 ISSN:2302-8149
e-ISSN:2540-9719

Clustering of Indonesian Provinces based on Demographic Characteristics using DBSCAN Algorithm

[1]Sasmita, [2]Rizal*, [3]Lidya Rosnita
[1,2,3]Informatics Engineering Study Program, Faculty of Engineering, Malikussaleh University
Lhokseumawe, Indonesia
*e-mail: rizal@unimal.ac.id

(***received*****:** 24 April 2026, ***revised*****:** 11 May 2026, ***accepted*****:** 15 May 2026)

Abstract

This research aims to implement the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm for clustering provinces in Indonesia based on demographic characteristics using a Web GIS system. The study utilizes demographic data of 34 provinces from 2020–2023, covering population size, population growth rate, poverty percentage, population density, and gender ratio. DBSCAN was applied with parameters epsilon (ε) = 0.50 and minimum samples = 2, evaluated using the Davies-Bouldin Index (DBI). The results show that DBSCAN successfully identified two consistent main clusters throughout the period: the High Cluster consisting of three provinces (West Java, Central Java, East Java), the Medium Cluster including 29–30 provinces, and DKI Jakarta as a single noise point.. Clustering quality varied with DBI values ranging from 0.3839 to 0.4123, with the best quality in 2020 and the highest internal diversification in 2023. Riau Islands fluctuated as a noise point in 2021 before returning to stability. The developed Web GIS system successfully integrated interactive map visualization using OpenStreetMap and Leaflet, providing a comprehensive spatio-temporal analysis dashboard. This study demonstrates the effectiveness of DBSCAN in detecting demographic patterns and regional anomalies, while producing a visualization tool that supports regional development planning and national demographic policies in Indonesia.

Keywords: DBSCAN algorithm, demographic characteristics, demographic clustering, Indonesian provinces, web GIS system

1 Introduction

Indonesia, as the world's largest archipelagic country consisting of 34 provinces, possesses exceptionally high and unique demographic complexity. Each province exhibits distinct characteristics in terms of population size, growth rate, population density, poverty levels, gender ratio, age structure, education levels, and welfare conditions. This diversity has resulted in significant regional disparities in human development and resource distribution. Such conditions create substantial challenges in formulating effective and equitable national development policies, including infrastructure planning, budget allocation, public service distribution, and community empowerment strategies, which ultimately affect the achievement of national development targets [1]. Demographic inequalities among provinces further exacerbate difficulties in optimal resource allocation and the design of targeted programs, as reflected in the persistent significant gaps in the Human Development Index (HDI) across regions over the past two decades. Fragmented migration data further complicates the understanding of socio-economic dynamics. A data-driven approach has become increasingly urgent to provide objective insights into regional needs and support evidence-based policymaking [2]. Cluster analysis has emerged as a powerful technique for grouping regions based on similarities in demographic profiles, thereby enabling prioritized interventions for densely populated areas or low-growth regions, as well as more systematic monitoring of development progress [3].

To overcome the limitations of traditional clustering methods such as K-Means, which require a predetermined number of clusters and are less effective in handling uneven data distributions, this study employs the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. DBSCAN groups data based on density and is capable of detecting clusters of arbitrary shapes while automatically identifying outliers (noise points), making it highly suitable for complex and non-uniform demographic data [4]. Previous applications of DBSCAN have demonstrated

 ISSN:2302-8149
e-ISSN:2540-9719

promising results, such as the identification of industrial centers in Situbondo Regency using spatial data from Google Maps and earthquake event clustering in Regional VII with evaluation using the silhouette coefficient [5] [6]. However, the comprehensive implementation of the DBSCAN algorithm for provincial-level demographic clustering in Indonesia remains very limited. Most existing studies focus primarily on parameter tuning or other domains, leaving a clear research gap in the systematic application of DBSCAN to multi-year demographic datasets for policy support [7] [8].

This study aims to address this gap through two interconnected core issues. First, it explores the application of the DBSCAN algorithm to uncover more accurate patterns and structures of demographic characteristics among provinces, including the detection of regional anomalies that have been insufficiently identified due to the limitations of conventional methods. Second, this research seeks to design and implement a Web GIS-based system that not only integrates the DBSCAN algorithm but also provides interactive spatio-temporal visualization, thereby serving as a practical instrument for stakeholders in supporting evidence-based and sustainable regional development planning [9]. This study is expected to contribute to the advancement of data mining methodologies in demographic analysis, provide a practical visualization tool for policymakers, and support more inclusive, targeted, and sustainable development policies through a better understanding of provincial cluster characteristics. The research is limited to the use of the DBSCAN algorithm, historical data from the period 2020–2023 sourced from www.bps.go.id, and five key variables: population size, population growth rate, poverty percentage, population density, and gender ratio. Clustering visualization employs color coding: red for the high cluster, yellow for the medium cluster, and green for the low cluster.

2 Literature Review

Recent studies on the application of clustering algorithms for demographic and spatial data in Indonesia show an increasing trend toward density-based approaches to handle uneven data distributions and the presence of outliers. Several studies have successfully utilized DBSCAN in public health contexts, such as determining priority scales for stunting intervention in toddlers using parameters epsilon 114 and minPts 2, resulting in three priority clusters with a silhouette coefficient of 0.512 [10] [11]. Similarly, the application of DBSCAN for earthquake data clustering in Indonesia successfully identified spatial patterns with silhouette coefficient and Davies-Bouldin Index evaluation, although it remained limited to seismic variables [12]. In the socio-economic domain, recent comparative studies applied DBSCAN alongside K-Means and Gaussian Mixture Model for provincial welfare clustering, where DBSCAN demonstrated superior performance in forming two main clusters (high-welfare and low-welfare) and detecting noise points more effectively than centroid-based methods (2025). Another recent work employed DBSCAN for clustering health worker ratios across provinces, successfully identifying provinces with extreme ratios as noise (DBSCAN Method in Clustering Provinces, 2025). However, most of these studies remain focused on single-sector domains or single-year datasets, thus failing to capture the temporal dynamics and multi-dimensional nature of demographic characteristics comprehensively [13].

Meanwhile, provincial-level clustering in Indonesia still predominantly relies on partitioning algorithms such as K-Means or K-Medoids, which require a predetermined number of clusters and show limitations in handling irregular distributions and non-spherical cluster shapes. For example, K-Means clustering of superior vegetables produced three clusters with a silhouette score of 0.6798 [14], while K-Medoids was used to compare academic absorption rates between urban and rural students during the pandemic. Although recent geospatial studies highlight DBSCAN's superiority in noise detection and irregular cluster formation, its integration with interactive visualization systems remains rare. Moreover, the combination of data mining techniques with Web GIS for provincial demographic analysis is still limited; most existing works produce only statistical outputs or static maps without providing interactive spatio-temporal dashboards that can be practically utilized by policymakers and stakeholders [15]. Previous website-based clustering studies

 ISSN:2302-8149
e-ISSN:2540-9719

tend to be domain-specific, such as scholarship selection or post-pandemic student interest analysis, and have not addressed national-scale provincial clustering using comprehensive multi-year demographic variables [16].

While DBSCAN is widely recognized for its efficacy in identifying density-based patterns, its potential in longitudinal demographic analysis remains underutilized. A critical gap persists: the absence of a systematic framework that applies this algorithm to multi-year, multi-variable datasets and embeds the results within a functional Web GIS environment. This study fulfills that dual objective. By implementing DBSCAN alongside an interactive OpenStreetMap and Leaflet dashboard, the research moves beyond isolated statistical evaluation to offer a dynamic spatio-temporal analysis. The core innovation is the convergence of sophisticated spatial data mining and intuitive geovisualization. Departing from earlier studies that treat clustering and visualization as separate endeavors, this integrated approach yields both methodological advancement and practical utility, equipping stakeholders with an empirical, replicable blueprint for provincial demographic planning.

3 Research Method

This study was conducted at Malikussaleh University, Lhokseumawe, Aceh, Indonesia, from October 2024 until completion. A quantitative approach utilizing secondary data was employed. Demographic data of 34 provinces for the period 2020–2023 were obtained from the official website of the Central Bureau of Statistics (*www.bps.go.id*) through online downloading. The variables used included population size, population growth rate, poverty percentage, population density, and gender ratio.

The system development adopted the sequential Waterfall method, covering stages of requirements analysis, design, implementation, and testing. The object of the study was the demographic characteristics of all provinces in Indonesia. The main tools used were Python with pandas, numpy, and scikit-learn libraries for DBSCAN implementation, along with PHP, MySQL, Leaflet.js, and OpenStreetMap for developing the Web GIS system.

All variables were normalized using Min-Max Normalization prior to analysis. The DBSCAN algorithm was applied with parameters $\varepsilon = 0.50$ and minimum samples = 2. Clustering quality was evaluated using the Davies-Bouldin Index (DBI). The analysis results were integrated into a Web GIS dashboard to provide interactive map visualization and spatio-temporal analysis. System testing was performed using white-box testing to ensure the functionality and accuracy of the algorithm.

4 Results and Analysis

4.1 Data Processing

Data processing in this study was carried out in several stages to ensure that the data were suitable for clustering using the DBSCAN algorithm. The stages include data collection, data cleaning, normalization, dataset construction, clustering, and evaluation.

1. Data Collection

The data used in this study are secondary data from 34 provinces in Indonesia for the period 2020–2023, obtained from the Central Statistics Agency (BPS). The variables used include:

a. Total population
b. Population growth rate
c. Poverty percentage
d. Population density
e. Sex ratio

The data were collected annually, forming a simple time-series dataset.

2. Data Cleaning

This stage aims to ensure data quality before analysis. The processes include:

a. Checking for missing values
b. Removing duplicate data
c. Adjusting numerical formats

 ISSN:2302-8149
e-ISSN:2540-9719

The results show that the dataset is complete and contains no missing values; therefore, no imputation process was required.

3. Data Normalization

Since each variable has a different scale, normalization was performed using the Min-Max. Normalization ensures that each variable contributes equally to the distance calculation.

Table 1 Example of normalized data

Province	Population	Growth	Poverty	Density	Sex Ratio
West Java	0.95	0.60	0.45	0.90	0.52
Central Java	0.90	0.55	0.50	0.85	0.51
Province	**Population**	**Growth**	**Poverty**	**Density**	**Sex Ratio**
DKI Jakarta	0.70	0.40	0.20	1.00	0.50
Aceh	0.30	0.35	0.60	0.25	0.50

The normalization results show that all values fall within the range of 0–1.

4. Dataset Construction for Clustering

The final dataset was structured as a numerical matrix consisting of:

a. 34 rows (provinces)
b. columns (demographic variables)

This dataset was used as the main input for the DBSCAN algorithm.

5. DBSCAN Clustering Process

Clustering was performed using the DBSCAN algorithm with the following parameters:

a. Epsilon (ε) = 0.50
b. Minimum samples = 2

The clustering process includes:

a. Calculating distances between data points using Euclidean Distance
b. Identifying core points based on ε and minimum samples
c. Grouping data based on density
d. Detecting noise (outliers)

6. Clustering Results

Table 2 DBSCAN clustering results (2020–2023)

Year	High Cluster	Medium Cluster	Noise
2020	3 provinces	30 provinces	1 province
2021	3 provinces	29 provinces	2 provinces
2022	3 provinces	30 provinces	1 province
2023	3 provinces	30 provinces	1 province

Table 3 Cluster composition

Cluster	Provinces	Characteristics
High	West Java, Central Java, East Java	High population and density
Medium	Majority of other provinces	Relatively homogeneous
Noise	DKI Jakarta	Outlier

7. Clustering Evaluation (DBI)

Evaluation was conducted using the Davies-Bouldin Index (DBI).

Table 4 DBI values

Year	DBI
2020	0.3839
2021	0.39
2022	0.40

ISSN:2302-8149
e-ISSN:2540-9719

2023	0.4123

The relatively low DBI values indicate that the clustering results have good quality.

8. Web GIS Implementation

The clustering results were integrated into a Web GIS system that provides:

a. Interactive map visualization using OpenStreetMap
b. Color-based cluster representation
c. Year-based analytical dashboard

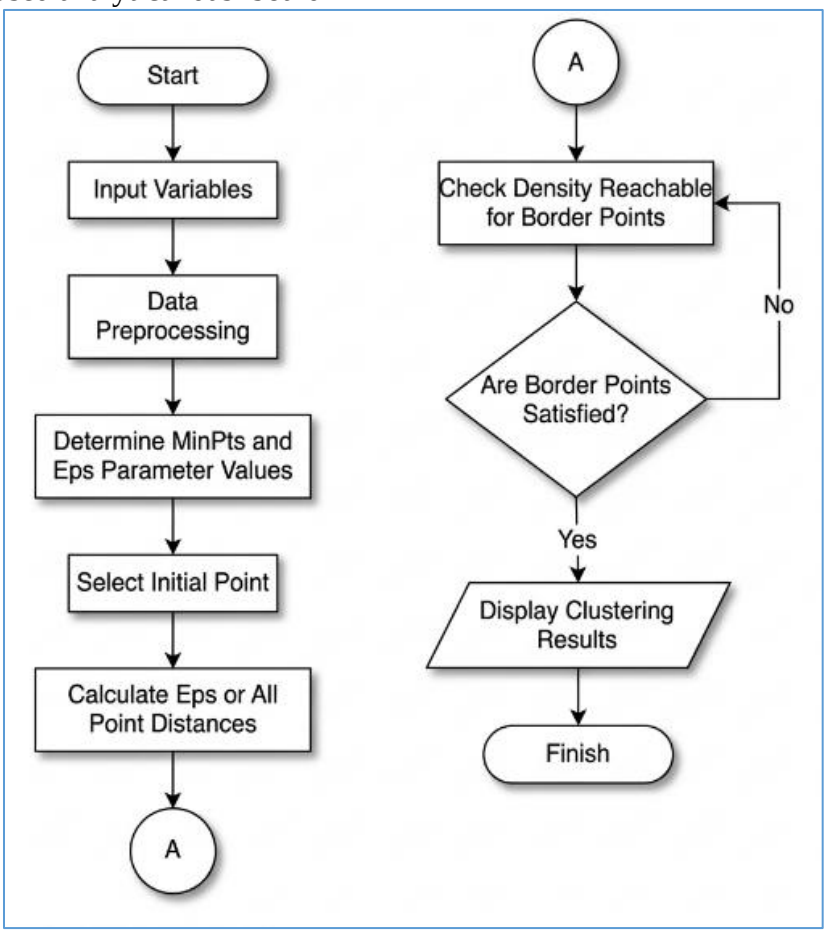

Figure 3 System schematic

The results indicate that the DBSCAN algorithm is effective in clustering Indonesian provinces based on demographic characteristics. Provinces with large populations and high density, such as West Java, Central Java, and East Java, consistently form the high cluster, indicating that population size and density are dominant factors.

DKI Jakarta is identified as noise due to its extreme characteristics, particularly in population density, which significantly differs from other provinces. This demonstrates DBSCAN's strength in automatically detecting outliers.

The stability of clusters from 2020 to 2023 indicates that Indonesia's demographic patterns are relatively consistent. However, the change observed in the Riau Islands in 2021 suggests that local dynamics can influence clustering results.

Compared to K-Means, DBSCAN has advantages as it does not require predefined cluster numbers and can handle non-uniform data distributions. This makes it more suitable for complex demographic data.

 ISSN:2302-8149
e-ISSN:2540-9719

This study has several strengths, including the use of multi-year data (2020–2023), five key demographic variables, and evaluation using the Davies-Bouldin Index. Additionally, the integration with a Web GIS system provides enhanced visualization capabilities.
The novelty of this research lies in the integration of density-based clustering with spatio-temporal visualization, offering not only statistical insights but also a decision-support tool.

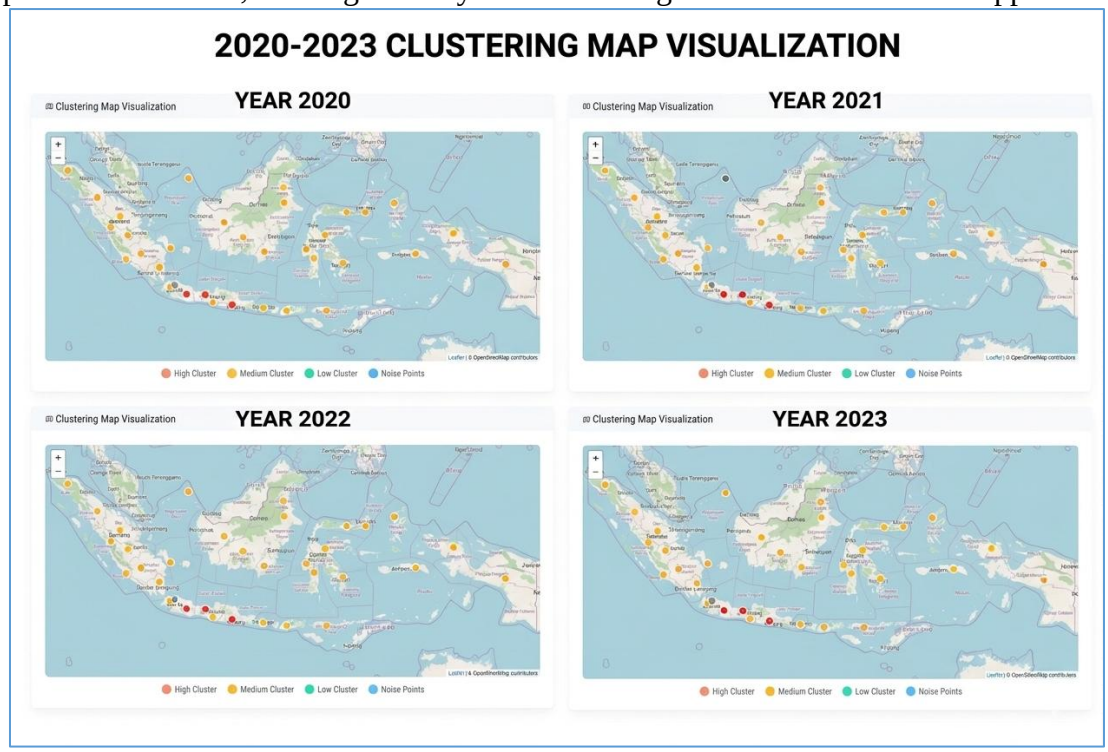

Figure 4 Cluster map visualization 2020-2023

Figure 5 DBSCAN clustering results display for 2020-2023

5 Conclusion

The conclusion of this study shows that the DBSCAN algorithm was successfully applied to cluster provinces in Indonesia based on demographic characteristics effectively. The analysis results indicate the formation of two main clusters that remained consistent throughout the 2020–2023 period, namely a high cluster dominated by provinces with large populations and high population

 ISSN:2302-8149
e-ISSN:2540-9719

density, and a medium cluster comprising the majority of other provinces, while DKI Jakarta was identified as a noise point due to its extreme characteristics. The relatively low Davies-Bouldin Index (DBI) values indicate that the clustering quality is good. In addition, the integration with a Web GIS system provides interactive and informative spatio-temporal visualization. Therefore, this study demonstrates that DBSCAN is not only effective in detecting patterns and anomalies in demographic data but also capable of supporting data-driven decision-making in regional development planning.

6. Reference

[1] A. F. Abdillah, "*Implementation Clustering Diabetes Suffering Areas using Web-based DBSCAN Algorithm North Aceh District*," *Jurnal Informasi dan Teknologi*, 2025.

[2] R. T. Adek, "*Online Newspaper Clustering in Aceh using the Agglomerative Hierarchical Clustering Method*," *International Journal of Engineering, Science & Information Technology (IJESTY)*, 2022.

[3] A. N. Fauzan, "*Analysis of Hotels Spatial Clustering in Bali: Density-based Spatial Clustering of Application Noise (DBSCAN) Algorithm Approach*," *Journal of Sciences and Data Analyst*, 2022.

[4] R. A. Fahmi, "*A Theil Decomposition of Regional Grouping in Indonesia's Human Development Index*," *Economics Development Analysis Journal*, 2024.

[5] T. D. Harjanto, "Analisis Penetapan Skala Prioritas Penanganan Balita Stunting menggunakan Metode DBSCAN Clustering," *Jurnal Rekursif*, 2021.

[6] L. M. Harahap, "Klastering Sayuran Unggulan menggunakan Algoritma K-Means," *Jurnal Teknik Informatika dan Sistem Informasi*, 2022.

[7] W. P. Handayani, "*Subduction and Local Fault Earthquake Analysis using ST-DBSCAN Clustering Algorithm in the Special Region of Yogyakarta (DIY)*," *Interconnection of Islam and Science Journal*, 2025.

[8] A. Kadir, *Pengenalan Sistem Informasi Edisi Revisi*. Yogyakarta, Indonesia: Penerbit Andi, 2013.

[9] L. M. Maharani, "*DBSCAN Method in Clustering Provinces in Indonesia based on Ratio of Health and Medical Personnel in 2023*," *UNP Journal of Statistics and Data Science*, 2025.

[10] S. M. Sabrina, "*Clustering Analysis of Provincial in Indonesia based on the 2023 Human Development Index Indicators using the K-Medoids Algorithm*," *Jurnal Matematika UNAND*, 2025.

[11] N. P. Sutramiani, "*The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping based on Asset Value and Turnover*," *Journal of Information Systems Engineering and Business Intelligence*, 2024.

[12] H. Tohari, *Analisis Serta Perancangan Sistem Informasi Melalui Pendekatan UML*. Yogyakarta, Indonesia: Penerbit Andi, 2014.

[13] Y. Homaidi and A. L. Yanto, "Implementasi Metode *Clustering* dengan Algoritma DBSCAN untuk Identifikasi Sentra Industri berbasis *Google Map*," *G-Tech: Jurnal Teknologi Terapan*, 2024.

[14] L. S. Lestari, "Analisis Clustering Penduduk berdasarkan Kelompok Umur dengan K-Means dan Hierarchical Clustering untuk Perencanaan Demografi," *UNTAR (Jurnal Komputer dan Informatika)*, 2025.

[15] R. T. Dewanto, "*Leveraging Big Data for Indonesia's Immigration Policy: Opportunities and Limitations*," *JATI (Jurnal Mahasiswa Teknik Informatika)*, 2025.

[16]. R. Tjut adek, "Jurnal Sistem dan Teknologi Informasi Sistem Informasi Geografis Pemetaan dan Penentuan Lokasi Wisata Alam Strategis dengan Metode *Simple Additive Weighting* (SAW)," *Positif : Jurnal Sistem dan Teknologi Informasi*, 2023.