Online-ISSN 3025-3125

Provincial Segmentation Based on District Road Stability in Indonesia: K-Means and Hierarchical Clustering Approach

Heldiansyah [a,1,*], Novi Shintia [b,2], Rustaniah [b,3], Hadi Gunawan [c,4], Muchtar Salim [c,5]

[a] Digital Business Study Program, Politeknik Negeri Banjarmasin, Brigjen Hasan Basri Street, Banjarmasin 70124, Indonesia
[b] Business Administration Study Program, Politeknik Negeri Banjarmasin, Brigjen Hasan Basri Street, Banjarmasin 70124, Indonesia
[c] Civil Engineering Study Program, Politeknik Negeri Banjarmasin, Brigjen Hasan Basri Street, Banjarmasin 70124, Indonesia
[1] heldiansyah@poliban.ac.id *; [2] novi221177@poliban.ac.id; [3] rustaniah@poliban.ac.id; [4] hadi.gunawan@poliban.ac.id; [5] salim@poliban.ac.id
* corresponding author

ARTICLE INFO

ABSTRACT (10PT)

Article history
Received
Revised
Accepted

Keywords
district road stability
provincial clustering
k-means clustering
hierarchical clustering
infrastructure policy

This study segments Indonesian provinces based on district road stability characteristics using K-Means and Hierarchical Clustering approaches. We analyzed district road stability data from 34 provinces during 2016-2023, including total road length, stable road conditions, and unstable road conditions. Data preprocessing included cleaning, normalization using min-max scaling, and feature selection. Results showed optimal clustering with k=4, achieving silhouette coefficient of 0.647 for K-Means and 0.623 for Hierarchical Clustering. Four distinct provincial clusters emerged: Optimal Infrastructure Provinces (>80% stability), Developing Infrastructure Provinces (60-80% stability), Infrastructure Challenge Provinces (<60% stability with extensive networks), and Limited Infrastructure Provinces (small networks with variable stability). The Adjusted Rand Index of 0.78 demonstrated high agreement between methods. This segmentation provides evidence-based insights for targeted infrastructure policy formulation in Indonesia.

1. Introduction

Transportation infrastructure serves as the backbone of economic development in Indonesia, with district roads playing a strategic role in supporting local connectivity and community development. The quality and stability of district roads significantly impact regional economic growth, access to public services, and overall quality of life. Indonesia's district road stability data reveals substantial variations across provinces, with considerable disparities influenced by geographical characteristics, fiscal capacity, and infrastructure management effectiveness.

Road stability assessment in Indonesia follows the International Roughness Index (IRI) methodology, where roads with IRI values ≤8 m/km are classified as stable, while roads exceeding this threshold are considered unstable [1]. This standardized measurement enables consistent evaluation across provinces and facilitates comparative analysis of infrastructure conditions nationwide.

Recent advances in machine learning have revolutionized infrastructure analysis and policy formulation. Clustering algorithms, particularly K-Means and Hierarchical Clustering, have demonstrated effectiveness in transportation infrastructure analysis by identifying hidden patterns and grouping regions with similar characteristics [2], [3]. These unsupervised learning techniques enable evidence-based policy development by revealing natural groupings within complex infrastructure datasets.

Previous infrastructure studies in Indonesia have predominantly focused on technical aspects with limited comprehensive analysis using machine learning approaches for provincial clustering. International studies have successfully applied clustering techniques for transportation infrastructure analysis, demonstrating significant potential for policy applications [4]. This study addresses the gap by applying machine learning clustering to segment Indonesian provinces based on district road stability characteristics, providing evidence-based insights for targeted policy formulation and resource allocation strategies.

2. Method

2.1. Data Collection

This study employs a quantitative research approach with exploratory design to identify patterns in Indonesian district road stability data using unsupervised machine learning techniques [5]. The research utilizes secondary data from the Directorate General of Highways, Ministry of Public Works of the Republic of Indonesia, ensuring reliability and consistency in measurement methodologies.

The study population comprises all 34 Indonesian provinces during 2016-2023, capturing recent infrastructure development trends. The dataset comprises six key variables: province code, province name, total district road length (kilometers), stable road length (kilometers), stable road percentage, unstable road length (kilometers), and unstable road percentage. Data validation included completeness verification, internal consistency checks, and temporal consistency analysis.

2.2. Data Preprocessing

Data preprocessing procedures include data cleaning, missing value imputation, outlier detection and treatment, and normalization [6]. Missing value analysis revealed less than 5% missing data across variables. Missing value treatment employed temporal interpolation for random missing values and mean imputation based on similar provinces for systematic missing data.

Outlier detection using Interquartile Range methods identified potential anomalies, which underwent contextual analysis considering geographical and administrative factors. Data normalization employed min-max scaling to transform all numerical variables to [0,1] range using the formula (x - min)/(max - min), ensuring equal contribution of all variables in distance-based clustering algorithms.

2.3. Clustering Implementation

Clustering implementation utilized Python programming environment with scikit-learn, pandas, and numpy libraries [7]. K-Means clustering implementation began with optimal cluster determination using elbow method analysis, examining Sum of Squared Errors values for k ranging from 2 to 10 [8]. Silhouette analysis validated cluster number selection by considering both cohesion and separation aspects.

Algorithm configuration employed k-means++ initialization with maximum iterations of 300 and convergence tolerance of 1e-4. Multiple random initializations (10 runs) addressed initialization sensitivity. Hierarchical clustering implementation utilized agglomerative approach with Ward linkage method, which minimizes within-cluster variance and produces balanced clusters [9].

2.4. Evaluation and Validation

Clustering evaluation employed multiple validation metrics including silhouette coefficient, Davies-Bouldin index, and Calinski-Harabasz index. Silhouette coefficient measures clustering quality by comparing average intra-cluster distance with average nearest-cluster distance [10]. Davies-Bouldin index provides complementary assessment by measuring the ratio of within-cluster scatter to between-cluster separation [11]. Stability validation assessed clustering consistency across different algorithm runs and parameter settings using Adjusted Rand Index [12].

3. Results and Discussion

3.1. Optimal Cluster Determination and Clustering Results

Optimal cluster number determination through elbow method analysis revealed clear inflection point at k=4, where Sum of Squared Errors reduction began to level off significantly. Silhouette analysis corroborated this finding, demonstrating maximum average silhouette coefficient of 0.647 for k=4 configuration, indicating reasonable to good clustering structure.

K-Means clustering with k=4 successfully segmented the 34 provinces into four distinct clusters with balanced distribution: Cluster 1 (9 provinces), Cluster 2 (8 provinces), Cluster 3 (10 provinces), and Cluster 4 (7 provinces). Hierarchical clustering produced comparable results with slight variations in provincial assignments, particularly for provinces with boundary characteristics.

The first cluster is optimal infrastructure provinces, comprising 9 provinces characterized by high road stability rates exceeding 80% and moderate total road lengths. This cluster includes several Java-based provinces and regions with advanced infrastructure development programs. The second cluster is developing infrastructure provinces, including 8 provinces with moderate road stability rates (60-80%) and steady improvement trends. The third cluster is infrastructure challenge provinces, containing 10 provinces featuring extensive district road networks but relatively low stability rates below 60%. The fourth cluster is limited infrastructure provinces, encompassing 7 provinces with smaller total road networks but variable stability rates.

3.2. Method Comparison and Policy Implications

Comparative analysis between K-Means and Hierarchical clustering demonstrated high agreement with Adjusted Rand Index of 0.78, indicating substantial consistency between methods. Performance evaluation showed K-Means achieving silhouette coefficient of 0.647 and Davies-Bouldin index of 0.89, while Hierarchical clustering achieved 0.623 and 0.95 respectively. K-Means demonstrated superior computational efficiency (0.23 seconds vs 1.47 seconds) and cluster compactness.

The clustering results provide strong evidence base for differentiated policy approaches tailored to specific provincial characteristics. For Cluster 1 provinces, policy focus should emphasize knowledge sharing and capacity building to support other provinces' development. Cluster 2 provinces require sustained support for ongoing development initiatives with attention to maintaining improvement momentum. Cluster 3 provinces demand comprehensive restructuring of infrastructure management approaches, combining increased resource allocation with technical capacity development.

The four-cluster structure reveals meaningful patterns that align with known geographical, economic, and administrative characteristics across provinces. The clustering pattern demonstrates that provinces with larger road networks do not automatically achieve better stability rates, suggesting that expansion strategies must be balanced with maintenance capacity development. This finding has significant implications for infrastructure policy, emphasizing sustainable development approaches considering long-term maintenance requirements.

4. Conclusion

This study demonstrated the effectiveness of machine learning clustering approaches for segmenting Indonesian provinces based on district road stability characteristics. The analysis revealed four distinct provincial clusters with clear policy implications: Optimal Infrastructure Provinces, Developing Infrastructure Provinces, Infrastructure Challenge Provinces, and Limited Infrastructure Provinces.

Both K-Means and Hierarchical clustering methods proved effective, with K-Means demonstrating superior computational efficiency and cluster compactness, while Hierarchical clustering provided superior visualization through dendrograms. The high agreement between methods validates the robustness of identified provincial groupings and supports confident policy recommendations.

The clustering results provide evidence base for differentiated policy approaches tailored to each cluster's specific characteristics rather than uniform national policies. This segmentation transcends simple geographical divisions, revealing patterns influenced by economic capacity, administrative efficiency, and geographical challenges. Key recommendations include establishing knowledge sharing mechanisms, providing sustained technical assistance, implementing comprehensive infrastructure management restructuring, and developing tailored approaches for different provincial segments.

Future research opportunities include extending analysis to multimodal transportation systems, incorporating additional performance indicators, and conducting longitudinal studies to assess clustering stability and policy impact over time.

Declarations

Author contribution. Conducting study conceptualization, methodology design, data collection and analysis, results interpretation, and manuscript preparation.

Funding statement. This research was supported by the P3M Internal Research Grant from Politeknik Negeri Banjarmasin.

Data and Software Availability Statements

The district road stability datasets used in this study were obtained from official government sources, specifically the Ministry of Public Works of the Republic of Indonesia. Raw datasets are subject to government data sharing policies and are available upon reasonable request through proper institutional channels. Analysis code implemented in Python using scikit-learn, pandas, and numpy libraries is available from the corresponding author upon request for research purposes.

References

[1] O. G. Dela Cruz, C. A. Mendoza, and K. D. Lopez, “International Roughness Index as Road Performance Indicator: A Literature Review,” IOP Conf Ser Earth Environ Sci, vol. 822, no. 1, p. 012016, Jul. 2021, doi: 10.1088/1755-1315/822/1/012016.

[2] M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” Inf Sci (N Y), vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[3] M. Vichi, C. Cavicchia, and P. J. F. Groenen, “Hierarchical Means Clustering,” J Classif, vol. 39, no. 3, pp. 553–577, Nov. 2022, doi: 10.1007/s00357-022-09419-7.

[4] R. Saha, M. T. Tariq, M. Hadi, and Y. Xiao, “Pattern Recognition Using Clustering Analysis to Support Transportation System Management, Operations, and Modeling,” J Adv Transp, vol. 2019, pp. 1–12, Dec. 2019, doi: 10.1155/2019/1628417.

[5] W. Suo and J. Zhao, “Exploring the Streetscape Perceptions from the Perspective of Salient Landscape Element Combination: An Interpretable Machine Learning Approach for Optimizing Visual Quality of Streetscapes,” Land (Basel), vol. 14, no. 7, p. 1408, Jul. 2025, doi: 10.3390/land14071408.

[6] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan, “An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity,” Decision Analytics Journal, vol. 9, p. 100341, Dec. 2023, doi: 10.1016/j.dajour.2023.100341

[7] F. AlShammari, “Implementation of Clustering using K-Means in Python,” Int J Comput Appl, vol. 186, no. 40, pp. 12–17, Sep. 2024, doi: 10.5120/ijca2024923990.

[8] E. Umargono, J. E. Suseno, and S. K. Vincensius Gunawan, “K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula,” in Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019), Paris, France: Atlantis Press, 2020. doi: 10.2991/assehr.k.201010.019.

[9] P. Yildirim and D. Birant, “K-Linkage: A New Agglomerative Approach for Hierarchical Clustering,” Advances in Electrical and Computer Engineering, vol. 17, no. 4, pp. 77–88, 2017, doi: 10.4316/AECE.2017.04010.

[10] F. Batool and C. Hennig, “Clustering with the Average Silhouette Width,” Comput Stat Data Anal, vol. 158, p. 107190, Jun. 2021, doi: 10.1016/j.csda.2021.107190.

[11] F. Ros, R. Riad, and S. Guillaume, “PDBI: A partitioning Davies-Bouldin index for clustering evaluation,” Neurocomputing, vol. 528, pp. 178–199, Apr. 2023, doi: 10.1016/j.neucom.2023.01.043.

[12] J. M. Santos and M. Embrechts, “On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification,” 2009, pp. 175–184. doi: 10.1007/978-3-642-04277-5_18.