Available at https://jtsiskom.undip.ac.id (26 April 2021)
DOI:10.14710/jtsiskom.2021.13998

Jurnal Teknologi dan Sistem Komputer, 9(3), 2021, 133-141

Identification of the distribution village maturation: Village classification using Density-based spatial clustering of applications with noise

Okfalisa[1*)], Angraini[2,4)], Shella Novi[1)], Hidayati Rusnedy[1)], Lestari Handayani[1,3)], Mustakim[4)]

[1)]*Informatics Engineering Department, Universitas Islam Negeri Sultan Syarif Kasim Riau*
Jl. HR. Soebrantas Panam Km. 15 No. 155, Tuah Madani, Kec. Tampan, Kampar Regency, Riau, Indonesia 28293

[2)]*School of Computing, Faculty Engineering, Universiti Teknologi Malaysia*
UTM Johor Bahru, Johor, Malaysia 81310

[3)]*Prism Lab, Insa Center Val de Loire*
88 Boulevard Lahitolle, Bourges, France 18000

[4)]*Information System Department, Universitas Islam Negeri Sultan Syarif Kasim Riau*
Jl. HR. Soebrantas Panam Km. 15 No. 155, Tuah Madani, Kec. Tampan, Kampar Regency, Riau, Indonesia 28293

How to cite: O. Okfalisa, A. Angraini, S. Novi, H. Rusnedy, L. Handayani, and M. Mustakim. "Identification of the distribution village maturation: Village classification using Density-based spatial clustering of applications with noise," *Jurnal Teknologi dan Sistem Komputer*, vol. 9. no. 3, pp. 133-141, 2021. doi: 10.14710/jtsiskom.2021.13998, [Online].

Abstract* - *The rural development measurement is undoubtedly not easy due to its particular needs and conditions. This study classifies village performance from social, economic, and ecological indices. One thousand five hundred ninety-one villages from the Community and Village Empowerment Office at Riau Province, Indonesia, are grouped into five village maturation classes: very under-developed village, under-developed village, developing village, developed village, and independent village. To date, Density-based spatial clustering of applications with noise (DBSCAN) is utilized in mining 13 of the villages' attributes. Python programming is applied to analyze and evaluate the DBSCAN activities. The study reveals the grouping's silhouette coefficient values at 0.8231, thus indicating the well-being clustering performance. The epsilon and minimum points values are considered in DBSCAN evaluation with percentage splits simulation. This grouping can be used as guidelines for governments in analyzing the distribution of rural development subsidies more optimal.

Keywords *- clustering; density-based spatial clustering of applications with noise; Python; silhouette coefficient; village maturity*

I. INTRODUCTION

A village is a law unit with the territory and approved a regulation in governance, the community's interests, and legal rights recognized and respected by the government system in the Negara Kesatuan Republik Indonesia (NKRI) [1]. As an archipelagic country, Indonesia faces complex interaction and access to the physical environment. Thus, this geospatial condition causes the village's situation to be separated from rural areas and identic with poverty and underdevelopment. Moreover, the limited livelihoods of the village community triggered the condition that further away from prosperity status.

To support village development, the government issued various regulations and laws, including statute No. 6, 2014 concerning the village's status, and government regulation No. 22, 2015 for village fund management arrangements. Moreover, Rencana Pembangunan Jangka Menengah Nasional (RPJMN) declaration in 2015-2019 released strengthening efforts to achieve villages and rural areas' development goals by introducing the villages index mechanism. The program seeks to cut the number of under-developed villages to 5,000 villages and boost the percentage of autonomous villages to at least 2,000 population by 2019. The villages index categorized five villages: under-developed village, under-developed village, developing village, developed village, and independent village. The aggregation classes were determined based on the progress and independence status, with the scoring range between 0.4907 to 0.8155. This village index mapped how were the conditions and characteristics of the village. Thus, the government can use this index to design village development plans to be more efficient and targeted.

Meanwhile, Badan Perencanaan Pembangunan Nasional (National Development Planning Agency Republic of Indonesia) or Bappenas, as the government agency that regulates the national development planning, issued another village index category. The villages were classified into three categories, viz. under-developed village, developing village, and independent village. However, it found several differences between the two above government measurements, especially that related to the indicator's performance and the percentage calculation of the village mapping [2]. Moreover, the two leveling indexes above are carried out using only the simple arithmetic nominal dimension

*) Corresponding author (Okfalisa)
Email: okfalisa@uin-suska.ac.id

Submitted: 3 December 2020; Revised: 19 March 2021; Accepted: 24 April 2021; Published: 31 July 2021

scale [3]. This condition prompted the measurement analysis's obligation to map the village grouping with more optimal, accurate, and comprehensive without ignoring the significant standard factors proposed by government appraisal indicators.

Riau province has a strategic geographical position, flanked by three neighboring countries: Malaysia, Singapore, and Thailand. This position certainly has a positive impact, especially concerning the development of the industrial market economy in Riau, thus accelerate the growth and progress of villages in this province. It is reported that there were 1591 villages data in Riau province that categorized 736 villages into developing village class, 69 villages in a developed village class, four villages in independent villages class, 121 villages in very under-developed, and 661 villages are grouping into an under-developed category. In a nutshell, these data indicated that the village's construction conditions in Riau province are far from evenly distributed, and many towns are found in the cluster behind.

Clustering is among the most critical data mining methods to treat and group unsupervised data in similar ways [4]. Clustering provides a valid analytical for solving complex problems by finding specific interesting data patterns to support the knowledge discovery process [5]. The contribution of clustering covers the limitation of statistical analysis, especially for a considerable data analysis. The various domain studies have been shown the effectiveness of this approach in clustering the medical imaging data and image segmentation [6], [7], the digital marketing analysis and performance metrics [8], [9], the education and performance prediction [10], the chemical process analysis [11], [12], and the manufacturing process and analytical [13], [14]. In a nutshell, the previous studies reflected on the potential of data mining tools and clustering techniques to increase the visibility and responsiveness of distributed knowledge discovery data.

Previous studies reviewed the three commonly used techniques in data clustering, including partition-based [15], hierarchy-based [16], and density-based [17]. Density-based clustering algorithms are widely used in several areas [18] that highlighting the arbitrary-shaped clusters and data noise. Density-based clustering distinguishes the different groups or clusters in a dataset relying on the idea that clusters are densely contiguous areas within the total data space, separated from other clusters by adjacent regions with relatively lower data density [19]. Data points with a softer object density ratio in the scattering area are typically classified as noise or outlier [19], [20].

Meanwhile, Density-Based Spatial Clustering of Application with Noise (DBSCAN) is a density-based clustering method that creates population densities linked to high and low-density deliberation [21]. DBSCAN generates the numbers of data within the radius of Eps (ε) and the minimum number of contiguous data points (*minpts*) to be grouped into clusters. Thus, DBSCAN is perceived as the most rugged and cited cluster algorithm for density that recognizes significant random shapes and sizes clusters in massive, nuisance-damaged databases [22]. Since DBSCAN accomplishes the disturbance points correctly and effectively, this method defines a group surrounded by noise and separates it into different categories [23], [24]. However, the current DBSCAN algorithm still has many shortcomings, such as unwillingness to locate multi-density clusters [25], [26], the issues with specifying the appropriate density thresholds [27], a scarcity of computational parallel design, the time spent in finding the nearest neighbors inside the cluster expansion [28], and the inability to group gradually [29].

Several new changes have been made to DBSCAN to overcome the original DBSCAN and effectively deal with ambient queries. Andrade et al. [30] used a graphics processing unit to parallel G-DBSCAN. Yinghua et al. [28] established a DBSCAN-Influence Space for a complex data set. Several studies optimized and rapidly generated DBSCAN with R, a novel DBSCAN hybridization and fuzzy earthworm optimization algorithm for data cube clustering [28], [30], [31].

After investigating the reviews of the DBSCAN's advantages and the opportunities of this method advancement in analytical data mining, this study aims to employ the DBSCAN method to cluster the development of villages status to provide a more comprehensive and accurate analytical solution in measuring development villages index. Here, Python programming is applied for interpreting the calculation and clustering theorem. Thus, the mapping and identification of villages' characteristics grow into more precise and optimal.

The remaining portion of this paper is structured accordingly: Section 2 outlined the procedures used in this paper, such as data mining and the DBSCAN formula. Section 3 considered the outcome and assessment of the DBSCAN adoption in the clustering of villages. The final declaration and contribution of this paper concluded with the new part in Section 4.

II. RESEARCH METHODS

Systematically, this research was conducted through several activities, including problem identification by exploring literature reviews associated with the topic, the observation, and interviews at the community and village empowerment office, Riau province. Five stakeholders from the agency were asked about their functions and work operations, activities, strategic planning, and supporting regulations for developing and empowering rural communities in Riau province.

Certain supporting documents were studied, especially the villages mapping data based on the development villages index's value. As primary data, 1591 villages from the year 2018 with 13 attributes were analyzed by focusing on the three main attributes, namely the social resilience index (IKS), the environmental resilience index (IKL), and the economic

resilience index (IKE). The above three main attributes are chosen by referring to the development villages index set up within the government regulation No. 2, 2016 that concerning the dimensions of the development village index [32].

The IKS is measured by considering the dimension in health (service, health, community empowerment for health), education (educational access to middle and high school, the road to non-formal education, and admittance to knowledge), and social capital (solidarity sensitivity, tolerance awareness, and sense of citizens protection). Each dimension is determined into several key performance indicators (KPIs).

The IKL is defined by the ecological dimension, including the environmental quality and disaster response. The availability of water, soil, air pollution, and the numbers of river waste affected assume KPIs' form in environmental quality. Meanwhile, natural disasters such as floods, landslides, forest fires, and handling such disasters were resolved as KPIs for disaster response.

The IKE deliberated economic (production diversity, the availability of service center, trading, access to financial credit institutions, and economic institutions), social welfare (the availability of special schools, numbers of people with social welfare, the numbers of suicide people), and settlement (access to clean water, sanitary, electricity, communication, and regional openness) dimensions whereby measured by its KPIs.

The government regulation also groups the villages into five village statuses with the scale distribution defined in Table 1. The calculation of the total value of the development villages index (IDM) is carried out with a simple formula in (1) by adding up the total values of IKS, IKL, and IKE.

$$IDM = \frac{1}{3}(IKL + IKE + IKS) \quad (1)$$

Meanwhile, the adoption of knowledge discovery data (KDD) mining in this study generates a new contribution to the IDM measurement and classification. Subsequently, this study follows the KDD concepts, consisting of data selection, preprocessing/cleaning, transformation, data mining, and interpretation [33], [34]. KDD is a noticeable method to find new relevant patterns from large quantities of potentially useful and meaningful [35]. Data mining is an unavoidable step of the KDD process in harvesting useful knowledge from the dataset. For mining, this study applied the DBSCAN method in clustering the village data. The tracking algorithm of DBSCAN is stated below [36].

1. Determine the point ***p*** as an object randomly.
2. Calculate the Euclidean distance with (2), where *x* and *y* as objects and *n* as numbers of objects. This calculation respects the similarity measurement between objects in cluster analysis.

$$E(x, y) = \sqrt{\sum_{i=0}^{n-1} (x_i - y_i)^2} \quad (2)$$

Table 1. The Category of villages status

No.	Villages Status	Development Villages Index
1.	Independent	> 0.8155
2.	Developed	0.7072 – 0.8155
3.	Developing	0.5989 – 0.7072
4.	Under-developed	0.4907 – 0.5989
5.	Very under-developed	< 0.4907

Algorithm 1. DBSCAN a density-based clustering algorithm

Input:
D: a data set containing n objects,
Eps: the radius parameter, and
MinPts: the neighborhood density threshold.
Output: A set of density-based clusters.
1: mark all objects as unvisited;
2: **do**
3: randomly select an unvisited object *p*;
4: mark *p* as visited;
5: **if** the *Eps*-neighbourhood of *p* has at least *MinPts* objects
6: create a new cluster *C* and add *p* to *C*;
7: let *N* be the set of objects in the *Eps*-neighbourhood of *p*;
8: **for** each point *p*0 in *N*
9: **if** *p*0 is unvisited
10: mark *p*0 as visited;
11: **if** the *Eps*-neighbourhood of *p*0 has at least *MinPts* points
12: add those points to *N*;
13: **if** *p*0 is not yet a member of any cluster
14: add *p*0 to *C*;
end for
15: output *C*;
16: **else** mark as noise;
until no object is unvisited;

3. Determine the value *Eps* and *MinPts* by considering the values of noise in (3), directly-density-reachable in (4), and density-connected in (5). The *x* denotes the data cluster, C_i the first cluster, N_{Eps} (*y*) the point around *y* in the radius, *Eps MinPts* as the minimum point in the cluster, N_{Eps} (*x*) as the surrounding point of *x* in the radius *Eps*, *D* as the data set, *dist (x, y)* as the Euclidean distance, and *Eps* as the radius parameter. The algorithm tracks the following as Algorithm 1.

$$Noise = \{x \in X \cup \forall i : x \in C_i\} \quad (3)$$

$$x \in N_{Eps}(y) \cap |N_{Eps}| \geq MinPts \quad (4)$$

$$N_{Eps}(x) = \{y \in D \cup dist(x, y) \leq Eps\} \quad (5)$$

In order to test the accuracy of village grouping, the silhouette validity index was applied with a ratio percentage split of training data and testing data at 90:10, 80:20, and 70:30, respectively [37]. Silhouette metric simultaneously tests cluster segregation and cohesiveness [38]. The visual object outcomes generally apply the silhouette approach to discover the cluster's

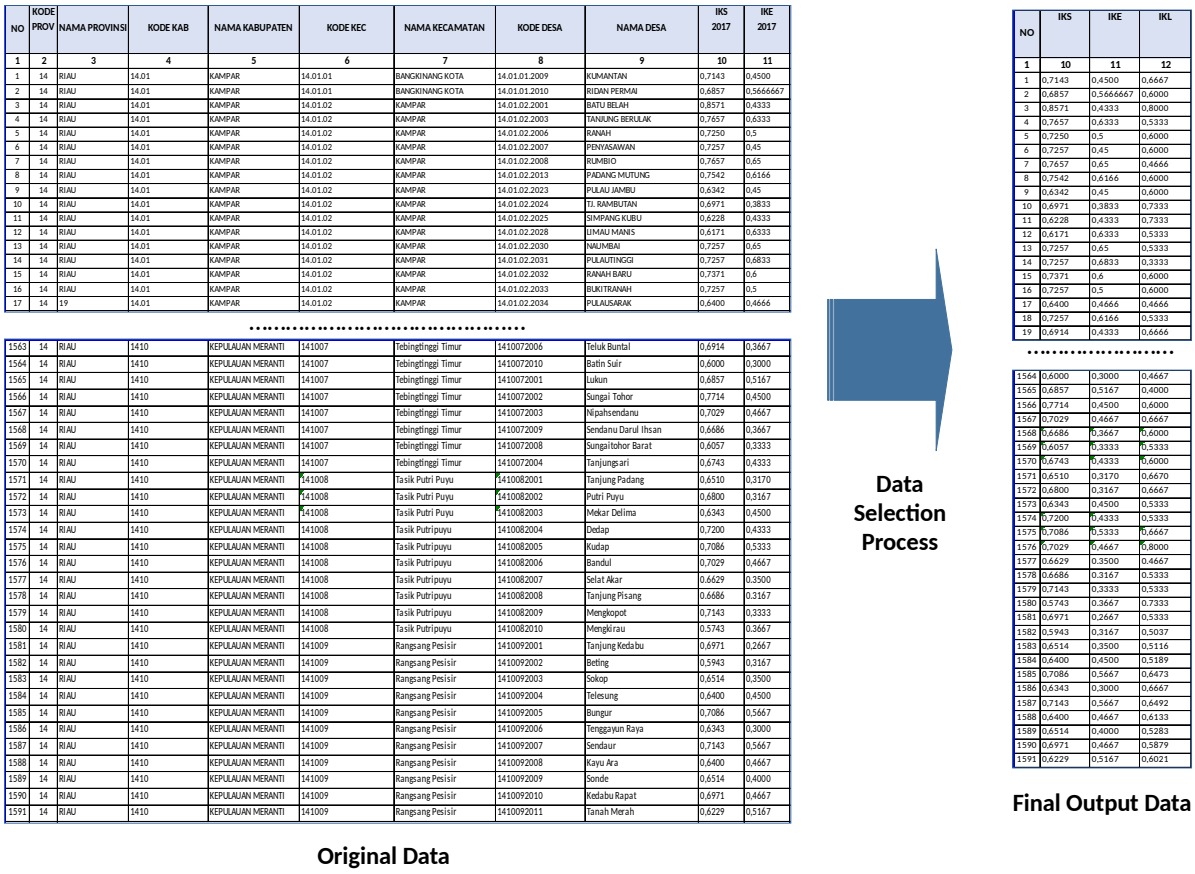


Figure 1. Illustration of the village data selection process

intensity and consistency [39]. The silhouette coefficient enumerates the average distance between data points in a similar cluster compared to other clusters [40]. The measurement of the silhouette validity index pursues (6)-(8).

$$a(o)=\frac{\sum_{o\in C_i, o\neq o'} dist(o,o')}{|C_i|-1} \quad (6)$$

$$b(o)=\min_{C_j:1\leq k, j\neq i}\left\{\frac{\sum_{o'\in C_j} dist(o,o')}{|C_j|}\right\} \quad (7)$$

$$s(o)=\frac{b(o)-a(o)}{max\{a(o),b(o)\}} \quad (8)$$

D is partitioned into k clusters ($C_1 \ldots C_k$); for each object $\boldsymbol{o} \in$ D, an ($\boldsymbol{o}$) is gauged as the average distance between $\boldsymbol{o}$ and all other objects in the cluster to which $\boldsymbol{o}$ belongs. Similarly, $b(\boldsymbol{o})$ is the minimum average distance from $\boldsymbol{o}$ to all clusters to which $\boldsymbol{o}$ does not belong. Formally, suppose $o \in C_i$ ($1 \leq i \leq k$). The value of the silhouette coefficient is between -1 and 1. The value of an ($\boldsymbol{o}$) indicates the cluster to which $\boldsymbol{o}$ belongs compactly. The less the value, the more the group is lightweight [41], [38].

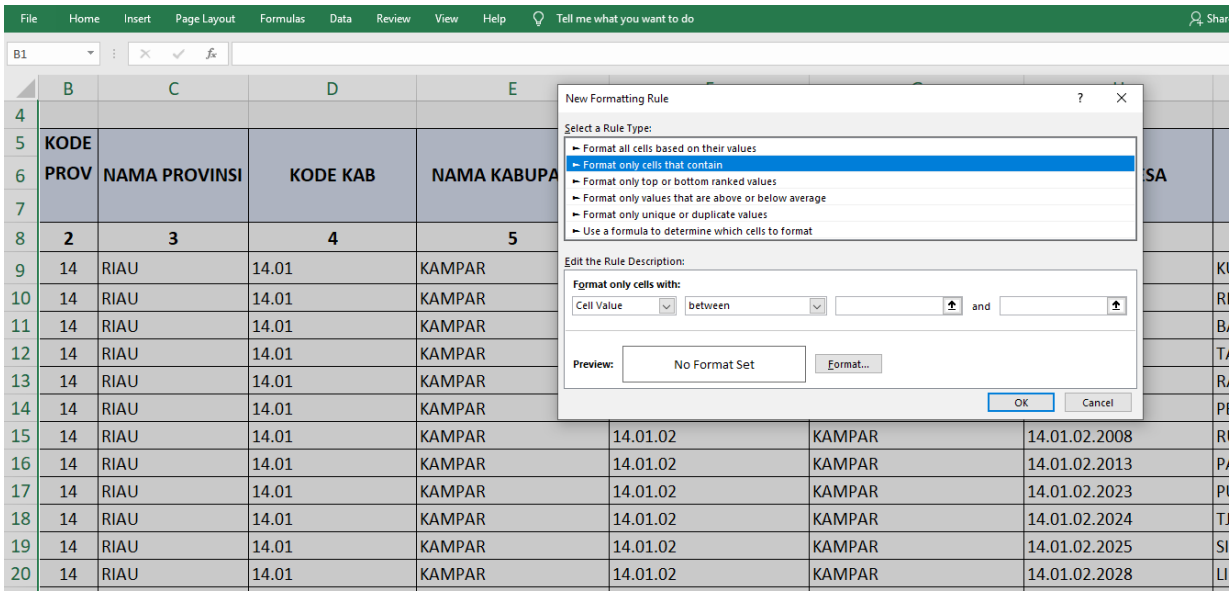
Python programming is planned and executed for the total calculation and clustering. Python is open-source packaging that carries out unsupervised graph data learning. This model offers community identification, node integration, and whole graph incorporation techniques, particularly data mining. [42]. The programming language embraces refinement, aggregation, interpolation, eigenvalues problems, algebraic equations, differential equations, and many other problems. In addition, Python's language has emerged long-term favorable and has culminated in the entire library ecosystem of related programs and social activities being interfered with [43].

III. Results and Discussion

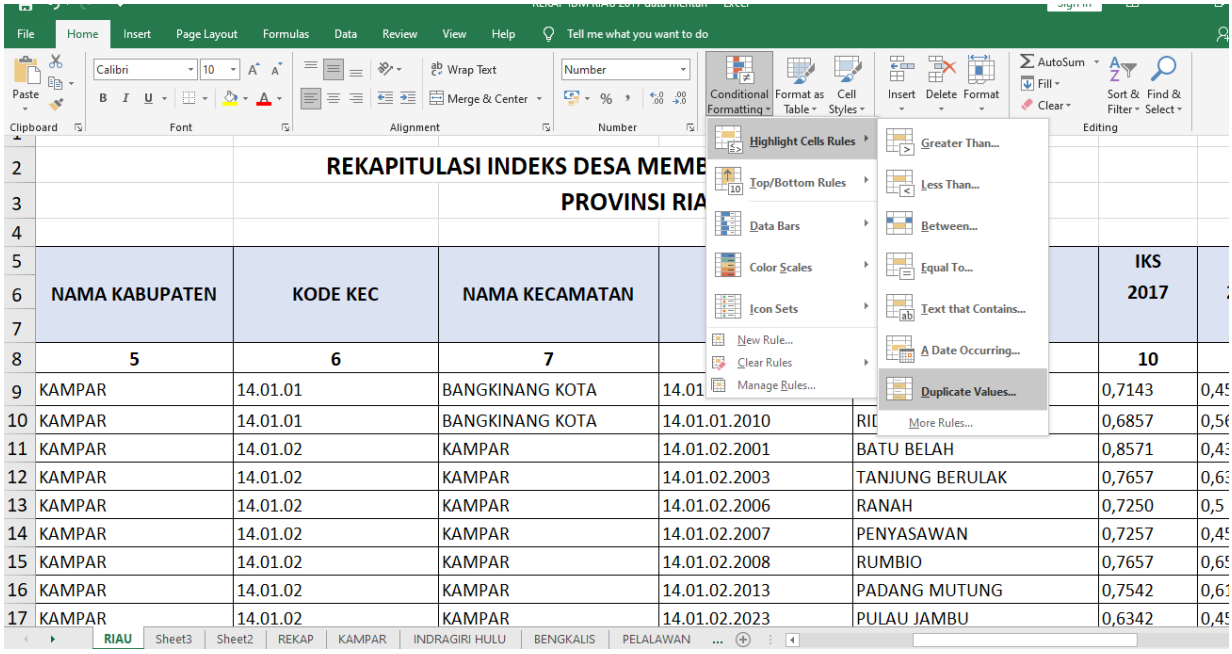
A. The execution of KDD

For further review, the KDD method was preferred in Section 2 for 1591 villages with 13 different attributes, including Provincial Code, Provincial Name, District Code, Regional Name, Sub District Name, Village Code, Village Name, IKS, IKE, IKL, IDM, and status. The data selection stage emphasizes the three main attributes as the village group's references, namely IKS, IKE, and IKL attributes. The illustration of the data selection process can be depicted in Figure 1.

Furthermore, the preprocessing stage is conducted by investigating the missing values and duplicate data. Several Excel programming functions are executed, thus found no mislead (Figure 2a) and duplex data (Figure



(a) Checking the missing values



(b) Checking the duplicate data

Figure 2. Data preprocessing

2b). Meanwhile, the data transformation stage was dismissed due to the numeric values IKS, IKE, and IKL attributes.

B. DBSCAN analysis

Since the DBSCAN tracking algorithm allows the execution of Equation (2) to (5) and the measurement of silhouette indexes in Equation (6) to (8), where *p* is set as the first data, the obtained values of Euclidean as shown in Table 2.

Hence, the toolkit library for Python is executed with an epsilon value at 0.100 and *minpts* at 0.1. As a result, the outstanding silhouette index value is determined at 0.82308. This outcome indicated the right point of data grouping due to enclose of the above value into 1. The index value has been successfully clustered the data into five groups. Cluster 0 is a group of very under-developed villages with 1577 villages numbers (silhouette coefficient = 0.47768), cluster 1 as an under-developed village with six numbers villages (silhouette coefficient = 0.81104), cluster 2 as developing villages with only one village (silhouette coefficient = 1.0), cluster 3 is a developed village with total six villages (silhouette coefficient = 0.82669), and cluster 4 as an independent village consisting of 1 village data (silhouette coefficient = 1.0). The interpretation of the clustering pattern for five villages grouping can be seen in Figure 3.

Table 2. The category of villages status

No	IKS	IKE	IKL	IDM	Euclidean
1	0.714	0.45	0.666	3	0
2	0.685	0.566	0.6	3	0.137
3	0.857	0.433	0.8	3	0.196
4	0.765	0.633	0.533	3	0.232
5	0.725	0.5	0.6	3	0.084
6	0.725	0.45	0.6	4	0.067
7	0.765	0.65	0.466	3	0.287
8	0.754	0.616	0.6	3	0.183
9	0.634	0.45	0.6	4	0.104
10	0.697	0.383	0.733	3	0.095
...	...	...	...		...
101	0.954	0833	0867	1	0.494
102	0.794	0.666	0.666	2	0.231
103	0.76	0.583	0.667	3	0.140
104	0.817	0.633	0.533	3	0.248
105	0.725	0.516	0.866	3	0.211
...	...	...	...		...
701	0.754	0.866	0.6	2	0.423
702	0.645	0.516	0.666	3	0.095
703	0.64	0.433	0.6	4	0.101
704	0.594	0.283	0.666	4	0.205
705	0.645	0.45	0.533	4	0.149
...	...	...	...		...
1587	0.714	0.566	0.649	3	0.117
1588	0.64	0.466	0.613	3	0.092
1589	0.651	0.4	0.528	4	0.160
1590	0.697	0.466	0.587	4	0.082
1591	0.622	0.516	0.602	3	0.130

This study reveals that the DBSCAN and government IDM calculation index pomp the significant differences in clustering the village's status. The comparison of the two above classifications can be seen in Figure 4. The running of (1) for the government IDM estimation index fails to explain the reasonable computation in grouping the villages and ensuring grouping validation.

Furthermore, the reasoning for DBSCAN grouping offered a more practical design with the current village conditions at the community and village empowerment agency at Riau Province, Indonesia.

C. DBSCAN evaluation

Subsequently, the silhouette calculation index on Equation (6) to (8) is compared depending on the percentage splits' equipment, as disclosed in Table 3. Table 3 showed the highest achievement of silhouette index value with randomized testing of epsilon and *minpts* rates. It reveals the percentage splits at 90:10 as the top of the silhouette index into grouping 5 clusters of villages (0.84351). The pattern performed from this interpretation can be seen in Figure 5. Hence, the comparison of villages grouping on each percentage split is outlined in Table 4.

Since the series of percentage splits tests with variant values of *minpts* and epsilon, the DBSCAN algorithm has successfully delivered the optimum numbers of silhouette index scores approaching 1. These values indicate that the epsilon and *minpts* values' determination directly affects the clustering amount produced [44]. The noise volume or outliers can be decreased or increased by the epsilon numbers' value [45]. Herein, DBSCAN presents an advantage in finding the clusters of arbitrary shapes efficiently, especially in massive grouping databases, by emphasizing the minimal needs of domain knowledge for parameters input [46].

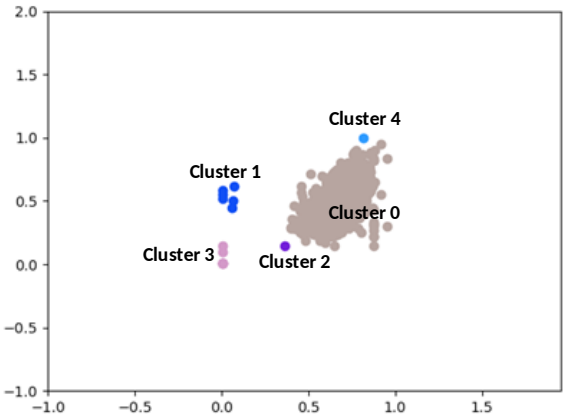


Figure 3. Interpretation of village clustering patterns

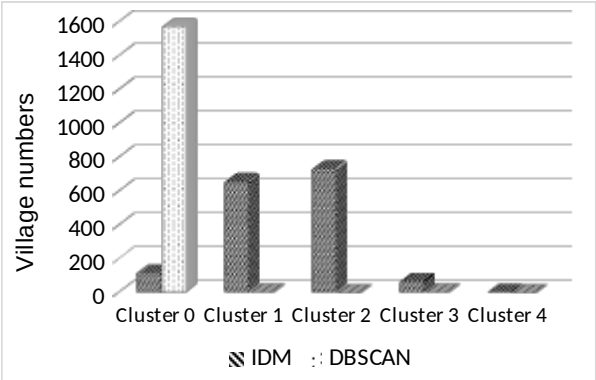


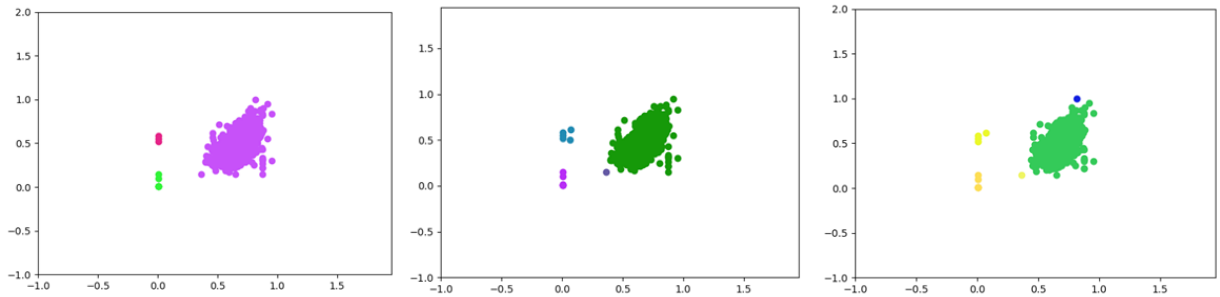
Figure 4. Government IDM and DBSCAN classification comparisons

Table 3. Clustering test results

Data Comparison	Epsilon	MinPts	Cluster	Silhouette
90:10	0.100	0.5	5	0.84351
80:20	0.100	0.05	4	0.80575
70:30	0.100	0.7	5	0.83698

Table 4. Village grouping results

Cluster	Training and test data ratio		
	90:10	80:20	70:30
Cluster 0	1419	1260	1101
Cluster 1	1	5	4
Cluster 2	6	6	1
Cluster 3	3	1	6
Cluster 4	1	-	1



(a) 90:10 pattern (b) 80: 20 pattern (c) 70:30 pattern

Figure 5. Interpretation of clustering using various training and test data ratio

IV. Conclusion

This study reveals that the DBSCAN algorithm has succeeded in handing over a novelty calculation in clustering the development villages index with unbroken reference to government regulation and standardization. DBSCAN accomplishes in grouping five villages in Riau province with the highest accuracy at 0.82308 silhouette coefficient value with epsilon and *minpts* values at 0.100 and 0.1, respectively. The evaluation presents the significant numbers of epsilon and *minpts* that directly affect many clusters composed with the percentage splits' divergent simulation.

In a nutshell, the deployment of DBSCAN in this case study provides a significant contribution to the village clustering with a better level of accuracy than ordinary mathematical calculations. This village's clustering will benefit village development planning and budget allocations and village development activity programs.

Acknowledgments

The authors acknowledged the generous assistance provided to the Faculty of Science and Technology of the State University of Islam Sultan Syarif Kasim Riau and the Riau Province Community and Village Empowerment Service, Indonesia. They contribute significantly to the feasibility and efficacy of this study. The authors also would like to thank the entire contributors, including the Universiti Teknologi Malaysia and Insa Center Val de Loire, Bourges, France.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi: 10.14710/jtsiskom.2021.13998.

References

[1] H. S. Bakti, "Identifikasi masalah dan potensi desa berbasis indek desa membangun (IDM) di desa Gondowangi kecamatan Wagir kabupaten Malang," *Wiga: Jurnal Penelitian Ilmu Ekonomi.*, vol. 7, no. 1, pp. 1–14, 2018. doi: 10.30741/wiga.v7i1.331

[2] M. Stit, N. Kusuma, And E. Purwanti, "Village index analysis building to know the village development in Gadingrejo district Pringsewu District," *Inovasi Pembangunan: Jurnal Kelitbangan*, vol. 6, no. 2, pp. 179–190, 2018. doi: 10.30741/wiga.v7i1.331

[3] A. Aprianti, M. Marliani, Y. Yunindyawati, and F. Nomaini, "Pengaruh program satu desa satu PAUD," *thesis*, Sriwijaya University, Indonesia. 2018.

[4] G. Bathla, H. Aggarwal, And R. Rani, "A novel approach for clustering big data based on Mapreduce," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 3, pp. 1711–1719, 2018. doi: 10.11591/ijece.v8i3.pp1711-1719

[5] A. Amelio and A. Tagarelli, *Data Mining : Clustering*. Encyclopedia of Bioinformatics and Computational Biology, 2018.

[6] R. Filipovych et al., "Semi-supervised cluster analysis of imaging data," *NeuroImage*, vol. 54, pp. 2185-2197, 2011. doi: 10.1016/j.neuroimage.2010.09.074

[7] A. Bewley and B. Upcroft, "Advantages of exploiting projection structure for segmenting dense 3D point clouds," in *Australasian Conference on Robotics and Automation*, Sydney, Australia, Dec. 2013, pp. 2–4.

[8] J. R. Saura, "Using data sciences in digital marketing: framework, methods, and performance metrics," *Journal of Innovation & Knowledge*, vol. 6, no. 2, pp. 92-102, 2020. doi: 10.1016/j.jik.2020.08.001

[9] Y. Yang, E. W. K. See-To, and S. Papagiannidis, "You have not been archiving emails for no reason! Using big data analytics to cluster B2B interest in products and services and link clusters to financial performance," *Industrial Marketing Management*, vol. 86, 2018, pp. 16–29, 2020. doi: 10.1016/j.indmarman.2019.01.016

[10] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Computers & Education*, vol. 143, 103676, 2020. doi: 10.1016/j.compedu.2019.103676

[11] M. C. Thomas, W. Zhu, and J. A. Romagnoli, "Data mining and clustering in chemical process databases for monitoring and knowledge discovery," *Journal*

of Process Control, vol. 67, pp. 160–175, 2018. doi: 10.1016/j.jprocont.2017.02.006
[12] S. Zheng and J. Zhao, "A new unsupervised data mining method based on the stacked autoencoder for chemical process fault diagnosis," *Computers and Chemical Engineering*, vol. 135, 106755, 2020. doi: 10.1016/j.compchemeng.2020.106755
[13] Y. Guo, N. Wang, Z. Y. Xu, and K. Wu, "The internet of things-based decision support system for information processing in intelligent manufacturing using data mining technology," *Mechanical Systems and Signal Processing*, vol. 142, 106630, 2020. doi: 10.1016/j.ymssp.2020.106630
[14] G. Grigoras and F. Scarlatache, "An assessment of the renewable energy potential using a clustering based data mining method. Case study in Romania," *Energy*, vol. 81, pp. 416–429, 2015. doi: 10.1016/j.energy.2014.12.054
[15] L. Kaufman and P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, volume (344). John Wiley & Sons, 2009.
[16] G. Karypis, E. H. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75, 1999. doi: 10.1109/2.781637
[17] D. M. Saputra, D. Saputra, and L. D. Oswari, "Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method," in *Sriwijaya International Conference on Information Technology and Its Applications*, Palembang, Indonesia, Nov. 2019, pp. 341–346. doi: 10.2991/aisr.k.200424.051
[18] S. Wang, D. Wang, C. Li, Y. Li, and G. Ding, "Clustering by fast search and find of density peaks with data field," *Chinese Journal of Electronics*, vol. 25, no. 3, pp. 397–402, 2016. doi: 10.1049/cje.2016.05.001
[19] H. P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011. doi: 10.1002/widm.30
[20] M. M. R. Khan, M. A. B. Siddique, R. B. Arif, and M. R. Oishe, "ADBSCAN: Adaptive density-based spatial clustering of applications with noise for identifying clusters with varying densities," in *4th International Conference on Electrical Engineering and Information and Communication Technology*, Dhaka, Bangladesh, Sept. 2019, pp. 107–111. doi: 10.1109/CEEICT.2018.8628138
[21] P. B. Nagpa and P. A. Mann, "Comparative study of density-based clustering algorithms," *International Journal of Computer Applications*, vol. 27, no. 11, pp. 44–47, 2011. doi: 10.5120/3341-4600
[22] M. Esther, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD-96 Proceedings*, vol. 96, no. 34, pp. 226–231, 1996.
[23] R. Arya and G. Sikka, "An optimized approach for density based spatial clustering application with noise," *in ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of the Computer Society of India*, vol I, 2014, pp. 695-702. doi: 10.1007/978-3-319-03107-1_76
[24] B. Borah and D. Bhattacharyya, "An improved sampling-based DBSCAN for large spatial databases," *in Intelligent Sensing and Information Processing*, Chennai, India, Jan. 2004, pp. 92-96. doi: 10.1109/ICISIP.2004.1287631
[25] B.Z. Qiu, X.Z. Zhang, and J.Y.I Shen, "Grid-based clustering algorithm for multi-density," in *International Conference on Machine Learning and Cybernetics*, Guangzhou, China, Aug. 2005, pp. 1509–1512. doi: 10.1109/ICMLC.2005.1527183
[26] C. Xiaoyun, M. Yufang, Z. Yan, and W. Ping, "GMDBSCAN: Multi-density DBSCAN cluster based on grid," *in IEEE International Conference on e-Business Engineering*, Xi'an, China, Oct. 2008, pp. 780–783. doi: 10.1109/ICEBE.2008.54
[27] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, Vol. 344, no. 6191, pp. 1492–1496, 2014. doi: 10.1126/science.1242072
[28] L. Yinghua *et al.*, "An efficient and scalable density-based clustering algorithm for datasets with complex structures," *Neurocomputing*, vol. 171, pp. 9–22, 2016. doi: 10.1016/j.neucom.2015.05.109
[29] C. Deng, J. Song, R. Sun, S. Cai, and Y. Shi, "Griden: An effective grid-based and density-based spatial clustering algorithm to support parallel computing," *Pattern Recognition Letters*, vol. 109, pp. 81–88, 2018. doi: 10.1016/j.patrec.2017.11.011
[30] G. Andrade, G. Ramos, D. Madeira, R. Sachetto, R. Ferreira, and L. Rocha, "G-DBSCAN: A GPU accelerated algorithm for density-based clustering," *Procedia Computer Science*, vol. 18, pp. 369–378, 2013. doi: 10.1016/j.procs.2013.05.200
[31] M. Hosseini-Rad and M. Abdolrazzagh-Nezhad, "A new hybridization of DBSCAN and fuzzy earthworm optimization algorithm for data cube clustering," *Soft Computing*, vol. 24, no. 20, pp. 15529–15549, 2020. doi: 10.1007/s00500-020-04881-0
[32] H. Hanibal et al., *Indeks desa membangun kementrian desa, pembangunan daerah tertinggal dan transmigrasi*. Jakarta, Indonesia, 2015.
[33] O. Okfalisa, R. Fitriani, and Y. Vitriani, "The comparison of linear regression method and k-nearest neighbors in scholarship recipient," in *19th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Busan, Korea, Jun. 2018, pp. 194–199. doi: 10.1109/SNPD.2018.8441068
[34] O. Okfalisa, I. Gazalba, M. Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor

and modified k-nearest neighbor algorithm for data classification," in *International Conferences on Information Technology, Information Systems and Electrical Engineering*, Yogyakarta, Indonesia, Nov. 2017, pp. 294–298. doi: 10.1109/ICITISEE.2017.8285514

[35] H. Yan, N. Yang, Y. Peng, and Y. Ren, "Data mining in the construction industry: Present status, opportunities, and future trends," *Automation in Construction*, vol. 119, no. August 2019, 103331, 2020. doi: 10.1016/j.autcon.2020.103331

[36] Han, Jiawei, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[37] E. Sharma, M. Mussetta and W. Elmenreich, "Investigating the impact of data quality on the energy yield forecast using data mining techniques," in *2020 IEEE PES Innovative Smart Grid Technologies Europe*, The Hague, Netherlands, Oct. 2020, pp. 599-603. doi: 10.1109/ISGT-Europe47291.2020.9248920

[38] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," *in International Conference on Electrical, Electronics, and Optimization Techniques*, Chennai, India, Mar. 2016. doi: 10.1109/ICEEOT.2016.7754750

[39] S.R. Kannan, "A new segmentation system for MR brain images based on fuzzy techniques," *Applied Soft Computing Journal*, vol. 8, no. 4, pp. 1599–1606, 2008. doi: 10.1016/j.asoc.2007.10.025

[40] P.J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. doi: 10.1016/0377-0427(87)90125-7

[41] V. T. P. Swindiarto, R. Sarno, and D. C. R. Novitasari, "Integration of Fuzzy C-Means Clustering and TOPSIS (FCM-TOPSIS) with silhouette analysis for multi criteria parameter data," in *International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia, Sept. 2018, pp. 463–468. doi: 10.1109/ISEMANTIC.2018.8549844

[42] B. Rozemberczki, O. Kiss, and R. Sarkar, "Karate club: an api oriented open-source python framework for unsupervised learning on graphs," in *29th ACM International Conference on Information & Knowledge Management*, Virtual Event, Ireland, Oct. 2020, pp. 3125–3132. doi: 10.1145/3340531.3412757

[43] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[44] Y. M. Elbarawy, R. F. Mohamed, and N. I. Ghali, "Improving social network community detection using DBSCAN algorithm," in *World Symposium on Computer Applications and Research*, Sousse, Tunisia, Jan. 2014, pp. 1-6. doi: 10.1109/WSCAR.2014.6916792

[45] M. Khatoon and W. A. Banu, "An efficient method to detect communities in social networks using DBSCAN algorithm," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1-12, 2019. doi: 10.1007/s13278-019-0554-1

[46] Y. Xie and S. Shekhar, "Significant DBSCan towards statistically robust clustering," in *ACM International Conference Proceeding Series*, Vienna, Austria, Aug. 2019, pp. 31–40. doi: 10.1145/3340964.3340968