Knowledge Engineering and Data Science (KEDS)
Vol 7, No 2, December 2024, pp. 187–199

pISSN 2597-4602
eISSN 2597-4637

A Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) Approach for Identifying Potential Villages in Buleleng Regency

Dina Nur Amalina [1], Achmad Fauzan [2,*]

Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia
Kaliurang Street No.Km. 14,5, Krawitan, Umbulmartani, Ngemplak, Sleman, Yogyakarta 55584, Indonesia
[1] *dinanuramalina12@gmail.com;* [2] *achmadfauzan@uii.ac.id**
** corresponding author*

ARTICLE INFO

Article history:
Received 19 September 2024
Revised 14 November
Accepted 19 December 2024
Published online 24 December 2024

Keywords:
HDBSCAN
Minimum Spanning Tree
Noise
Village Potential

ABSTRACT

Buleleng Regency, located in Bali Province, possesses diverse village potential, including agricultural production and tourist attractions. However, this potential has not been fully optimized. Therefore, it is important to enhance village potential by clustering villages based on their specific characteristics to identify and prioritize those requiring special attention. This approach aims to promote equitable village development and reduce poverty levels. This study clusters villages in Buleleng Regency based on their potential using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) method. The data utilized in this study comprises village potential data obtained from the Buleleng Regency Statistics Office (BPS) for all districts and the Statistical Service Information System. The variables used in this study are based on aspects of population, communication, tourism, trade, health, religion, social affairs, and public welfare. Tuning parameters were performed to determine the optimal parameters, resulting in optimal parameters, such as minimum cluster size = five and minimum samples = 2, which produced two main clusters. The first cluster comprises six villages, while the second includes 118 villages. Additionally, a noise cluster representing outliers, consisting of 24 villages, was identified. The findings indicate that the first cluster exhibits higher village potential than the second cluster. Based on these results, it is recommended that the government prioritize the second cluster when designing and implementing targeted programs and policies to reduce poverty by developing village potential.

I. Introduction

Buleleng Regency, located in the northern part of Bali Island, covers an area of 1,365.88 km². Due to its geographical position along Bali's northern coastline, several villages directly border the sea. Fifty-three villages, accounting for 35% of all villages in Buleleng Regency, are in coastal areas. This geographic characteristic indicates that Buleleng Regency possesses significant potential in marine resources. Such potential presents an opportunity for the local government to enhance economic development through agricultural production and by leveraging fisheries and marine-based tourism as key economic sectors [1][2]. However, this potential remains underutilized. Systematic efforts are needed to identify and group villages based on their potential to optimize resources, reduce poverty, and promote village development.

Cluster analysis is a powerful method for grouping villages based on their potential, classifying villages into homogeneous clusters according to their unique characteristics. Cluster analysis, or clustering, refers to organizing a given collection of data objects into distinct groups. Within each group, objects share similarities, while those in different groups exhibit dissimilar characteristics [3]. Clustering algorithms based on density provide several benefits compared to conventional clustering techniques like k-means and hierarchical clustering. Clustering methods have evolved into a diverse range of approaches, including agglomerative clustering, divisive clustering, fuzzy clustering, and density-based clustering [4]. Among the various density-based clustering methods, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is favored for its

https://doi.org/10.17977/um018v7i22024p187-199

robustness in handling noise within datasets [5][6]. This algorithm effectively manages datasets with varying densities and cluster sizes.

HDBSCAN enhances the DBSCAN algorithm by integrating hierarchical clustering and identifying flat clusters based on cluster stability. Additionally, it can detect varying densities across geographic regions [7]. In contrast to DBSCAN, which relies on a fixed density threshold, HDBSCAN leverages its hierarchical structure and the minimum spanning tree approach to detect clusters with varying densities [8]. Some studies related to HDBSCAN include in the implementation of HDBSCAN in machine-learning libraries [6], concerning ship trajectory clustering [9], to create Gaussian mixture models [10], and in text insertion clustering [11]. HDBSCAN has applied in economic and energy-related clustering at the national level [12]. HDBSCAN was also utilized to detect urban areas of interest by analyzing Flickr data from Seoul collected in 2019 and 2020 [13].

Nevertheless, applying the HDBSCAN method remains relatively rare in clustering the potential of a region, such as a village potential. Most studies still rely on hierarchical or partition-based methods that utilize square error clustering, which employ K-Means [14], Self-Organizing Map (SOM) [15], as well as results from GWR and Moran’s I [16]. As an alternative, density-based approaches have gained increasing attention in clustering analysis, mainly due to their ability to identify more complex patterns without requiring a predetermined number of clusters. One such method that offers this advantage is HDBSCAN, which can effectively group villages based on their potential in a more adaptive manner. Given its flexibility in handling variations in cluster density and its robustness against outliers, HDBSCAN has the potential to serve as a more accurate solution for spatial analysis in identifying potential villages.

By applying the HDBSCAN method, areas with similar characteristics can be grouped, facilitating the government in identifying and prioritizing villages or sub-districts in Buleleng Regency that require special attention based on their specific needs and potential. This approach aims to promote village development, alleviate poverty, address regional disparities, and ensure equitable opportunities for all villages to enhance the welfare of their residents. The findings are expected to provide comprehensive insights into the characteristics of village potential, facilitating the formulation of more targeted and efficient development policies. Such insights will help identify areas requiring special attention and enable the design of tailored development strategies. The clustering results using HDBSCAN can assist local governments in designing data-driven and targeted rural development policies. Accurate cluster identification enables the optimization of resource allocation, more effective development interventions, and increased efficiency in community empowerment programs. Additionally, this approach supports implementing Smart Village initiatives and sustainable planning. Thus, this study serves as a scientific foundation for more strategic decision-making in rural development in Buleleng Regency.

II. Method

This section describes the research methodology used in this study to identify and cluster village potential in Buleleng Regency. Selecting an appropriate method is crucial in ensuring accurate and relevant analysis. Considering the data’s complexity and the villages' varying characteristics, this study employs the HDBSCAN approach as the primary clustering technique. The research process includes data collection, sample selection, data preprocessing, and implementing clustering algorithms to obtain optimal results. The following subsections provide a detailed explanation of the study population, research sample, and the data analysis stages.

A. Population and Research Sample

The population in this study comprises all potential villages and sub-districts in Buleleng Regency, Bali Province. The data in this study were obtained from village-level statistical data in Buleleng Regency for the year 2022 published by the Buleleng Regency Statistics Office (BPS) and Statistical Service Information System (https://bulelengkab.bps.go.id/). The data encompasses various social, economic, and demographic indicators relevant for clustering analysis using HDBSCAN. The variables used are based on strategic issues related to the Buleleng Regency Regional Development Plan for 2023–2026 [26] and relevant previous studies [17].

The sampling method employed in this study is purposive sampling, a technique in which samples are selected based on specific criteria that align with the research objectives [18]. The sample includes village potential data from all villages and sub-districts in Buleleng Regency, amounting to 148 data points. This study utilizes secondary data, specifically the 2021 village potential data for Buleleng Regency. The variables used are as follows: population density (X_1), number of cellular towers (X_2), number of mobile communication operators (X_3), number of accommodation facilities (X_4), number of trade facilities (X_5), number of active cooperatives (X_6), number of bank facilities (X_7), number of worship places (X_8), number of state electricity company (PLN) users (X_9), number of educational facilities (X_{10}), number of health facilities (X_{11}), number of health workers (X_{12}).

B. Research Stage

Figure 1 is a research flowchart of data preprocessing and analysis using the HDBSCAN method. The data analysis in this study follows multiple stages. First, village potential data for Buleleng Regency (2021) is collected from the BPS Buleleng Regency website and the BPS Silastic database. The data is then processed in Python, where descriptive analysis is performed using boxplot visualizations to provide an overview of village potential. Next, data preprocessing is conducted, including multicollinearity testing, data standardization, and outlier detection.

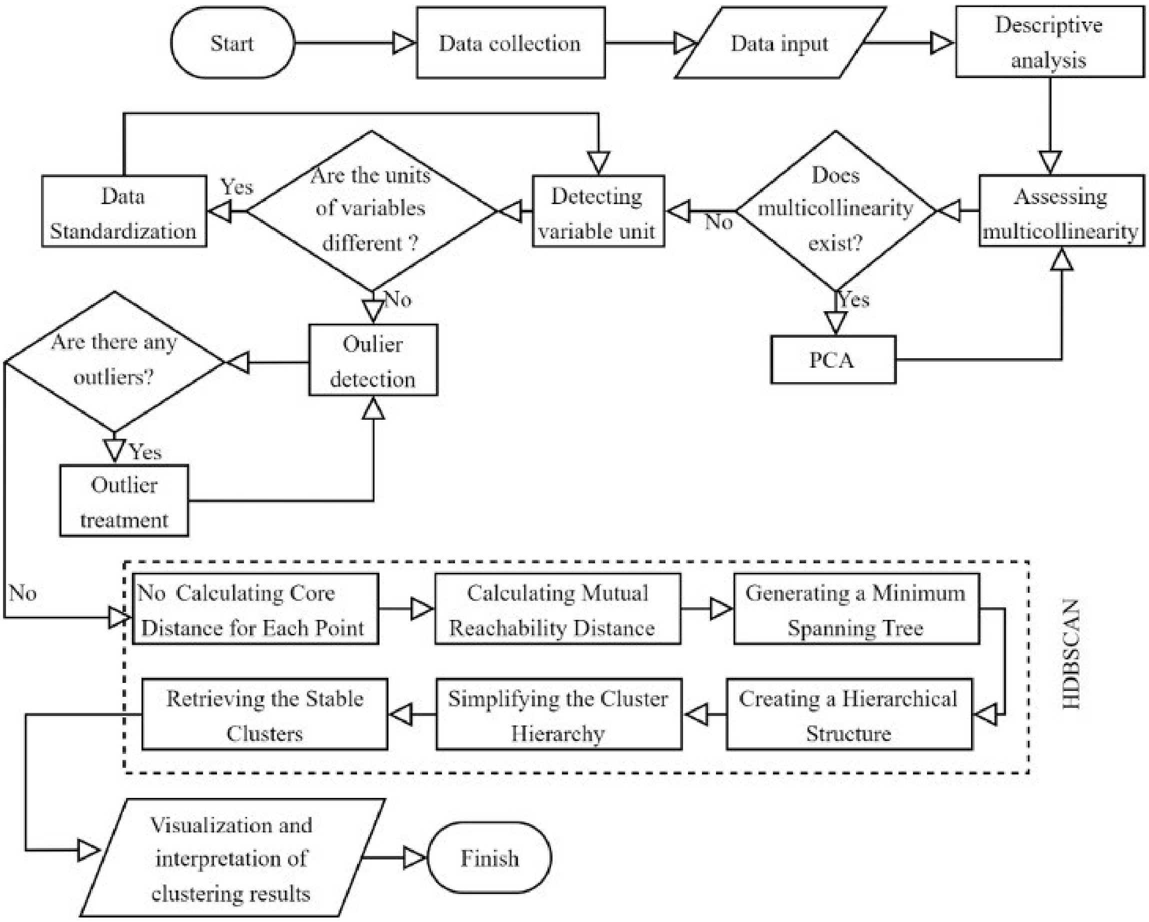

Start
Data collection
Data input
Descriptive analysis
Assessing multicollinearity
Does multicollinearity exist?
Yes
PCA
No
Detecting variable unit
Are the units of variables different ?
Yes
Data Standardization
No
Oulier detection
Are there any outliers?
Yes
Outlier treatment
No
No Calculating Core Distance for Each Point
Calculating Mutual Reachability Distance
Generating a Minimum Spanning Tree
Creating a Hierarchical Structure
Simplifying the Cluster Hierarchy
Retrieving the Stable Clusters
HDBSCAN
Visualization and interpretation of clustering results
Finish

Fig. 1. Research stages

The absence of multicollinearity is a key assumption that must be satisfied when conducting cluster analysis [19]. Multicollinearity occurs when there is a high correlation or linear relationship between independent variables [20]. A multicollinearity test is employed to identify whether any independent variables are similar or redundant with other independent variables [21]. Symptoms of multicollinearity can be detected by examining the correlation values and the Variance Inflation Factor (VIF). The VIF, calculated as the main diagonal element of the inverse correlation matrix, is used to quantify the extent to which a variable is explained as a linear combination of other variables [22].

Meanwhile, an outlier (anomalous data) is an observation that deviates significantly from other data points in a random population sample. When unusual combinations of values occur across multiple variables, the observation is referred to as a multivariate outlier [23]. Detecting outliers is crucial before performing cluster analysis, as cluster analysis is susceptible to irrelevant variables [24].

The Mahalanobis distance is a commonly used distance metric for detecting multivariate outliers [25]. However, a limitation of the Mahalanobis distance is its reliance on the arithmetic mean and covariance matrix, making it highly sensitive to outliers [26]. A robust estimator, such as the Minimum Covariance Determinant (MCD), is required to address this issue. The MCD method estimates the mean and covariance matrix while minimizing the influence of outliers [27].

The HDBSCAN parameters, specifically minimum cluster size and minimum sample values, are then determined. The HDBSCAN clustering process consists of several steps: calculating core distances for each data point, computing mutual reachability distance as the distance metric, constructing a minimum spanning tree (MST), building a cluster hierarchy, compacting the hierarchy based on minimum cluster size, and extracting stable clusters from the dendrogram. The quality of the generated clusters is evaluated based on the Silhouette Coefficient (SC) value. SC is a method used to evaluate the quality and cohesiveness of a cluster by assessing how well an object is assigned to its cluster [28]. The calculation of the SC value refers to the study by Akbar (2023) [29]. The higher the positive SC value of an element, the greater the probability of being assigned to the correct cluster. Conversely, elements with negative SC values are more likely to be misclassified into incorrect clusters [30]. In this study, the SC value is iteratively evaluated to identify the combination that yields the optimal SC score. Once the optimal combination is determined, the best clustering results are obtained using the HDBSCAN method.

C. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

HDBSCAN is a widely used cluster analysis algorithm known for its robustness in handling noise within datasets [31]. It is particularly effective for datasets with varying densities and cluster sizes [8]. HDBSCAN extends the DBSCAN algorithm by incorporating hierarchical clustering and employing a technique to extract flat clusters based on cluster stability [32]. HDBSCAN algorithm requires two key input parameters: (1) minimum samples (min samples) and (2) minimum cluster size (min cluster size). The min samples parameter specifies the minimum number of points needed to designate a core point in cluster formation. A core point has at least the specified number of neighbors determined by the minimum sample value. The min cluster size parameter determines the minimum number of points required to form a cluster. The following outlines the HDBSCAN algorithm's steps, as McInnes et al. des [33].

- *Adjust the space based on the density or sparsity of the data*

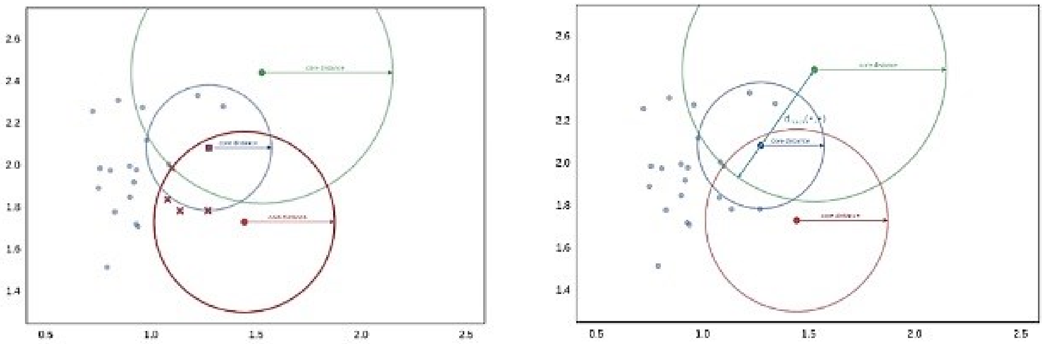
This algorithm identifies clusters by detecting areas of higher data density surrounded by areas of lower data density. The density estimation can be measured by core distance, the distance to the k^{th} the nearest neighbor determined by min samples. The mutual reachability distance metric is used to separate points in low-density regions (high core distances), which is defined as:

$$d_{mreach-k}(a,b) = max\ \{core_k(a), core_k(b), d(a,b)\} \tag{1}$$

Where $d(a,b)$ is the original metric distance between a and b. As an illustration with a value of k = 5, Figure 2 (a) shows the core distance at points A, B, and C. Then, between point A (blue) and point B (green), an arrow can be drawn between the two points, as illustrated in Figure 2 (b). Since the core distance of point B (green) is greater than the core distance of point A (blue) and the distance between the two points. Consequently, the mutual reachability distance between point A and point B is equal to the core distance of point B.

- *Develop a minimum spanning tree based on the weighted distance graph and building a cluster*

After calculating the mutual reachability distance, the next step is to build a weighted graph, where the data points are nodes, and the edges connecting the nodes are weighted based on the mutual reachability distance. From this graph, MST connects all nodes with the most negligible total weight without forming a cycle [34]. Next, the MST is transformed into a connected component hierarchy by sorting and pruning the edges from the highest to the lowest weight. Edges with the same weight are pruned simultaneously. The resulting hierarchy can be visualized as a dendrogram.



(a) (b)

Fig. 2. (a) Illustration of 3 core points, (b) Illustration of mutual reachability distance 1

- *Compacting the cluster hierarchy based on the minimum cluster size and extracting stable cluster*

The first step in cluster extraction is to condense the large and complex cluster hierarchy into a smaller, simplified tree. Next, set the HDBSCAN parameter for the minimum cluster size. Once this parameter is defined, each split in the cluster hierarchy is examined. If a new cluster formed by a split contains fewer points than the minimum cluster size, those points are excluded from the cluster. Conversely, if the split results in two clusters containing at least the minimum number of points, the split is recognized as a valid cluster division. The cluster with the largest ink area can be selected, provided no derived clusters are chosen. A different distance measure is used to evaluate cluster persistence, defined as $\lambda = \frac{1}{distance}$. The stability of each cluster can then be calculated using Equation (2).

$$Stability = \sum_{p\in Cluster} (\lambda_p - \lambda_{birth}) \tag{2}$$

Suppose the sum of the stability values for the child clusters exceeds the stability of the parent cluster. In that case, the cluster stability is set to the sum of the child stabilities. Conversely, suppose the stability of the parent cluster is greater than the sum of its child stabilities. In that case, the parent cluster is selected, and all its child clusters are deselected. The Pseudocode 1 shows the illustration of the HDBSCAN.

III. Result and Discussion

This section presents the results of the data analysis and the interpretation of the research findings obtained through the application of the HDBSCAN method. The analysis aims to identify clustering patterns of villages based on their potential characteristics and evaluate the accuracy of the clustering results. The findings are compared with previous studies to assess their validity and relevance. This section also discusses the implications of the clustering results for village development policies, particularly in designing more targeted and data-driven programs. The following subsections provide a detailed explanation of the results of the analysis, as well as their interpretation and discussion.

A. *Descriptive Statistical Analysis*

Figure 3 visualizes the boxplots of two research variables, highlighting the presence of outliers in all variables, with varying numbers of outliers. These variables tend to be asymmetrical (skewed), as evidenced by the median line being off-center within the box, longer upper whiskers, and outliers above the box plot. These characteristics suggest positive skewness, where most data points have lower values while a few have very high values (outliers).

Figure 3 presents several example variables from the study observations. The figure indicates the presence of certain regions classified as outliers in the population density (X_1) and the number of families using PLN electricity (X_2) variables. Some areas exhibit significantly higher population density than others; a similar pattern is observed in $certain X_2 the\ case\ of\ certain\ regions\ having$

greater electricity access than others. The presence of outliers in these variables is an initial intuition for employing the HDBSCAN clustering method in this study.

PSEUDOCODE 1. HDBSCAN

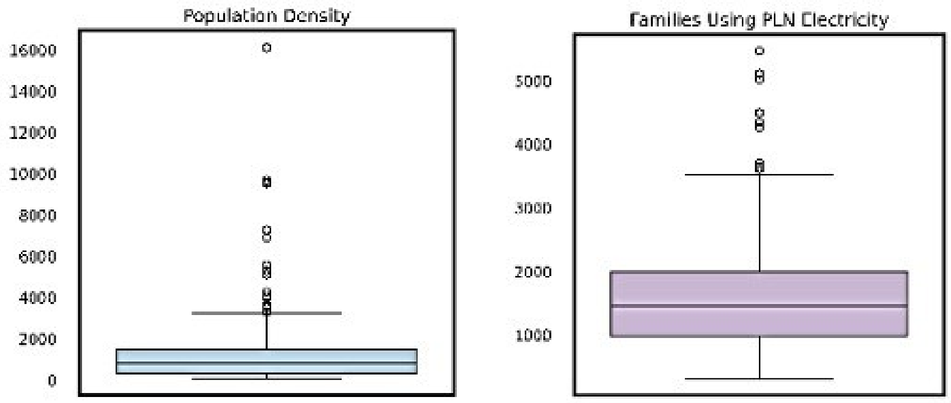
```
Data Collection
   buleleng = pd.read_excel("buleleng.xlsx", header=None)

Preprocessing Data
   scaler = StandardScaler()
   buleleng_std = scaler.fit_transform(buleleng)
   bulelengstd = pd.DataFrame(buleleng_std, columns=[f'Variable_{i}'
   for i in range(1, 13)])
Clustering HDBSCAN
   clusterer=hdbscan.HDBSCAN(min_cluster_size=5,min_samples=2,gen_min
   _span_tree=True)
   clusterer.fit(bulelengstd)
   plt.figure(figsize=(8, 6))
   clusterer.minimum_spanning_tree_.plot(edge_cmap='viridis',edge_alp
   ha=0.6,node_size=80,edge_linewidth=2)
   plt.figure(figsize=(8, 6))
   clusterer.single_linkage_tree_.plot(cmap='viridis', colorbar=True)
   plt.figure(figsize=(8, 6))
   clusterer.condensed_tree_.plot()
   plt.figure(figsize=(8, 6))
   clusterer.condensed_tree_.plot(select_clusters=True,selection_pale
   tte=sns.color_palette())

Clustering Results
   bulelengstd['Cluster'] = clusterer.labels_
   filtered_labels = clusterer.labels_[clusterer.labels_ != -1]
   filtered_probabilities =
   clusterer.probabilities_[clusterer.labels_ != -1]
   bulelengstd_filtered = bulelengstd.query('Cluster != -1')
   if 'Cluster' in bulelengstd_filtered.columns:
   bulelengakhir = bulelengstd_filtered.drop(columns=['Cluster'])

Calculating the Silhouette Coefficient Value
   silhouette_score = silhouette_score(bulelengakhir,
   filtered_labels)

print("Silhouette Score: ", silhouette_score)
```



(a) (b)

Fig. 3. Boxplot of several research variables (source: data processed)

B. Multicollinearity Assessment, Data Standardization, and Outlier Detection

This test aimed to identify whether linear relationships or high correlations existed between the variables. The results of the multicollinearity test indicated that the VIF values for all 12 variables were below 10, indicating no multicollinearity issues [35]. This indicates that the data satisfies the assumption of non-multicollinearity, allowing it to proceed to the subsequent stages of analysis. Then, in clustering analysis, data standardization is necessary when the variables in a study have varying units of measurement [36]. This study's data included variables with different units, such as population density (people per km²) and the number of families using PLN electricity. Therefore, standardization was conducted to ensure comparability across variables. Data standardization aims to rescale the data with a mean of 0 (centered) and a standard deviation 1 [37].

Before clustering with HDBSCAN, multivariate outlier detection was performed using the robust Mahalanobis distance with the Minimum Covariance Determinant (MCD) method. The analysis identified 41.22% of villages (61 data points) in Buleleng Regency as outliers. These outliers can be treated as noise or assigned to a separate cluster if they represent significant characteristics. This study retained outlier data to evaluate their presence and how HDBSCAN classifies them.

C. HDBSCAN Analysis

In this study, to determine the optimal HDBSCAN parameters, various combinations of the minimum cluster size (ranging from 2 to 7) and minimum samples (ranging from 1 to 5) were tested. The clustering results were then evaluated using the SC metric, which measures the clustering quality. In this study, to determine the optimal HDBSCAN parameters, various combinations of minimum cluster sizes (ranging from 2 to 7) and minimum samples (ranging from 1 to 5) were tested. The clustering results were then evaluated using the SC metric, which measures the clustering quality. Table 1 presents the results of parameter combinations applied to the 2021 village potential dataset in Buleleng Regency. Based on Table 1, the parameter combination results indicate that the highest silhouette coefficient achieved in HDBSCAN clustering is 0.37. Consequently, the optimal HDBSCAN parameters in this study are a minimum cluster size of 5 and a minimum sample value of 2, leading to the formation of two clusters.

Table 1. Parameter combination results

No	Minimum Cluster Size	Minimum Samples	Number of Clusters	Number of Noise	SC
1	2	1	23	45	0.19
⋮	⋮	⋮	⋮	⋮	⋮
17	5	2	2	24	0.37
⋮	⋮	⋮	⋮	⋮	⋮
30	7	5	2	101	0.28

These parameters were selected to balance the number of clusters, the proportion of noise, and the silhouette coefficient value. While a silhouette coefficient 0.37 suggests a weak cluster structure, HDBSCAN remains valuable. It successfully identifies two qualitatively distinct clusters, providing meaningful insights for grouping village potentials in line with the research objectives. Although HDBSCAN is robust to noise and complex data distributions, the heterogeneity of village potential in Buleleng Regency contributes to significant variations within clusters. Additionally, the dataset comprises multiple variables with different scales, making achieving a higher silhouette coefficient challenging. Parameter testing further reveals that excessively high parameter values significantly increase the number of villages classified as noise (90–100 out of 148 villages), reducing clustering effectiveness and lowering the silhouette score. Therefore, the selected parameter combination is the most appropriate for generating meaningful and informative clusters.

The first step in clustering analysis using the HDBSCAN method in Python is to determine the optimal parameters: a minimum cluster size of 5 and a minimum sample value of 2. The core distance for each data point is then calculated, where the min samples parameter specifies the k^{th} nearest neighbor to be used. For min samples = 2, the core distance of a point corresponds to the distance to its second nearest neighbor. Next, the mutual reachability distance between points is computed, and a minimum spanning tree is constructed using these distances. Figure 4 (a) illustrates the MST of village potential data in Buleleng Regency for 2021. Based on Figure 4 (a), the graph provides a visualization of the MST, which uses the mutual reachability distance (MRD) to calculate the distances between points. The MRD values in the graph range from 1 to 8. An MRD value close to 1 indicates a shorter distance between points, suggesting that these points are likely within the same cluster.

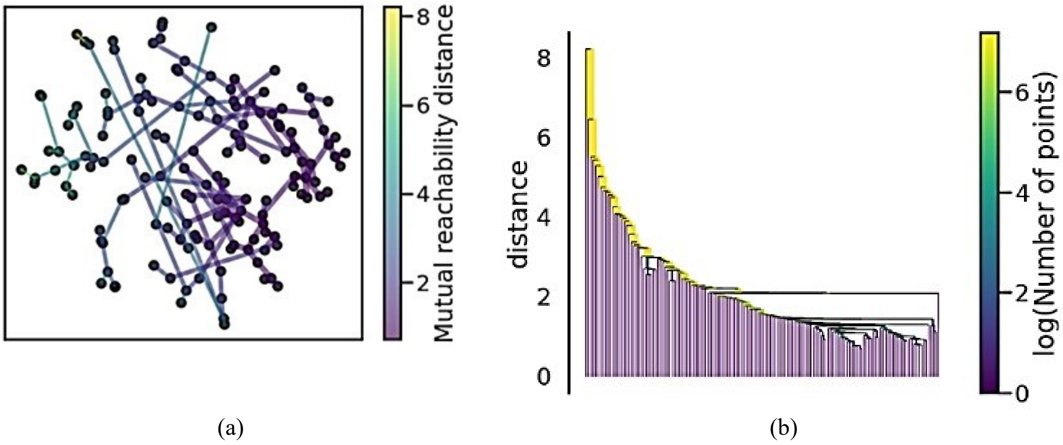


Fig. 4. (a) Minimum spanning tree (MST) graph, (b) Cluster hierarchy graph (source: data processed)

In contrast, an MRD value closer to 8 represents a greater distance between points, implying that these points are more likely to belong to different clusters. Regions with densely interconnected nodes will likely form village clusters with similar economic or social potential. In contrast, villages with longer connections and weaker linkages may indicate more isolated areas or those with distinct characteristics than the main clusters.

Next, construct a cluster hierarchy of connected components by sorting or removing the edges of the minimum spanning tree in descending order of their weights. This process results in a dendrogram, which is visualized as a single linkage tree graph, as shown in Figure 4 (b). Based on Figure 4 (b), the graph visualizes a single linkage tree plot that clusters the village potential data based on distances between points. This visualization includes two y-axes: the left y-axis represents the distance between clusters, ranging from 0 to 8, calculated using the mutual reachability distance. Lower distance values indicate more remarkable similarity or proximity between clusters. The right y-axis displays the logarithm of the number of points in each cluster, ranging from 0 to 7. The dendrogram’s color gradient, from purple to yellow, represents the number of points within each cluster, as indicated by the branches or points in the dendrogram. Points with large distances are likely to be outliers. Denser clusters may indicate villages with similar economic potential, infrastructure, or accessibility.

In contrast, outliers may represent villages with unique characteristics or limited resource access. Previously, the MST was used to establish village relationships based on mutual reachability distance. The Cluster Hierarchy graph serves as the next step in HDBSCAN, facilitating determining the optimal cluster structure. The cluster hierarchy is refined by condensing the dendrogram using a min_cluster_size parameter of 5, resulting in the condensed tree graph shown in Figure 5 (a).

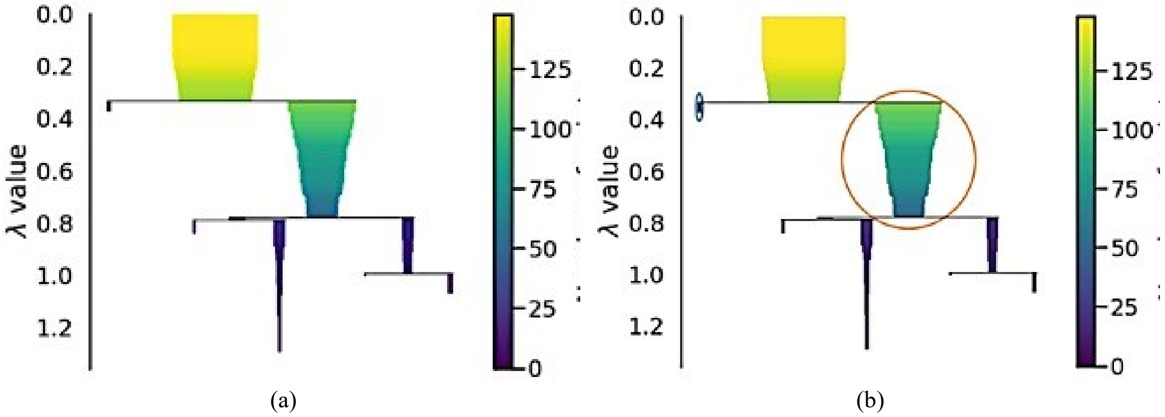


Fig. 5. (a) Cluster hierarchy compaction graph, (b) Cluster extraction graph (source: data processed)

Based on Figure 5 (a), the graph above presents a condensed tree visualization, illustrating clusters’ hierarchical formation and merging within the village potential data. This visualization includes two y-axes: the left y-axis represents the λ values, ranging from 0.0 to 1.2, indicating the clusters’ hierarchical level. The λ value, the inverse of the mutual reachability distance, reflects cluster depth within the dendrogram hierarchy. Lower λ values (e.g., 0.0) correspond to points not included in a cluster or those in the lowest hierarchical cluster. In comparison, higher λ values (e.g., 1.2) represent points belonging to the topmost clusters. The right y-axis displays the logarithm of the number of points within each cluster, ranging from 0 to 150. The next step involves extracting stable clusters from the condensed dendrogram tree.

The condensed tree plot below demonstrates the results of cluster compaction by selecting specific clusters to identify stable or more significant groupings. These stable clusters are extracted from the previously condensed hierarchy, with the selection based on clusters occupying the largest area in Figure 5 (b). Based on Figure 5 (b), HDBSCAN analysis with minimum cluster size = five and minimum samples = 2 parameters on the 2021 village potential data of Buleleng Regency is formed into 2 clusters. The λ value indicates the hierarchical level of the cluster. Cluster 0 is marked with a blue circle containing a purple plot, indicating that the number of points in the cluster is small. In contrast, cluster 1 is marked with an orange circle containing a green plot, indicating the number of points in the cluster is large.

D. *Cluster Result Interpretation*

Figure 6 provides a visualization of the HDBSCAN clustering results for 2021, illustrating the distribution and grouping of villages based on their potential characteristics. This visualization helps us understand the spatial patterns formed by the clustering process. It highlights the distinct groupings identified through the analysis. By examining these results, we can gain deeper insights into the similarities and differences among clusters, which can be helpful for further policy development and strategic planning.

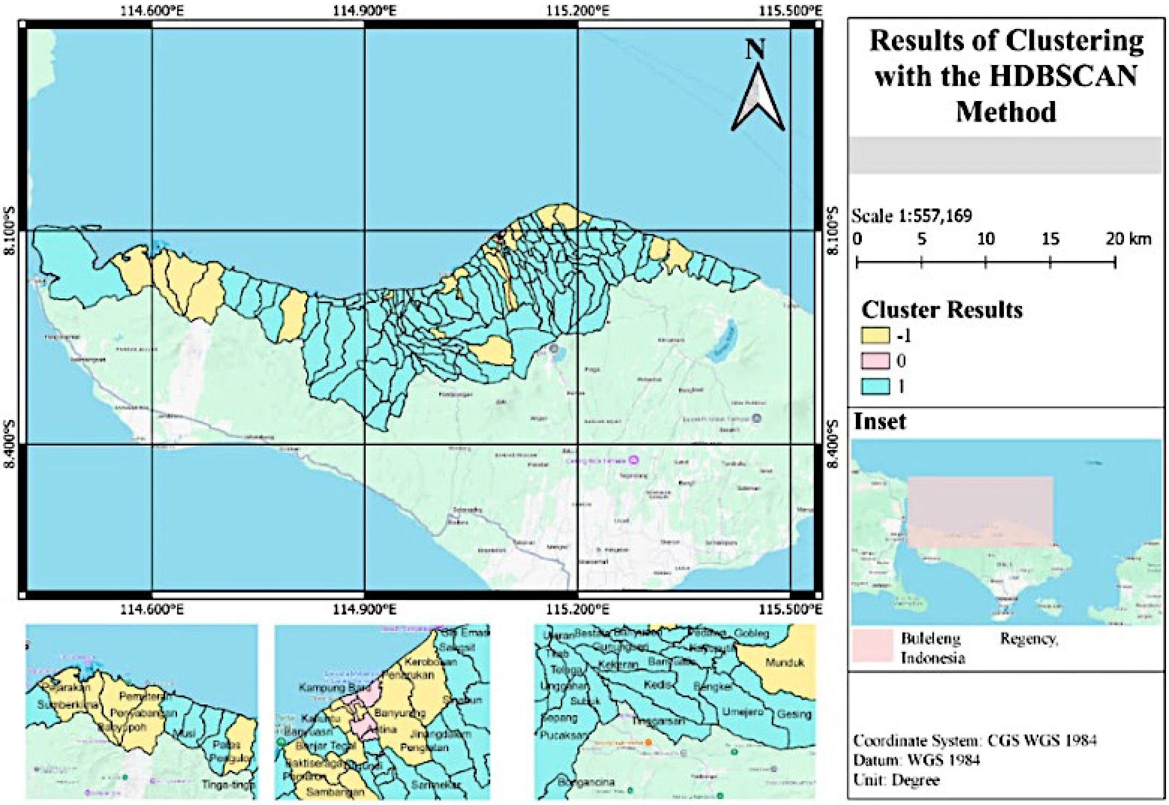


Fig. 6. Cluster results visualization (source: data processed)

Based on Figure 6, the HDBSCAN analysis with parameters of minimum cluster size = five and minimum samples = 2 was applied to the 2021 village potential data in Buleleng Regency, which produces three groups: two clusters and a noise cluster. The first cluster, represented in red, contains six villages/sub-districts. In contrast, the second cluster, shown in blue, comprises 118 villages/sub-districts. The noise cluster, marked in green, includes 24 villages/sub-districts. In HDBSCAN, the cluster labeled as -1 indicates noise or data points that lack sufficient density to form a distinct cluster. Based on Figure 6, the distribution of villages presented in Table 2 is categorized into three clusters:

Cluster -1, Cluster 1, and Cluster 2. The results reveal that outlier data, identified using the MCD method, are distributed across the noise and first clusters. In contrast, the remaining data points are grouped into the second cluster.

Table 2. Distribution of villages based on the HDBSCAN method

Cluster	**Villages**
-1	Pejarakan, Sumberkima, Pemuteran, Banyupoh, Patas, Munduk, Kaliasem, Bondalem, Tejakula, Kalibukbuk, Baktiseraga, Banyuasri, Banjar Tegal, Banjar Bali, Kaliuntu, Kampung Anyar, Banyuning, Penarukaan, Kututambahan, Bungkulan, Seririt, Sambangan, Panji, and Kayuputih (in Sukasada district).
1	Kampung Singaraja, Astina, Banjar Jawa, Kampung Kajanan, Kampung Bugis, Kampung Baru.
2	Sumber Klampok, Penyabangan, Musi, Sanggalangit, Gerokgak, Pengulon, Tinga Tinga, Celukan Bawang, Tukad Sumaga, Banyuatis, Gesing, Gobleg, Kayuputih (in Banjar district), Tirtasari, Banyusri, Pedawa, Tigawasa, Cempaga, Sidetapa, Tampekan, Banjar Tegeha, Banjar, Dencarik, Temukus, Sembiran, Pacung, Julah, Madenan, Les, Penuktukan, Sambirenteng, Tembok, Anturan, Tukadmungga, Pemaron, Paket Agung, Beratan, Liligundi, Kendran, Jinengdalem, Penglatan, Petandakan, Sari Mekar, Nagasepaha, Alasangker, Poh Bergong, Sepang Kelod, Tista, Bongancina, Pucaksari, Sepang, Telaga, Titab, Kekeran, Busungbiu, Pelapuan, Subuk, Tinggarsari, Kedis, Bengkel, Umejero, Tambakan, Pakisan, Bontihing, Tajun, Tunjung, Depeha, Bulian, Tamblang, Bila, Bukti, Mengening, Lemukih, Galungan, Sekumpul, Bebetin, Sudaji, Sawan, Menyali, Suwug, Jagaraga, Sinabun, Kerobokan, Sangsit, Giri Emas, Unggahan, Gunungsari, Munduk Bestala, Bestala, Mayong, Rangdu, Ularan, Ringdikit, Joanyar, Kalianget, Tangguwisia, Sulanyah, Bubunan, Pengastulan, Patemon, Lokapaksa, Umeanyar, Banjar Asem, Kalisada, Pangkungparuk, Pancasari, Wanagiri, Ambegan, Gitgit, Pegayaman, Silangjana, Pegadungan, Padangbulia, Sukasada, Panji Anom, Tegal Linggah, and Selat.

This demonstrates that HDBSCAN effectively separates less dense data points (noise) and groups the remaining data based on their density and characteristics. Next, the HDBSCAN cluster results will be profiled to identify the characteristics of each cluster. Table 3 presents the descriptive statistics for each cluster. Based on the profiling results, Cluster 1 exhibits higher village potential than Cluster 2, as indicated by its average population density of 8,021 people/km². Health facilities (6 per village) and trade facilities (10 per village) contribute more significantly to village potential in Cluster 1 than in Cluster 2. Cluster 1 outperforms Cluster 2 in key variables, including mobile communication service operators, trading facilities, active cooperatives, banking institutions, PLN electricity user families, educational and health facilities, and healthcare workers.

Table 3. Descriptive statistics for each cluster

Variables	**Clusters**											
	-1				**1**				**2**			
	Mean	***Max***	***Min***	***Std.dev***	***Mean***	***Max***	***Min***	***Std.dev***	***Mean***	***Max***	***Min***	***Std.dev***
X_1	2570.4	16122.2	247.2	3264.8	8020.5	9688.9	5118.5	1717.6	920.5	4113.3	89.2	826.3
X_2	3.5	8.0	0.0	2.2	1.2	2.0	0.0	0.9	1.4	7.0	0.0	1.5
X_3	5.0	7.0	4.0	0.8	5.7	6.0	5.0	0.5	4.6	6.0	2.0	0.8
X_4	19.3	103.0	0.0	25.7	0.5	1.0	0.0	0.5	2.9	43.0	0.0	6.2
X_5	14.9	67.0	3.0	13.6	9.8	18.0	1.0	6.2	3.1	19.0	0.0	3.7
X_6	3.4	11.0	0.0	2.8	1.0	2.0	0.0	0.8	1.0	5.0	0.0	1.1
X_7	3.0	10.0	0.0	2.6	1.3	3.0	0.0	1.1	1.3	5.0	0.0	1.0
X_8	11.6	39.0	3.0	7.7	4.8	8.0	2.0	2.0	9.5	34.0	1.0	5.9
X_9	2858.7	5467.0	805.0	1319.0	1401.5	2584.0	403.0	681.9	1398.8	3702.0	317.0	679.7
X_{10}	13.7	27.0	5.0	5.8	6.2	11.0	1.0	3.2	6.0	19.0	1.0	3.3
X_{11}	7.4	19.0	0.0	5.0	5.5	10.0	0.0	4.0	3.2	11.0	0.0	2.0
X_{12}	16.4	53.0	1.0	14.8	7.2	11.0	1.0	3.2	4.8	17.0	0.0	3.9

However, Cluster 2 exhibits higher values in mobile phone towers, accommodation facilities, and places of worship. Despite these differences, Cluster 1 generally represents villages with more significant overall potential than Cluster 2. Cluster -1 (noise) notably demonstrates the highest village potential, outperforming both clusters across all variables except for population density and mobile communication service operators. Table 4 summarizes the clustering results and policy recommendations derived from the cluster analysis. Based on Table 4, villages in Cluster 1 have higher potential due to their strategic location near Singaraja City, ensuring better access to public facilities and government services. These areas should focus on strengthening the economic sector, improving infrastructure, and optimizing public services. In contrast, Cluster 2 consists of remote or mountainous villages with limited infrastructure and economic facilities. The government should prioritize developing basic infrastructure, including communication services, banking, trade, education, healthcare, and cooperatives.

Table 4. Summarizes the clustering results and policy recommendations derived from the cluster analysis

Cluster	Main Characteristics	Challenges	Policy Recommendations
1 (6 villages)	High village potential, better access to public facilities and government services	Economic competition, need for infrastructure strengthening	Strengthening the economic sector, strategic infrastructure, and optimizing public services
2 (118 villages)	Lower village potential, mostly in remote/mountainous areas	Limited infrastructure and economic facilities, as well as minimal access to essential services	Prioritizing basic infrastructure development (communication, banking, trade, education, healthcare)
-1 (24 villages)	Very high village potential, dominated by tourism, agriculture, and fisheries	Disparities with other villages, management of key sectors	Developing community-based tourism, supporting agribusiness, and improving supporting infrastructure

Additionally, accommodation facilities and places of worship require attention to address the specific needs of these areas. Villages in Cluster -1 (noise) exhibit higher village potential than those in Clusters 1 and 2, often featuring key tourist destinations such as Pemuteran. Additionally, agriculture and fisheries serve as the primary economic sectors in these areas. The government can promote community-based tourism, support agribusiness, and improve market access and infrastructure to enhance their development. This clustering provides valuable insights for equitable village development in Buleleng Regency, helping the government prioritize needy areas, particularly those in Cluster 2. The government can enhance community welfare more effectively by focusing on targeted programs and strategic resource allocation.

IV. Conclusion

An analysis of the village potential data in Buleleng Regency reveals significant variation in variables such as population density, accommodation facilities, and the number of families using PLN electricity. Other variables exhibit moderate and low variability, reflecting disparities in infrastructure facilities and services across regional villages. Clustering analysis using the HDBSCAN method with optimal parameters (minimum cluster size=5 and minimum samples=2) identified two primary and one noise cluster. The first cluster comprises six villages/sub-districts, the second comprises 118 villages/sub-districts, and the noise cluster (labeled cluster -1) contains 24 villages/sub-districts. The clustering results yielded a Silhouette Coefficient value of 0.37, indicating a weak cluster structure. Despite this, the HDBSCAN method remains valuable, as it produces distinct clusters that provide meaningful and relevant insights for grouping village potential according to the study’s objectives. Profiling results reveal that the first cluster represents villages/sub-districts with higher potential than those in the second cluster. These villages benefit from their strategic location near the center of Singaraja City, offering better access to public facilities and government services. Development efforts should focus on enhancing the economy, infrastructure, and services for these areas. In contrast, villages/sub-districts in the second cluster demonstrate lower potential due to their location in mountainous or remote areas, which limits infrastructure and public service access. These areas require targeted improvements in infrastructure and public services to enhance their development and overall quality of life.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We extend our deepest gratitude to the computer science faculty of Universitas Muslim Indonesia. Their guidance, expertise, and unwavering support have been key in realizing this research.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

Publisher's Note: Department of Electrical Engineering and Informatics - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.

References

[1] S. Berdej and D. Armitage, "Bridging for Better Conservation Fit in Indonesia's Coastal-Marine Systems," Front. Mar. Sci., vol. 3, Jun. 2016.

[2] T. L. Wanadjaja and P. L. Samputra, "Examining tri hita karana as the critic to the triple bottom line of sustainable development," IOP Conf. Ser. Earth Environ. Sci., vol. 716, no. 1, p. 012121, Mar. 2021.

[3] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," Artif. Intell. Rev., vol. 56, no. 7, pp. 6439–6475, Jul. 2023.

[4] R. Mondal, E. Ignatova, D. Walke, D. Broneske, G. Saake, and R. Heyer, "Clustering graph data: the roadmap to spectral techniques," Discov. Artif. Intell., vol. 4, no. 1, p. 7, Jan. 2024.

[5] J. C. Munguía Mondragón, E. Rendón Lara, R. Alejo Eleuterio, E. E. Granda Gutirrez, and F. Del Razo López, "Density-Based Clustering to Deal with Highly Imbalanced Data in Multi-Class Problems," Mathematics, vol. 11, no. 18, p. 4008, Sep. 2023.

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

[7] N. M. Nhat, "Applied Density-Based Clustering Techniques for Classifying High-Risk Customers: A Case Study of Commercial Banks in Vietnam," J. Appl. Data Sci., vol. 5, no. 4, pp. 1639–1653, Dec. 2024.

[8] Y. Wang, S. Yu, Y. Gu, and J. Shun, "Fast Parallel Algorithms for Euclidean Minimum Spanning Tree and Hierarchical Spatial Clustering," in Proceedings of the 2021 International Conference on Management of Data, Jun. 2021, vol. 12, no. 5, pp. 1982–1995.

[9] L. Wang, P. Chen, L. Chen, and J. Mou, "Ship AIS Trajectory Clustering: An HDBSCAN-Based Approach," J. Mar. Sci. Eng., vol. 9, no. 6, p. 566, May 2021.

[10] M. Strobl, J. Sander, R. J. G. B. Campello, and O. Zaïane, "Model-Based Clustering with HDBSCAN*," in Machine Learning and Knowledge Discovery in Databases, 2021, pp. 364–379.

[11] L. Zhang, X. Su, Y. Wang, M. Wang, X. Yang, and Z. Xu, "HDBSCAN-based semantic clustering model in classifying incidents on security and environmental conservation management," in Ninth International Symposium on Advances in Electrical, Electronics, and Computer Engineering (ISAEECE 2024), Oct. 2024, p. 116.

[12] 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.

[13] Y. Kim and B. Yang, "Extracting Urban Areas of Interest Using HDBSCAN Clustering Method," Journal of the Korean Cartographic Association, vol. 1, pp. 67–77, 2023.

[14] M. Jahangoshai Rezaee, M. Eshkevari, M. Saberi, and O. Hussain, "GBK-means clustering algorithm: An improvement to the K-means algorithm based on the bargaining game," Knowledge-Based Syst., vol. 213, p. 106672, Feb. 2021.

[15] A. Bigdeli, A. Maghsoudi, and R. Ghezelbash, "Application of self-organizing map (SOM) and K-means clustering algorithms for portraying geochemical anomaly patterns in Moalleman district, NE Iran," J Geochem Explor, vol. 233, p. 106923, Feb. 2022.

[16] A. Islam, Md. A. Sayeed, Md. K. Rahman, J. Ferdous, S. Islam, and M. M. Hassan, "Geospatial dynamics of COVID-19 clusters and hotspots in Bangladesh," Transbound Emerg Dis, vol. 68, no. 6, pp. 3643–3657, Nov. 2021.

[17] I Made Gunamantha, I Gede Astra Wesnawa, Ni Made Oviantari, Ni Wayan Yuningrat, Putu Lilik Pratami Kristiyanti, and Komang Widiadnyana, "Estimating Circular Economic Potential of Organic Fraction of Municipal Solid Waste in Small City," J. Environ. Sci. Econ., vol. 2, no. 4, pp. 80–96, Dec. 2023.

[18] N. A. Wahyuni, M. N. Hayati, and N. A. Rizki, "Metode hierarchical density-based spatial clustering of application with Noise (HDBSCAN) pada wilayah desa/kelurahan tertinggal di kabupaten Kutai Kartanegara," EKSPONENSIAL, vol. 12, no. 1, p. 47, Jun. 2021.

[19] C. Andrade, "The Inconvenient Truth About Convenience and Purposive Samples," Indian J Psychol Med, vol. 43, no. 1, pp. 86–88, Jan. 2021.

[20] D. T. Utari and D. S. Hanun, "Hierarchical Clustering Approach for Region Analysis of Contraceptive Users," EKSAKTA: Journal of Sciences and Data Analysis, pp. 99–108, Sep. 2021.

[21] T. Kyriazos and M. Poga, "Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions," Open J Stat, vol. 13, no. 03, pp. 404–424, 2023.

[22] S. Streukens and S. Leroi-Werelds, “Multicollinearity: An Overview and Introduction of Ridge PLS-SEM Estimation,” in Partial Least Squares Path Modeling, Cham: Springer International Publishing, 2023, pp. 183–207.

[23] P. Pradhan and A. D. Setyawan, “Filtering multi-collinear predictor variables from multi-resolution rasters of WorldClim 2.1 for Ecological Niche Modeling in Indonesian context,” Asian Journal of Forestry, vol. 5, no. 2, Sep. 2021.

[24] M. M. El-Masri, F. I. Mowbray, S. M. Fox-Wasylyshyn, and D. Kanters, “Multivariate Outliers: A Conceptual and Practical Overview for the Nurse and Health Researcher,” Canadian Journal of Nursing Research, vol. 53, no. 3, pp. 316–321, Sep. 2021.

[25] B. Ray, S. Ghosh, S. Ahmed, R. Sarkar, and M. Nasipuri, “Outlier detection using an ensemble of clustering algorithms,” Multimed Tools Appl, vol. 81, no. 2, pp. 2681–2709, Jan. 2022.

[26] K. Dashdondov and M.-H. Kim, “Mahalanobis Distance Based Multivariate Outlier Detection to Improve Performance of Hypertension Prediction,” Neural Process Lett, vol. 55, no. 1, pp. 265–277, Feb. 2023.

[27] S. Zahariah and H. Midi, “Minimum regularized covariance determinant and principal component analysis-based method for the identification of high leverage points in high dimensional sparse data,” J Appl Stat, vol. 50, no. 13, pp. 2817–2835, Oct. 2023.

[28] H. Bulut and T. Zaman, “An improved class of robust ratio estimators by using the minimum covariance determinant estimation,” Commun Stat Simul Comput, vol. 51, no. 5, pp. 2457–2463, May 2022.

[29] T. Akbar, G. M. Tinungki, and S. Siswanto, “Performance Comparison of K-Medoids and Density Based Spatial Clustering of Application with Noise using Silhouette Coefficient Test,” BAREKENG: Jurnal Ilmu Matematika dan Terapan, vol. 17, no. 3, pp. 1605–1616, Sep. 2023.

[30] Y. J. Kim, W. Nam, and J. Lee, “Multiclass anomaly detection for unsupervised and semi-supervised data based on a combination of negative selection and clonal selection algorithms,” Appl. Soft Comput., vol. 122, p. 108838, Jun. 2022.

[31] M. Shutaywi and N. N. Kachouie, “Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering,” Entropy, vol. 23, no. 6, p. 759, Jun. 2021.

[32] N. M. Nhat, “Applied Density-Based Clustering Techniques for Classifying High-Risk Customers: A Case Study of Commercial Banks in Vietnam,” Journal of Applied Data Sciences, vol. 5, no. 4, pp. 1639–1653, Dec. 2024.

[33] T.-H. Tran, T.-D. Cao, and T.-T.-H. Tran, “HDBSCAN: Evaluating the Performance of Hierarchical Clustering for Big Data,” 2021, pp. 273–283.

[34] A. Mashreghi and V. King, “Broadcast and minimum spanning tree with o(m) messages in the asynchronous CONGEST model,” Distrib. Comput., vol. 34, no. 4, pp. 283–299, Aug. 2021.

[35] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, Multivariate Data Analysis, Eigth edition. Annabel Ainscow, 2019.

[36] W. G. Hopkins and D. S. Rowlands, “Standardization and other approaches to meta-analyze differences in means,” Stat Med, vol. 43, no. 16, pp. 3092–3108, Jul. 2024.

[37] M. Oka, “Interpreting a standardized and normalized measure of neighborhood socioeconomic status for a better understanding of health differences,” Archives of Public Health, vol. 79, no. 1, p. 226, Dec. 2021.