the interpretability of the clustering results and underscore the importance of incorporating economic context into region-specific infrastructure planning and budget allocation.

Discussion

To synthesize the key findings from the clustering analysis, Table 6 summarizes the most salient characteristics observed across both KMeans and HDBSCAN results. This overview highlights statistical distinctions and spatial, temporal, and economic patterns supporting differentiated infrastructure strategies.

Table 6. Summary of Key Cluster Characteristics and Analytical Insights

Aspect	Cluster	Key Characteristics
Highest Average CCI	KMeans Cluster 1 / HDBSCAN Cluster 0	Persistently high costs, urban or logistically constrained areas
Most Stable Cost Behaviour	KMeans Cluster 0 / HDBSCAN Cluster 4 and 6	Moderate CCI, low variation, suitable for baseline budgeting
Declining CCI Trend	KMeans Cluster 2 / HDBSCAN Cluster 1	Indicates market correction or changing infrastructure intensity
High GRDP–CCI Correlation	HDBSCAN Cluster 32, 30	Strong alignment between cost and economic output
Negative GRDP–CCI Correlation	HDBSCAN Cluster 8, 24	High cost in regions with modest economic capacity
High Cluster Membership (Size)	KMeans Cluster 0	Majority of districts (n = 474), representing national baseline
Most Fragmented Structure	HDBSCAN	35 micro-clusters + 103 noise districts, captures local anomalies

A technical comparison between the two clustering strategies further clarifies why the hybrid design is necessary. Under KMeans, the partition is highly imbalanced: the dominant macro-cluster contains 474 districts (over 90% of observations), indicating that centroid-based optimization in a heterogeneous national dataset tends to absorb diverse local regimes into a single "mainstream" cluster. In contrast, HDBSCAN yields a more granular structure by identifying 35 micro-clusters and flagging 103 districts as noise, explicitly separating atypical trajectories rather than forcing them into global centroids. This difference is also reflected in statistical separability: one-way ANOVA on 2020 CCI indicates highly significant between-cluster differences for both methods, with KMeans producing $F = 1696.49$ ($p = 2.35 \times 10^{-226}$) and HDBSCAN producing $F = 370.91$ ($p = 9.22 \times 10^{-267}$, excluding noise). While the larger KMeans F-statistic is consistent with broad separation between coarse macro-groups, HDBSCAN's value reflects partitioning across many smaller groups, where between-cluster variance is distributed across multiple micro-structures rather than concentrated in a few centroids. From an algorithmic standpoint, these patterns align with known properties: KMeans is most effective when clusters are approximately convex and similarly sized in Euclidean space, whereas HDBSCAN is better suited to irregular cluster geometry and heterogeneous density, with the additional advantage of providing an explicit noise set that is directly actionable for targeted feasibility assessment and contingency budgeting. Therefore, the hybrid framework leverages KMeans for interpretable macro-zoning while using HDBSCAN to recover local irregularities and outliers that are most critical for differentiated infrastructure planning.

To complement the analytical summary, Table 7 lists selected districts and cities from the key HDBSCAN clusters referenced in the analysis. These clusters represent distinct regional typologies, ranging from high-cost urban zones to stable-cost rural or transitional regions, each with its own policy implications. By mapping actual jurisdictions to cluster labels, this table enhances the interpretability of the clustering results and facilitates future integration into policy instruments such as cost-zoned procurement frameworks, subsidy targeting, or differentiated monitoring strategies. The geographic diversity of cluster memberships, spanning both western and eastern Indonesia, underscores the utility of this clustering framework in capturing nationwide heterogeneity in infrastructure cost dynamics.

Table 7. Selected Districts and Cities from Key HDBSCAN Clusters.

Cluster	District
0	Jayawijaya, Lanny Jaya, Tolikara, Mamberamo Tengah, Yalimo, Bintang Highlands
1	Mentawai Islands, Anambas Islands, Mahakam Ulu, Southwest Maluku, Fakfak, Bintuni Bay, South Manokwari, Raja Ampat, Maybrat, Yapen Islands, Biak Numfor, Jayapura City, Mimika, Nabire
4	Lingga, Batam City, East Jakarta City, Central Jakarta City, Bogor, Sabu Raijua, Sanggau, Ketapang, Murung Raya, Kutai Kartanegara, Nunukan, Buru, West Halmahera, North Halmahera
6	Bintan, West Jakarta City, North Jakarta City, Badung, Denpasar City, Sambas, Sintang, Sukamara, Bontang City
8	Simeulue, Nias, North Nias, West Nias, Meranti Islands, Bandung, Bogor City, Bandung City, Gresik, Jembrana, Tabanan, Bangli, Buleleng, Sekadau, North Kayong, Kubu Raya, Singkawang City, Berau, Balikpapan City, Malinau, Tarakan City, Wakatobi, North Buton, Central Maluku, West Seram, South Halmahera
24	South Aceh, Southeast Aceh, Bukittinggi City, Pelalawan, Wonogiri, Central Lombok
30	West Aceh, South Solok, Dharmasraya, Ciamis, Banjar City, Klaten, Magelang City, Situbondo, Probolinggo, Serang City, Bone Bolango
32	East Aceh, Tasikmalaya City, Pemalang, Tegal City, Tuban

The spatial composition of several clusters, as shown in Table 7, further supports their economic plausibility. For example, Cluster 0 is composed entirely of districts in the Papua Highlands region, including Jayawijaya, Lanny Jaya, Tolikara, Mamberamo Tengah, Yalimo, and Bintang Highlands. This concentration in mountainous and relatively inaccessible areas is consistent with the exceptionally high construction costs observed in earlier sections, reflecting the combined effects of difficult terrain, limited transport infrastructure, and high logistics dependency. By contrast, Cluster 6 includes a mix of urban and economically active areas, such as West Jakarta City, North Jakarta City, Denpasar City, and Bontang City, together with several districts in Kalimantan and Bali. The composition of this cluster suggests a cost structure shaped less by extreme remoteness and more by urbanization, labor market intensity, and tourism- or trade-related construction demand.

The regional concentration visible in Table 7 indicates that the clusters are not arbitrary statistical groupings, but correspond to recognizable geographic and economic contexts. In this sense, the clustering results capture real-world disparities in infrastructure cost environments across Indonesia, ranging from remote highland districts to metropolitan and coastal growth centers.

The empirical analysis presented in the preceding sections reveals a structured pattern of regional construction cost dynamics across Indonesian districts. The clustering results demonstrate that construction costs are neither uniformly distributed nor randomly dispersed across regions. Instead, they form identifiable groupings characterized by distinct temporal trajectories and spatial concentrations. The KMeans analysis provides a macro-level segmentation of the national cost landscape, separating districts into dominant moderate-cost regions, high-cost outliers, and transitional zones. In contrast, HDBSCAN offers a more granular perspective, uncovering a larger number of micro-clusters as well as noise-labeled districts that exhibit irregular or unstable cost trajectories. Together, these complementary methods highlight both the structural consistency and localized variability present in Indonesia's construction cost environment.

Firstly, the identification of high-cost clusters (e.g., KMeans Cluster 1 and HDBSCAN Cluster 0) points to zones experiencing sustained cost pressure. These areas likely face structural constraints related to labor availability, logistical complexity, or material supply chains. For such regions, national planners may consider policy instruments such as bulk procurement mechanisms, pre-negotiated contract rates, or targeted subsidy buffers to mitigate the risk of cost overruns. Conversely, moderate-cost clusters (such as KMeans Cluster 0 or HDBSCAN Clusters 4 and 6) represent environments with relatively stable and predictable pricing dynamics, making them suitable for standard infrastructure budgeting frameworks.

Secondly, the direction of temporal trends within clusters offers strategic insights for planning. For example, districts belonging to HDBSCAN Cluster 1 exhibit a declining CCI trajectory over time, which may indicate market stabilization, improved supply efficiency, or reduced infrastructure demand. In such cases, policy interventions could focus on stimulating construction activity through incentive-based programs or targeted development initiatives in underdeveloped regions.

Thirdly, the classification of 103 districts as noise by HDBSCAN highlights the presence of irregular cost trajectories that do not conform to stable cluster structures. These outliers may reflect volatile market conditions, geographic isolation, or institutional factors that affect construction pricing dynamics. Rather than being excluded from analysis, such districts should be treated as priority cases for deeper diagnostic assessment, as their cost structures may require customized policy responses.

The clustering framework also enhances transparency and accountability in infrastructure budgeting. By assigning each district to a well-defined cost typology, planners and auditors gain a reference baseline for evaluating project cost proposals, benchmarking unit prices, and identifying abnormal deviations. This systematic classification improves fiscal oversight and supports more evidence-based infrastructure allocation decisions.

The complementary use of KMeans and HDBSCAN demonstrates the advantages of combining global and density-based clustering methods. While KMeans provides a clear and interpretable macro-level segmentation suitable for national policy formulation, HDBSCAN captures finer regional heterogeneity and localized anomalies. Together, these approaches provide a robust analytical foundation for differentiated infrastructure governance, moving beyond uniform budgeting practices toward more adaptive and region-sensitive planning.

In addition to validating the spatial coherence of the clusters, the observed correlations between regional economic output

(GRDP) and construction costs provide further interpretive insight. The presence of heterogeneous, cluster-specific relationships suggests that different economic mechanisms shape construction cost dynamics across regions. In economically active districts, higher costs may be associated with strong demand for labor and construction inputs. In contrast, some regions exhibit elevated costs despite relatively modest economic output, reflecting structural constraints such as remoteness, logistical barriers, or limited supplier networks. These distinctions reinforce the need for economically informed infrastructure planning that aligns investment strategies with both cost structures and regional development conditions.

From a policy implementation perspective, districts classified as noise by HDBSCAN should be treated as actionable signals rather than discarded observations. These cases can function as an early-warning layer for targeted feasibility assessments, where planners conduct detailed cost decomposition analyses, including materials, logistics, labor availability, and contractor capacity, to determine whether unusual CCI trajectories reflect genuine market constraints or measurement inconsistencies. Budgeting frameworks may therefore incorporate contingency allocations or risk buffers for such districts, while procurement strategies can be adapted through flexible contract packaging, supplier qualification adjustments, or modified delivery schedules. Integrating these noise indicators into routine monitoring systems would also support periodic data validation and recalibration of cost assumptions, ensuring that exceptional districts receive appropriate policy attention rather than being absorbed into national averages.

Policy Recommendation

Drawing from the outcomes of clustering analysis coupled with their temporal, spatial, and economic validations, this research delineates a series of policy recommendations aimed at fostering more adaptive and equitable frameworks for infrastructure planning. Initially, it is imperative to incorporate CCI-based clusters into the budgeting processes for infrastructure, thereby supplanting uniform cost assumptions with empirically substantiated regional typologies. It is recommended that national planners implement differentiated ceilings for procurement and varying levels of subsidies contingent upon the specific cost behaviours associated with each cluster. For example, districts categorized within high-cost clusters, such as KMeans Cluster 1 or HDBSCAN Cluster 0, may find substantial advantages in adopting centralized procurement systems, engaging in bulk contract negotiations, or implementing location-specific cost modifications to mitigate the risks of chronic underfunding.

Furthermore, development incentives ought to be strategically directed towards transitional clusters, illustrated by KMeans Cluster 2 or HDBSCAN Cluster 1, where a decline in construction costs is indicative of either enhanced efficiency or diminished activity levels. These regions may currently be undergoing phases of market correction or infrastructural stagnation and stand to benefit from performance-based grant mechanisms or catalytic investment initiatives designed to spur economic growth. In addition, districts classified as noise-labeled by HDBSCAN, which display erratic cost fluctuations, necessitate more rigorous fiscal oversight and may require bespoke interventions, including resilience-oriented infrastructure support or mechanisms for real-time price auditing.

Finally, the observed correlation between GRDP and CCI in certain clusters reinforces the need for regionally aligned budgeting logic. Clusters where economic output is tightly coupled with construction costs should be prioritized for scalable investment, while clusters with negative correlations may require deeper assessment of structural barriers or inefficiencies. By institutionalizing a data-driven, cluster-informed zoning mechanism, national and sub-national governments can allocate infrastructure resources more efficiently and equitably, improving not only cost-effectiveness but also regional development outcomes.

To operationalize the analytical insights derived from the clustering results, Table 8 outlines a set of tailored policy recommendations mapped to key cluster types. Each recommendation is informed by the cost profile and economic context of the cluster, aiming to support more precise, equitable, and responsive infrastructure planning across regions.

The results of this study generally support the findings in previous literature, which suggest that spatial and temporal variations in infrastructure costs need to be systematically identified and mapped to support adaptive and equitable development planning. For example, the clustering-based approach in this study reinforces the findings of Min et al. [28], who used KMeans to cluster traffic dynamics and found the strategic value of spatial data segmentation in data-driven policymaking [29].

Table 8. Recommended policy actions by cluster type and construction cost profile.

Cluster Type	Cost Profile	Policy Recommendation
KMeans Cluster 1 / HDBSCAN Cluster 0	Persistently high	Centralized procurement, cost ceiling adjustments, targeted subsidies
KMeans Cluster 2 / HDBSCAN Cluster 1	Declining trend	Incentive-based funding, catalytic investment programs, performance monitoring
KMeans Cluster 0 / HDBSCAN Cluster 4 & 6	Stable, moderate	Baseline budgeting, routine disbursement protocols, standard procurement procedures
HDBSCAN Noise	Erratic or inconsistent	Resilience-based infrastructure planning, real-time cost auditing, fiscal anomaly monitoring
HDBSCAN Cluster 30 & 32	High CCI and high GRDP	Scalable investment schemes aligned with economic productivity
HDBSCAN Cluster 8 & 24	High CCI but low GRDP	Structural support, logistic cost mitigation, focused capacity-building interventions

Likewise, this study is in line with Mauritsius et al. [30], who emphasized that EV infrastructure location selection in urban areas should be based on cluster analysis to ensure investment efficiency and sustainability.

In addition, the unsupervised learning-based approach to clustering regions based on the temporal pattern of the CCI (Construction Cost Index) from 2014 to 2024 also supports the idea proposed by Gold et al. [31] regarding the importance of exploratory modeling in dealing with spatial and temporal uncertainty in regional infrastructure planning [32][33]. This approach supports the area-based adaptive policy framework developed in their study.

This study also aligns with Trenggono et al. [34], who developed a multidimensional index to assess provincial heterogeneity in Indonesia and subsequently grouped provinces into distinct clusters using PCA-based analysis. Their findings illustrate that cluster-based regional differentiation can reveal meaningful disparities that may be obscured under uniform national assumptions, thereby supporting the broader rationale for data-driven, zone-specific policy design.

Although this study does not explicitly reject the results of previous studies, this study offers a methodological critique of the conventional approach still used in national infrastructure budget planning, namely the use of national average assumptions in determining construction costs [35]. This is in accordance with the criticism put forward by Widiputera et al. [36], who stated that the construction cost index needs to be combined with a regional approach so that the distribution of educational assistance is more targeted [26].

CONCLUSION

This study applied a dual-clustering framework, combining KMeans and HDBSCAN, to classify Indonesian districts and cities based on their multi-year Construction Cost Index (CCI) trajectories. Empirically, KMeans produced three macro-clusters but yielded a highly imbalanced partition in which the dominant cluster contains 474 districts (over 90% of observations). In contrast, HDBSCAN identified 35 micro-clusters and classified 103 districts as noise, thereby exposing finer heterogeneity and explicitly isolating anomalous cost trajectories. Statistical validation confirms that CCI differs significantly across clusters (one-way ANOVA on 2020 CCI: KMeans $F = 1696.49$, $p = 2.35 \times 10^{-226}$; HDBSCAN $F = 370.91$, $p = 9.22 \times 10^{-267}$, excluding noise), while spatial mapping indicates coherent geographic patterning and GRDP-based analysis reveals heterogeneous cost–output relationships that strengthen economic interpretability. Collectively, these findings provide empirical evidence that uniform construction cost assumptions are inadequate for national budgeting in a heterogeneous archipelagic context. Importantly, the hybrid integration of centroid-based and density-sensitive clustering offers a practically relevant basis for differentiated cost zoning, enabling adaptive infrastructure planning through cluster-specific budgeting assumptions, targeted feasibility assessments for atypical districts, and more equitable allocation of public investment. The proposed framework can be extended in future research by incorporating real-time data streams, richer socio-economic covariates, and dynamic clustering approaches to strengthen policy responsiveness over time.

REFERENCES

[1] A. Mahmud, V. V. Gayah, and S. I. Guler, “Modeling economic impacts of transportation infrastructure in Pennsylvania considering spatial dependency across counties,” *Int. J. Transp. Sci. Technol.*, Jan. 2025, doi: 10.1016/j.ijtst.2025.01.006.

[2] M. Suhrab, P. Chen, and A. Ullah, “Digital financial inclusion and income inequality nexus: Can technology innovation and infrastructure development help in achieving

sustainable development goals?," *Technol. Soc.*, vol. 76, p. 102411, Mar. 2024, doi: 10.1016/j.techsoc.2023.102411.

[3] W. P. Humang et al., "Hub-and-spoke network design to optimize government subsidy costs for maritime transportation in Indonesia," *Sustain. Futures*, vol. 9, p. 100602, Jun. 2025, doi: 10.1016/J.SFTR.2025.100602.

[4] R. Aditia and K. Széll, "Belonging matters: How context and inequalities shape student achievement in Indonesia," *Int. J. Educ. Res. Open*, vol. 9, p. 100512, Dec. 2025, doi: 10.1016/J.IJEDRO.2025.100512.

[5] I. S. Rosidin, H. L. F. de Groot, and P. Mulder, "Power Islands. The geography of electricity diffusion in Indonesia, 1975–2018," *Energy Strategy Rev.*, vol. 64, p. 102077, Mar. 2026, doi: 10.1016/J.ESR.2026.102077.

[6] B. V. Oommen, G. A. G. R. Godawatte, and Y. Neilsen, "ERP-Integrated Cost Value Reconciliation Automation in Construction Main Contracting Organizations Using a Hybrid Machine Learning Framework," *Results Eng.*, p. 109862, Mar. 2026, doi: 10.1016/J.RINENG.2026.109862.

[7] Z. Zhou, J. Zhou, B. Zhang, J. Alcalá, and V. Yepes, "The centennial sustainable assessment of regional construction industry under the multidisciplinary coupling model," *Sustain. Cities Soc.*, vol. 101, p. 105201, Feb. 2024, doi: 10.1016/J.SCS.2024.105201.

[8] I. Kousar and M. M. Alam, "Review of wind and earthquake vibrations of skyscrapers: trend, cost, vibration, failure, maintenance, and vibration suppression techniques," *J. Build. Eng.*, vol. 114, p. 114440, Nov. 2025, doi: 10.1016/J.JOBE.2025.114440.

[9] R. Meoded and I. BenDavid-Hadar, "Fiscal decentralization of education: A social network analysis of values underlying local decision-making processes," *Int. J. Educ. Dev.*, vol. 119, p. 103444, Nov. 2025, doi: 10.1016/J.IJEDUDEV.2025.103444.

[10] L. Huang, L. Zhao, and Z. C. Li, "Effects of functional spaces on small logistics facilities: machine learning-based decision-support tool," *Transp. Res. D Transp. Environ.*, vol. 153, p. 105246, Apr. 2026, doi: 10.1016/J.TRD.2026.105246.

[11] B. G. H. Drama, K. Soro, and M. M. Senou, "Sectoral allocation of public investment: Optimizing efficiency in education and health systems in developing economies," *Soc. Sci. Humanit. Open*, vol. 12, p. 102300, Jan. 2025, doi: 10.1016/J.SSAHO.2025.102300.

[12] C. Jacobs-Crisioni, M. Kompil, and L. Dijkstra, "Big in the neighbourhood: Identifying local and regional centres through their network position," *Pap. Reg. Sci.*, vol. 102, no. 2, pp. 421–458, Apr. 2023, doi: 10.1111/PIRS.12727.

[13] A. AlTalhoni, H. Liu, and O. Abudayyeh, "Forecasting Construction Cost Indices: Methods, Trends, and Influential Factors," *Buildings*, vol. 14, pp. 3272, Oct. 2024, doi: 10.3390/buildings14103272.

[14] B. Aslam, A. Maqsoom, H. Inam, M. ul Basharat, and F. Ullah, "Forecasting Construction Cost Index through Artificial Intelligence," *Societies*, vol. 13, no. 10, p. 219, Oct. 2023, doi: 10.3390/soc13100219.

[15] N. Bachmann, S. Tripathi, M. Brunner, and H. Jodlbauer, "The Contribution of Data-Driven Technologies in Achieving the Sustainable Development Goals," *Sustainability*, vol. 14, no. 5, p. 2497, Feb. 2022, doi: 10.3390/su14052497.

[16] S. Saeidlou and N. Ghadiminia, "A construction cost estimation framework using DNN and validation unit," *Build. Res. Inf.*, vol. 52, no. 1–2, pp. 38–48, Feb. 2024, doi: 10.1080/09613218.2023. 2196388.

[17] D. Balaram, B. Dufford, S. Martin, G. A. Negoita, M. Yen, and W. A. Paxton, "Graph-based two-level clustering for electric vehicle usage patterns," *Energy and AI*, vol. 21, p. 100539, Sep. 2025, doi: 10.1016/J.EGYAI.2025.100539.

[18] J. Enes, R. R. Expósito, J. Fuentes, J. L. Cacheiro, and J. Touriño, "A pipeline architecture for feature-based unsupervised clustering using multivariate time series from HPC jobs," *Inf. Fusion*, vol. 93, no. 6, pp. 1–20, May 2023, doi: 10.1016/j.inffus.2022.12.017.

[19] M. Saklani, D. K. Saini, M. Yadav, and P. Siano, "Scalable Data-Driven EV Charging Optimization Using HDBSCAN-LP for Real-Time Pricing Load Management," *Smart Cities*, vol. 8, no. 4, p. 139, Aug. 2025, doi: 10.3390/smartcities8040139.

[20] G. Stewart and M. Al-Khassaweneh, "An Implementation of the HDBSCAN* Clustering Algorithm," *Appl. Sci.*, vol. 12, no. 5, p. 2405, Feb. 2022, doi: 10.3390/app12052405.

[21] A. A. Abdulnassar and L. R. Nair, "Performance analysis of Kmeans with modified initial centroid selection algorithms and developed Kmeans9+ model," *Meas. Sens.*, vol. 25, p. 100666, Feb. 2023, doi: 10.1016/j.measen.2023.100666.

[22] C. A. Carrasco and A. Hernandez-del-Valle, “Energy intensity, economic structure, and capital goods imports in upper-middle income countries: Insights from HDBSCAN clustering,” *J. Environ. Manage*., vol. 339, p. 117840, Aug. 2023, doi: 10.1016/J.JENVMAN.2023.117840.

[23] Y. Song, M. Kalacska, M. Gašparović, J. Yao, and N. Najibi, “Advances in geocomputation and geospatial artificial intelligence (GeoAI) for mapping,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 120, p. 103300, Jun. 2023, doi: 10.1016/J.JAG.2023.103300.

[24] G. Jin and C. Yang, “A systematic intelligent prediction model for residential construction cost based on fuzzy AHP and GA-BP neural network,” *Adv. Eng. Inform.*, vol. 69, p. 103858, Jan. 2026, doi: 10.1016/J.AEI.2025.103858.

[25] L. H. T. Bandória, W. N. Silva, M. C. de Almeida, and B. H. Dias, “Enhancing electrical load profile segmentation for smart campus energy management,” *Energy Build*., vol. 329, p. 115232, Feb. 2025, doi: 10.1016/J.ENBUILD.2024.115232.

[26] “Construction Cost Index Province and Regency/Municipality 2024 - BPS-Statistics Indonesia.” Accessed: Mar. 04, 2026. [Online]. Available: https://www.bps.go.id/en/publication/2024/10/01/a0ce996786e332d81a43d95f/construction-cost-index-province-and-regency-municipality-2024.html

[27] M. Sari, F. Setyawati, M. A. Berawi, and P. Miraj, “Logistic Aerocity: Optimizing financial feasibility through public-private partnerships,” *Transp. Res. Interdiscip. Perspect*., vol. 34, p. 101675, Nov. 2025, doi: 10.1016/J.TRIP.2025.101675.

[28] Z. Min et al., “Enhancing 5G network slicing for IoT traffic with a novel clustering framework,” *Pervasive Mob. Comput*., vol. 104, p. 101974, Nov. 2024, doi: 10.1016/J.PMCJ.2024.101974.

[29] X. Gao, X. Ding, T. Han, and Y. Kang, “Analysis of influencing factors on excellent teachers’ professional growth based on DB-Kmeans method,” *EURASIP J. Adv. Signal Process*., vol. 2022, no. 1, pp. 117-, Dec. 2022, doi: 10.1186/S13634-022-00948-2/TABLES/6.

[30] T. Mauritsius et al., “An analytics framework for predicting electric vehicle charging station demand using spatial and interpretive methods,” *Decis. Anal. J.*, vol. 17, p. 100648, Dec. 2025, doi: 10.1016/J.DAJOUR.2025.100648.

[31] D. F. Gold, P. M. Reed, D. E. Gorelick, and G. W. Characklis, “Advancing Regional Water Supply Management and Infrastructure Investment Pathways That Are Equitable, Robust, Adaptive, and Cooperatively Stable,” *Water Resour. Res*., vol. 59, no. 9, p. e2022WR033671, Sep. 2023, doi: 10.1029/2022WR033671.

[32] R. Goodspeed, R. Liu, D. Gounaridis, C. Lizundia, and J. Newell, “A regional spatial planning model for multifunctional green infrastructure,” *Environ. Plan. B Urban Anal. City Sci*., vol. 49, no. 3, pp. 815–833, Mar. 2022, doi: 10.1177/23998083211033610/SUPPL_FILE/SJ-PDF-1-EPB-10.1177_23998083211033610.PDF.

[33] G. Rani and D. K. Saini, “Need of Integrated Regional Planning Approach for the Decentralisation and Optimisation of Renewable Energy Based Electric Vehicle Infrastructure: A Comprehensive Visualisation,” *Sustainability*, vol. 15, no. 18, p. 13315, Sep. 2023, doi: 10.3390/SU151813315.

[34] S. W. Trenggono et al., “Developing a marine and fisheries blue economy index for Indonesia: A comprehensive assessment framework,” *Sustainable Futures*, vol. 10, p. 101440, Dec. 2025, doi: 10.1016/J.SFTR.2025.101440.

[35] M. A. Masten, A. Poirier, and L. Zhang, “Assessing Sensitivity to Unconfoundedness: Estimation and Inference,” *J. Bus. Econ. Stat.*, vol. 42, no. 1, pp. 1–13, 2024, doi: 10.1080/07350015.2023.2183212.

[36] F. Widiputera, W. Rahayu, S. Suryadi, and R. Rumtini, “Calculation of the Equal Education Cost Index in Indonesia: Using the Relative Importance Regression and Winsorization Approach,” *J. Educ. Soc. Res.*, vol. 14, no. 6, p. 141, Nov. 2024, doi: 10.36941/jesr-2024-0162.