Scientific Journal of Informatics
Vol. 11, No. 2, May 2024
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Village Potential Mapping: Comprehensive Cluster Analysis of Continuous and Categorical Variables with Missing Values and Outliers Dataset in Bogor, West Java, Indonesia

Nafisa Berliana Indah Pratiwi[1*], Indahwati[2], Anwar Fitrianto[3]
[1,2,3]Department of Statistics and Data Science, Faculty of Mathematics and Natural Sciences, IPB University, Indonesia

Abstract.
Purpose: This research emphasizes the need to map villages' conditions and identify village potentials, evaluate the effectiveness of development capability, and address the rural-urban development gap with clustering algorithms. The study employs the village development index (IPD) indicators obtained from the village potential dataset, with various numerical and categorical indicators, to capture both tangible and intangible aspects of village potential. Challenges such as missing data and outliers in IPD data collection can be found. The study aims to evaluate the effectiveness of clustering algorithms, with integrated and separated imputation processes, in handling these data issues and to track the development of villages in the Bogor Regency, West Java, Indonesia, based on the village's potential (PODES) dataset.
Methods: Three clustering algorithms, such as k-prototype, simple k-medoids, and Clustering of Mixed Numerical and Categorical Data with Missing Values (k-CMM) are compared. The pre-processing data, which is the imputation process for the first two algorithms, is conducted separately, while the k-CMM has an integrated imputation process. Both imputation stages are tree-based algorithms. Cluster evaluation is based on internal criteria and external criteria. Clusters resulting from the k-prototype and simple k-medoids are selected by internal validity indices and compared to k-CMM using external validity indices for several numbers of clusters (k = 3,4,5).
Result: According to data exploration, the IPD of Bogor Regency, West Java, Indonesia dataset contains ± 5% of outliers and six missing values in some chosen variables. Tree-based imputation methods are applied separately in k-prototype and simple k-medoids, jointly in k-CMM. Based on the elbow and gap statistics methods, this research aims to determine the optimum number of clusters k = 3. The internal validity indices performed on k-prototype and simple k-medoids resulting in three clusters (k = 3) are optimum. Trials on several clusters (k = 3,4,5) for three algorithms show that the k-prototype with k = 3 performs the best and is most stable among the two other algorithms with IPD datasets containing many outliers; external validity indices evaluate cluster results.
Novelty: This research addresses issues commonly found in mixed datasets, including outliers and missing values, and how to treat problems before and during cluster analysis. An improvement of Gower distance is applied in the medoid-based clustering algorithm, and the k-CMM algorithm is the first algorithm to integrate the imputation process and clustering analysis, which is interesting to explore this algorithm's performance in clustering analysis.

Keywords: Clustering, Mixed dataset, Missing value, Outliers, Village mapping
Received April 2024 / **Revised** May 2024 / **Accepted** May 2024

INTRODUCTION

The distance-based clustering algorithm is divided into hierarchical clustering and non-hierarchical clustering. Hierarchical clustering produces a series of partitions in which, during each clustering process, two clusters are combined (agglomerative) or one cluster is divided into two (divisive) according to certain criteria, leading to hierarchical partitions represented by a dendrogram [1]. Non-hierarchical or partitional clustering divides data into k clusters, generally starting from an initialization and using an iterative relocation procedure. Hence, data moves from one cluster to another to obtain an optimal objective function [2]. The k-means algorithm is a partitional clustering algorithm for numerical data types that has faster computational time and produces denser clusters than hierarchical clustering algorithms [3]. An extension of the k-means algorithm, k-modes, is designed for clustering categorical data types by replacing the Euclidean distance measure with the Hamming distance [4]. The k-prototype integrates both the k-means and k-modes algorithms for clustering data consisting of numerical and categorical variables. The k-

*Corresponding author.
Email addresses: nafisaberliana@apps.ipb.ac.id (Pratiwi)
DOI: 10.15294/sji.v11i2.3903

prototype is useful in practice because objects often encountered in daily life problems are objects with mixed data types [5]. Clustering algorithms for both numerical and categorical data types continue to be developed to have the ability to be applied to mixed datasets. Simple k-medoids [6] is one of the mixed data clustering algorithms developed from k-medoids, designed for numerical data clustering. K-medoids is robust or less sensitive to outliers [7].

The data collection process for mixed data only sometimes yields complete data due to potential individual errors, data entry mistakes, malfunctioning equipment, inconsistencies in data, and data deletion, which lead to missing data and outliers. Observations potentially detected as outliers also unavoidably degrade the clustering performance [8]. Missing values might be examined as the key factors impacting the clustering performance [9], for instance, significantly degrading the clustering process [10] or leading to inaccuracies in clustering, undermining the reliability and validity of the analysis [11]. Several unsupervised outlier detection methods have been proposed from various perspectives to address the negative effects of outliers in clustering processes [8]. An initial adjustment has been made to the standard Gower's distance (commonly used for distance in mixed variables clustering analysis) to compensate for the effects of outliers in interval or ratio-scaled variables [12]. According to Dinh et al. 2021 [13], possible solutions for the clustering process in incomplete data include (i) performing data pre-processing such as data imputation to reduce the risk of decreased clustering algorithm performance, (ii) applying data to clustering algorithms that perform well on incomplete datasets. The clustering of mixed numerical and categorical data with missing values (k-CMM) algorithm is the first clustering algorithm to be integrated with an imputation process for missing data. It efficiently improves the quality of cluster results for mixed data with missing information [13].

Clustering algorithms designed for mixed data can assist in mapping the potential of villages, including natural and human resources, and assess the performance of development activities in villages. Mapping conditions and measuring village development performance is crucial in reducing the gap between rural and urban development, which tends to be urban-biased. To achieve this goal, improvements and services to residents are necessary, including infrastructure development and empowerment in various sectors such as the economy, education, health, energy, transportation, and communication facilities. The key to successful mapping is the accurate and detailed information obtained from the clustering algorithm results. In the last five years, villages or wide-ranging district mapping in various regions of Indonesia has been extensively conducted through varied cluster analysis with numerical variables or mixed variables (numerical and categorical variables). Several sectors discussed in area mapping with cluster analysis include tourism (e.g., [14]; health and education conditions [15]–[17] village social-economy [18], and the impact of natural disasters [19], [20], and the overall villages potential [21], [22]. Previously mentioned research mostly used the k-means or an improved clustering algorithm based on k-means, which works on numerical datasets.

Accurate village potential mapping can serve as a reference for the government and stakeholders to identify villages that require resource improvement. Moreover, village condition mapping with clustering algorithms opens opportunities for more innovative and measurable approaches in rural development, potentially guiding to greater and more sustainable impacts. The Village Development Index (IPD) consists of several indicators with numerical and categorical (mixed) data types. The IPD uses indicators obtained from both physical and non-physical village potentials. In the data collection process for IPD, challenges such as incomplete data or deviant data (outliers) are often encountered. This research is a continuation of Chrisinta et al. (2020) [21], which utilized the k-prototype for mixed data clustering and adopted variables chosen with an adjustment of variables according to IPD publications and includes further exploration related to outliers and missing data. This research investigates the actual condition of a real-case dataset, which may contain outliers or missing values in survey-based datasets as potential villages' (PODES) data, and its treatment before or during the cluster analysis. In addition, this research aims to compare the performance of clustering algorithms that integrate the imputation process with those that use a separate imputation process to ensure accurate mapping and measurement of village development performance.

METHODS

Data description

The empirical data used in this study are secondary data from the 2021 Village Potential Survey (PODES) collected by the Central Bureau of Statistics Indonesia (BPS). The observation unit comprises 433 Bogor Regency, West Java, Indonesia villages. There is a total of 32 mixed variables, including (i) 17 categorical variables (including ordinal, nominal, symmetric binary, and asymmetric binary) and (ii) 15 numerical variables. The variables selection is based on the latest Village Development Index (IPD) publication [23], which covers five dimensions. Details of dimensions and the number of numerical and categorical variables used in this research are shown in Table 1.

Table 1. Dimensions and number of mixed variables used in the research

Dimension	Categorical	Numerical
Basic services	0	3
Infrastructure condition	7	1
Accessibility and Transportation	2	4
Public services	3	3
Governance	3	2

The variables contained in each dimension used in this research continue the research by Christina et al. 2020 [21]. Table 2 provides details of the variable description involved in each dimension.

Table 2. The details of each variable involved in all dimensions

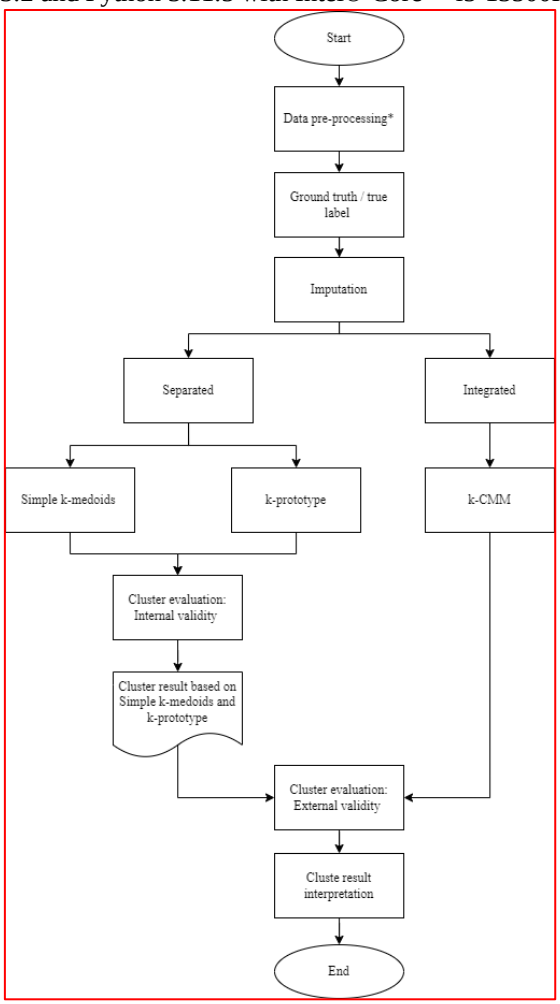
Variable	Description
Dimension: Basic services	
$X1_{(N)}$	The total number of facilities across all educational levels
$X2_{(N)}$	The total number of facilities across all health levels
$X3_{(N)}$	The total number of health professionals
$X4_{(N)}$	The total number of facilities across all economic sectors
Dimension: Infrastructure condition	
$X5_{(N)}$	The total amount of various types of cooking fuel used by the village's resident
$X6_{(C)}$	The primary sources of drinking water used by most of the village's resident
$X7_{(C)}$	The primary sources of bathing/washing water used by most of the village's resident
$X8_{(C)}$	The presence of village residents using handphones
$X9_{(C)}$	The Internet facilities at the village head's office/subdistrict office
$X10_{(C)}$	The operational status of postal office facilities/auxiliary post/post house
$X1_{(C)}$	The Operational status of private courier service companies/agents (for shipping goods/documents) facilities
Dimension: Accessibility and transportation	
$X12_{(C)}$	The most common type of road surface between villages/sub-districts
$X13_{(C)}$	The presence of public transportation
$X14_{(C)}$	Travel time per kilometer from the village head or sub-district office to the district office (Minutes)
$X15_{(C)}$	Transportation costs per kilometer from the village head or sub-district office to the district office (Rupiah)
$X16_{(C)}$	Travel time per kilometer from the village head or sub-district office to the mayor's office
$X17_{(C)}$	Transportation costs per kilometer from the village head or sub-district office to the mayor's office (Rupiah)
Dimension: Public services	
$X18_{(C)}$	The existence of dengue fever outbreaks
$X19_{(N)}$	The number of dengue fever outbreak cases
$X20_{(C)}$	The existence of Covid-19 outbreaks
$X21_{(N)}$	The number of COVID-19 outbreak cases
$X22_{(C)}$	The availability of sports facilities
$X23_{(N)}$	The number of sports team
Dimensions: Governance	
$X24_{(C)}$	The existence of the village's head
$X25_{(C)}$	The gender of the village's head
$X26_{(C)}$	The highest education level of the village's head
$X27_{(N)}$	The age of the village's head
$X28_{(C)}$	The existence of the village's secretary
$X29_{(C)}$	The gender of the village's secretary

Variable	Description
$X30_{(C)}$	The highest education level of the village's secretary
$X31_{(N)}$	The age of the village's secretary
$X32_{(N)}$	The number of village government officials

Each variable's indexing $_{(C)}$ refers to categorical and $_{(N)}$ to numerical variables.

Research procedures

Each procedure described above in this research can be shown as a flowchart in Figure 1. This research was conducted in R-4.3.2 and Python 3.11.5 with Intel® Core™ i5-13500H and RAM 16GB.



*Data pre-processing includes several stages such as (i) missing data examination and outlier identification, (ii) Gower distance transformation for data's ground truth, cluster structure discovery, and simple k-medoids input.

Figure 1. The flowchart of research procedures

The research procedure of clustering on village potential datasets of Bogor Regency is as follows:

1. Exploration and visualization of village potential datasets of Bogor Regency.
2. Missing data examination and outlier identification (pre-processing data).
3. Data Transformation into a Gower Distance Matrix.

The most commonly used metric for mixed-type variables is Gower's distance. This approach computes an average of individual variable distances, each normalized to fall between 0 (indicating the least distance) and 1 (representing the greatest distance). Gower's distance is particularly useful due to its ability to handle missing data and its flexibility in allowing the user to apply custom weightings when averaging the distances across various variables. The Gower's distance can be defined as the complement to one of the Gower's similarity coefficients in equation (1)

$$d_{G,ij} = 1 - s_{G,ij} = \frac{\sum_{t=1}^{p} \delta_{ijt} d_{ijt}}{\sum_{t=1}^{p} \delta_{ijt}} \quad (1)$$

A modification of the standard Gower's distance [12] to compensate for the impact of outliers in interval or ratio-scaled variables consists simply of the replacement of the range R_t with the inter-quartile range written in equation (2)

$$d_{ijt} = \begin{cases} \frac{|x_{it} - x_{jt}|}{IQR_t}, \text{ if } |x_{it} - x_{jt}| < IQR_t \\ 1, \text{ otherwise} \end{cases} \quad (2)$$

where $IQRt$ is the inter-Quartile range ($P75\% - P25\%$), this modification permits to account for outliers.

4. Tree-Based Imputation: Performing tree-based imputation separately (Step 5) and integrated (Step 7)
5. Separate imputation followed by clustering process with algorithms:
   i. Agglomerative Hierarchical Clustering: The basic steps involved in the agglomerative hierarchical clustering algorithm are as follows: (a) calculating the dissimilarity matrix on empirical data, (b) merging clusters so $C_{a \cup b} = C_a \cup C_b$, (c) determining the cardinality of the new cluster ($N_{a \cup b} = N_a + N_b$), (iv) the process continues to new rows and columns of the dissimilarity matrix representing the distance between the new cluster $C_{a \cup b}$ and remaining clusters until one maximum cluster remains. The cluster analysis with agglomerative hierarchical clustering aims to create data labels/ground truth. Data labels/ground truth are selected from the best cluster results and are then used as a reference compared with clustering results of algorithms (ii), (iii), and the algorithm in Step 8.
   ii. The k-prototype algorithm falls under non-hierarchical clustering and was first proposed by Huang [24]. The k-prototype algorithm integrates the processes of k-means and k-modes for clustering data with mixed numerical and categorical variables. It adopts an approach similar to the k-means algorithm, thus maintaining the efficiency of k-means. The distance measurement in equation (3) used in the k-prototype algorithm for numerical attributes is the squared Euclidean distance, and for categorical attributes, it utilizes simple matching dissimilarity measurement.

$$d(X,Y) = \sum_{j=1}^{p} \left(x_j - y_j\right)^2 + \gamma \sum_{j=p+1}^{m} \delta\left(x_j, y_j\right)$$

with

$$\delta\left(x_j, y_j\right) = \begin{cases} 0, & (x_j = y_j) \\ 1, & (x_j \neq y_j) \end{cases} \quad (3)$$

   Gamma (γ) is a single weighting for the overall contribution of categorical variables [25]. This algorithm aims to cluster *x* datasets into *k* clusters by minimizing the objective function. The k-prototype involves three processes: initialization of the initial prototype, initial allocation, and reallocation [26].
   iii. Simple k-medoids [6] is a distance-based clustering algorithm developed from k-medoids and simple and fast k-medoids. K-medoids and its developments are more robust against outliers [27]. Simple k-medoids address the weaknesses of previous medoids-based algorithms, thus producing better clusters than the simple and fast k-medoids. Simple k-medoid uses the generalized distance function (GDF) measure, which can measure more variations of categorical variables. The Gower distance is the commonly used distance measure for data with mixed variables [12]. The Gower distance can be formulated into the GDF in equation (4).

$$d_{ij} = \frac{1}{p_n + p_b + p_c} \left(\sum_{r=1}^{p_n} \frac{1}{R_r} \left| x_{ir} - x_{jr} \right| + p_b \sum_{t=1}^{p_b} \delta_b\left(x_{it}, x_{jt}\right) + p_c \sum_{s=1}^{p_c} \delta_c\left(x_{is}, x_{js}\right) \right) \quad (4)$$

   The Gower distance consists of a combination of weighting and distance. Numerical and categorical variables each contribute to the GDF function. Each observation with the same data type contributes equally to the distance calculation process.
6. Evaluating clusters resulted from i, ii, and iii, with an internal validity index: Algorithm (i) is evaluated based on the agglomerative coefficient, whereas algorithms (ii) and (iii) are evaluated using an internal validity index first, followed by Step 9.
7. Integrated imputation followed by clustering process with the algorithm:
   iv. Mixed Numerical and Categorical Data with Missing Values (k-CMM): The k-CMM is the first clustering algorithm that handles the clustering process on mixed data with missing data where data pre-processing is in the form of integrated data imputation within the clustering algorithm [13]. This algorithm is designed to improve clustering results by combining the imputation process

on missing data and can be applied to data with mixed variables without specific assumptions. Both distance measures used in the k-CMM are the Squared Euclidean Distance and an information-theoretic-based dissimilarity. The dissimilarity measure of a mixed object x_i to cluster center Z_j is formulated as equation (5).

$$dis(x_i, Z_j) = \sum_{d=1}^{m} dis_d(x_i, Z_j)$$

with

$$= \begin{cases} dis_d(x_i, Z_j) & \text{if } d \text{ is numerical attribute} \\ \sum_{o_{ld}^j \in O_d^j} P_d^j(O_{ld}^j) dsim_d(x_{id}, O_{ld}^j), & \text{if } d \text{ is a categorical attribute} \end{cases} \quad (5)$$

8. Evaluating the cluster results using an external validity index.
9. Visualization and interpretation of the cluster results.

RESULTS AND DISCUSSIONS

Table 2 summarizes the amount of missing data and detected outliers. The correlation coefficient used is Spearman's correlation coefficient (ρ). The range of the numeric variable is ρ = [-0.6, 0.6], with the lowest value indicating no correlation between variables, such as variable X17 (Transportation cost from the village/headquarters to the regency mayor's office) and variable X21 (Number of Covid-19 outbreak cases). There is a strong positive relationship between variables, such as X1 (Total number of educational facilities across all levels) and X2 (Total number of health facilities).

Table 3. Numbers of missing values and outliers in each variable

Variables	Missing values	Outliers
Numerical	2	285*
Categorical	4	0

*Total number of outliers across all numerical variables (n = 15)

The village potential data, imputed using a machine learning-based imputation method, includes integrated and separated tree-based imputations. Separate imputation utilizes the package Multiple Imputation by Chained Equations (MICE) [28] in R for both k-prototype and simple k-medoids, and the imputation integrated with k-CMM is decision-tree-based. To explore the cluster structure in the village's potential data transformed into a Gower distance matrix [12], this is represented by a heatmap in Figure 1. According to Kassambara [29], red indicates high similarity, and blue indicates low similarity. Figure 1 shows that the red and blue areas form a pattern rather than being randomly distributed, indicating the presence of a cluster structure within village potential (PODES) data of Bogor Regency. The Gower distance matrix is used as input to estimate the optimal number of clusters generated using the agglomerative hierarchical clustering algorithm.

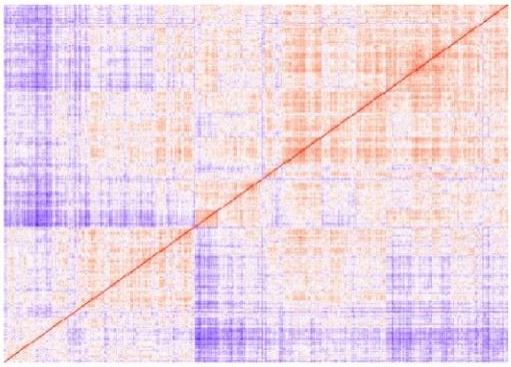
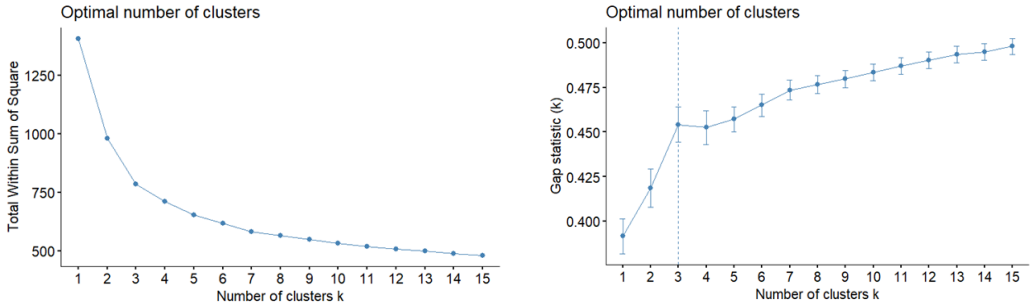


Figure 1. Heatmap for mixed variables village dataset transformed into Gower distance matrix

Estimating the optimal number of clusters is based on (i) the Elbow method and (ii) the Gap statistic, followed by determining the hierarchical clustering algorithm: complete linkage, single linkage, and

average linkage. According to the Elbow method, an Elbow point indicates that the optimum number of formed clusters is k = 3. Meanwhile, according to the Gap Statistics, which compares the change in within-cluster dispersion, the optimal number of clusters is also estimated to be k = 3. Therefore, the optimal number of clusters in this study will be predicted to be similar to the estimations of these two methods.



(i) Elbow (ii) Gap statistics

Figure 2. Number of optimal cluster approximation

Table 3 presents the agglomerative coefficient (AC) as an indicator for determining which hierarchical clustering algorithm will be used. The agglomerative coefficient (AC) ranges between [0,1] and describes the strength of the generated cluster's structure [30]. A coefficient value approaching 1 indicates that the resulting cluster is stronger or better [31]. Complete linkage has the highest agglomerative coefficient compared to the other two algorithms. The relatively high coefficient of complete linkage indicates that this method creates relatively dense and consistent clusters.

Table 4. Agglomerative coefficient (AC) of hierarchical clustering algorithms

Hierarchical clustering algorithm	Agglomerative coefficient (AC)
Complete linkage	0.722
Single linkage	0.428
Average linkage	0.577

Cluster validity is categorized into two main types: internal criterion validity and external criterion validity indexes. Both evaluate the data's degree of similarity or dissimilarity [32]. Internal validity assesses the goodness of a cluster structure without considering external cluster information. Table 4 shows the internal validity measures for the complete linkage algorithm. The internal validity measure, the Silhouette index, ranges between [-1,1], measuring the degree of a clustering's confidence in a particular observation; well-clustered observations are close to 1, and poorly clustered ones are close to -1. The Dunn Index ranges from [0 ∞), where higher values indicate better cluster density and separation. The Cophenetic Correlation Coefficient (CCC) measures the correlation between elements of the original dissimilarity matrix and elements of the dendrogram-produced matrix [33]. The CCC values range between [-1,1]. High CCC values indicate that the dendrogram in hierarchical clustering accurately reflects the original distances between samples, indicating precise clustering. The Connectivity coefficient ranges from [0,∞) [34]. A lower Connectivity coefficient indicates better-formed clusters.

Table 5. Internal validity index of complete linkage

Algorithm	Internal measurements			
	Silhouette	*Dunn*	CCC	*Connectivity*
Complete linkage	0.227	0.372	0.501	48.431

CCC: *cophenetic correlation coefficient*

Further cluster analysis with the k-prototype and simple k-medoids algorithms begins by determining the number of clusters, *k*. Determining the number of clusters refers to the rule $k = \sqrt{n/2}$ [35] for every *n* object in data and *p* variables. The possible number of clusters formed is $k = \sqrt{435/2} \approx 15$ clusters. A total of 100 iterations are used for each *k* value. The clusters (*k* = 2, …,15) generated by the k-prototype and simple k-medoids algorithms are first evaluated using the internal validity index, the Silhouette index. The choice of the Silhouette index to validate internal criteria is based on a study by Aschenbruck et al. 2020

[36], which concluded that the Silhouette index is the most suitable for clustering mixed data in terms of computational time and accuracy in the determination of the optimal number of clusters. The optimal number of clusters from the k-prototype and simple k-medoids algorithms are chosen based on the highest Silhouette Index (SI) value from trials of possible cluster numbers (k = 2, 3, …,15). The highest Silhouette Index (SI = 0.204) values were obtained at k = 3 for the k-prototype algorithm and k = 5 for simple k-medoids (SI = 0.177). Therefore, the clusters to be compared with each algorithm (k-prototype, simple k-medoids, and k-CMM) for further analysis are k = 3, 4, and 5.

The performance of the k-prototype, simple k-medoids, and k-CMM algorithms are measured based on the external validity index. External validity is a simple and widely applied measure that utilizes external information. External validity measures the clustering algorithm results' appropriateness with the data's ground truth (if available) or other clusters' results [37]. The external validity index aims to obtain the maximum score within a certain range of indices and is considered the standard for optimum clustering results [38]. Table 5 shows the five external validity index values of the k-prototype, simple k-medoids, and k-CMM algorithms at various k values (k = 3,4,5).

Table 6. External validity indexes for k-prototype, simple k-medoids, and k-CMM with complete linkage as the ground truth

	Purity	NMI	Homogeneity	Completeness	V-measure
k = 3					
k-prototype	0.695	0.251	0.440	0.369	0.401
Simple k-medoids	0.709	0.216	0.354	0.358	0.356
k-CMM	0.493	0.100	0.095	0.107	0.100
k = 4					
k-prototype	0.642	0.238	0.386	0.383	0.385
Simple k-medoids	0.524	0.183	0.308	0.311	0.309
k-CMM	0.509	0.132	0.128	0.137	0.132
k = 5					
k-prototype	0.503	0.201	0.331	0.340	0.335
Simple k-medoids	0.434	0.180	0.301	0.310	0.306
k-CMM	0.460	0.138	0.134	0.142	0.138

k-prototype clustering algorithm at k = 3 shows the highest values across all five indices compared to the other two algorithms at the same k value, becoming the best algorithm due to its consistent index results compared to simple k-medoids and k-CMM at different

The external validity measurement includes (i) purity, which measures the proportion of the largest member of one class in each cluster, indicating how dominant one class is in a cluster; (ii) Normalized Mutual Information (NMI), which measures the conformity between clustering results and original classes (ground truth), depicting how well the clustering reflects the original class structure; (iii) Homogeneity, which describes whether each cluster consists only of members of a single class, assessing a class's uniformity within each cluster; (iv) Completeness, which represents how well members of one class are grouped in the same cluster, indicating how effectively clustering gathers all members of the same class; and (v) V-measure, which quantifies the harmonic mean of homogeneity and completeness, providing a combined measure of both aspects in evaluating the quality of clustering.

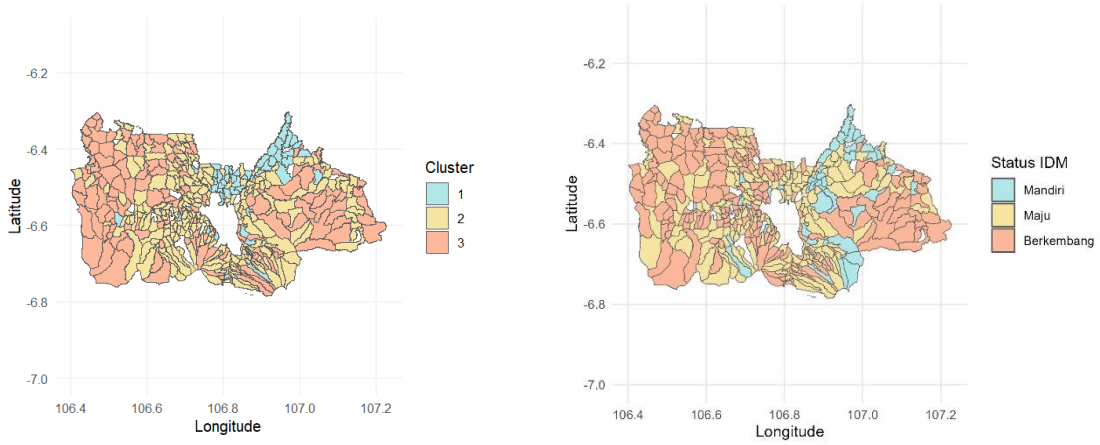
The external validity measures for clustering with the previous three algorithms further utilize ground truth information from the 2021 Village Development Index (Indeks Desa Membangun [IDM]) publication for the village status in Bogor Regency, West Java, Indonesia. The exploration results show there are 49 villages with a "Mandiri" (Self-reliant) status, 203 villages with a "Maju" (Advanced) status, and 181 villages with a "Berkembang" (Developing) status.

Table 7. External validity indexes of k-prototype, simple k-medoids, and k-CMM with ground truth: Village Development Index (IDM)

	Purity	NMI	Homogeneity	Completeness	V-measure
k = 3					
k-prototype	0.559	0.059	0.116	0.109	0.112
Simple k-medoids	0.388	0.044	0.080	0.091	0.085
k-CMM	0.562	0.046	0.045	0.047	0.046

The k-prototype clustering algorithm with k = 3 demonstrates the highest values across all five indices compared to the other two algorithms.

Figure 3 shows the distribution of villages clustered by the k-prototype and IDM status. The k-prototype algorithm results in three clusters: Cluster 1, Cluster 2, and Cluster 3, which have a similar location pattern to the villages' distribution according to IDM status. Thus, each cluster of the k-prototype results is an approximation to clusters derived from the IDM status, namely "Mandiri" (Self-reliant), "Maju" (Advanced), and "Berkembang" (Developing).



(i) k-prototype (ii) IDM status

Figure 3. Distribution of villages in Bogor Regency

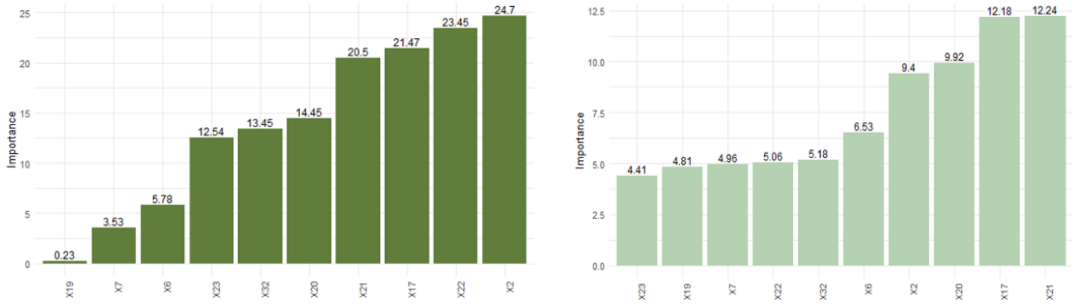
The number of cluster members from the k-prototype cluster result (Cluster 1, 2, 3) and their correspondence based on IDM status: "Mandiri" (Self-reliant), "Maju" (Advanced), and "Berkembang" (Developing) are shown in a cross-tabulation in Table 7.

Table 8. Cross tabulation of k-prototype clustering result and IDM status

Cluster	Self-reliant	Advanced	Developing	Total observation
1	15	32	3	50
2	31	141	92	264
3	3	30	86	119
Total observation	49	203	181	433

This research reveals that Cluster 1 consists of 15 villages clustered as "Mandiri" (Self-Reliant), but the majority, 32 villages, are categorized under the "Maju" (Advanced) village status. Out of a total of 433 villages, Cluster 1 is predominantly composed of "Maju" (Advanced) and "Mandiri" (Self-Reliant) villages. Cluster 2 contains the largest number of villages, most of which fall under the "Maju" (Advanced) and "Berkembang" (Developing) statuses. Meanwhile, Cluster 3 predominantly focuses on the "Berkembang" (Developing) status and has the smallest number of "Mandiri" (Self-Reliant) villages.

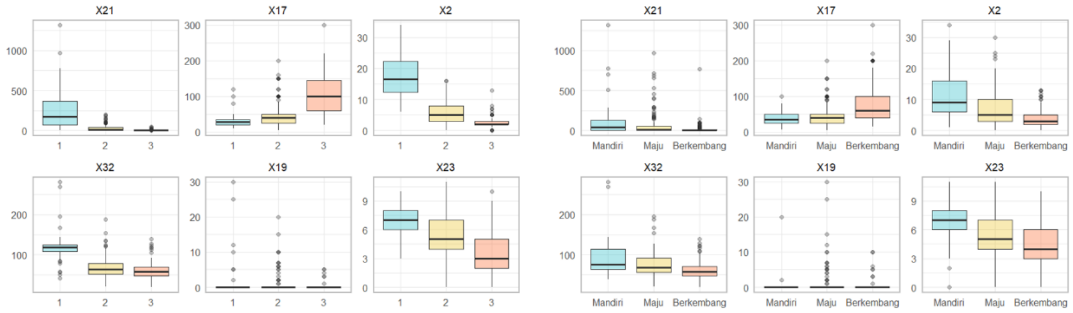
The distinguishing variables in clusters resulting from the k-prototype and IDM statuses were identified through feature importance using the random forest algorithm. Figure 4 shows the 10 variables with the highest importance scores from each ground truth



(i) k-prototype (ii) IDM status

Figure 4. Distinguishing variables of k-prototype and IDM status based on PODES data

Based on (i) and (ii), the 10 distinguishing variables for each cluster result consist of the same numerical and categorical variables, yet in a different order of contributions. The variable representing the total number of healthcare facility levels (X2) emerged with the highest importance score in the k-prototype clustering result. In contrast, variable X21, which represents the number of extraordinary occurrences such as COVID-19, has the highest importance value based on the IDM status cluster. The distinct characteristics of the identified clusters by the algorithm k-prototype and IDM status are observed in the numeric variables illustrated in Figure 5 and the categorical variables shown in Figure 6.

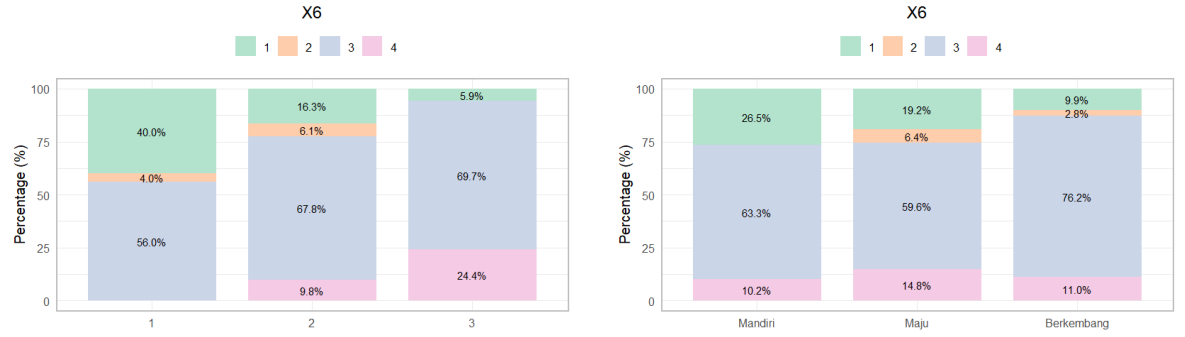


(i) k-prototype (ii) IDM status

Number of extraordinary events (KLB): Covid-19 (X21), transportation costs from the village head's office to the regency/mayor's office (X17), total number of health facility levels (X2), number of village officials (X32), number of extraordinary events (KLB): Dengue Fever (DBD) (X19), number of sports groups (X23).

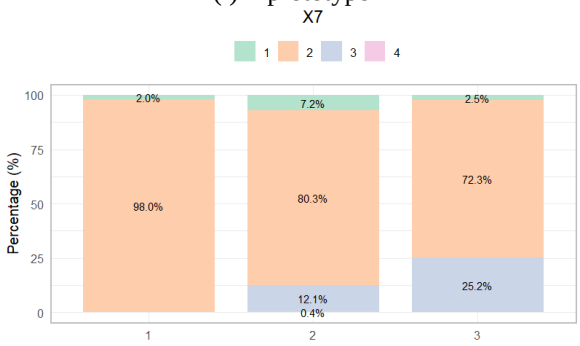
Figure 5. Boxplot of numerical variables distinguishing the clusters

Cluster characteristics resulting from the k-prototype and IDM status show similar distribution patterns for each numeric variable. X21 has a wide range in Cluster 1 and the “Mandiri” (Self-reliant) cluster, indicating significant variability in the spread of COVID-19 across different villages. Related to transportation costs (X17), villages classified under the “Berkembang” (Developing) IDM status or Cluster 3 tend to have higher transportation costs compared to those in the “Mandiri” (Self-reliant) cluster (k-prototype: Cluster 1) and the "Maju" (Advanced) cluster (k-prototype: Cluster 2). The total number of health facilities (X2) does not show significant differences among clusters in the k-prototype and IDM status.

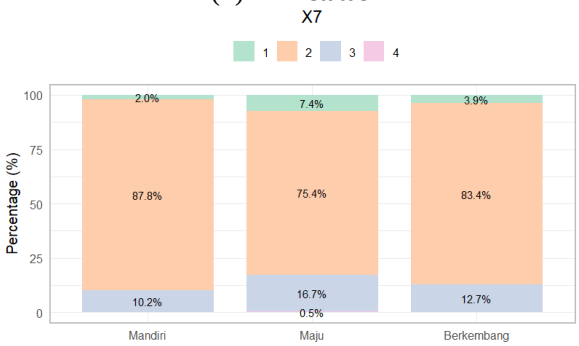


(i) k-prototype

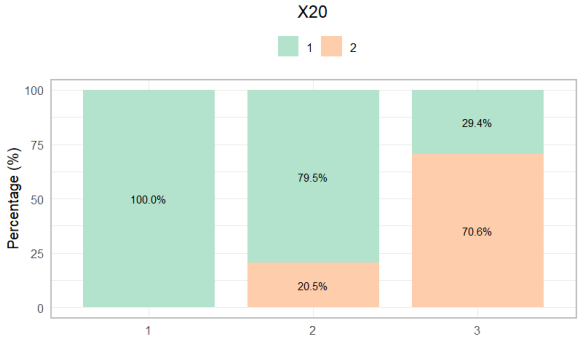
(ii) IDM status



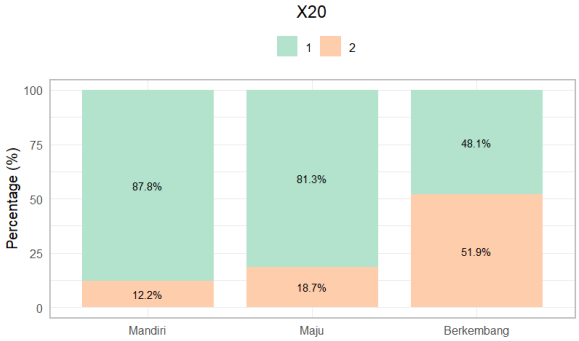
(i) k-prototype



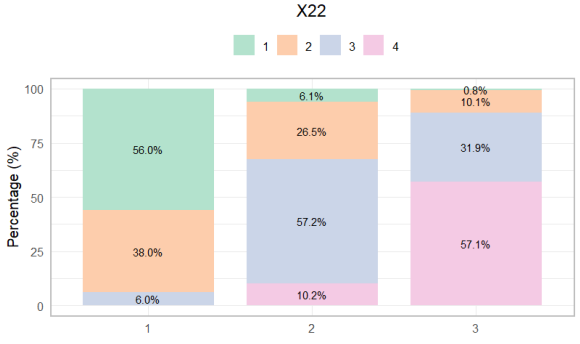
(ii) IDM status



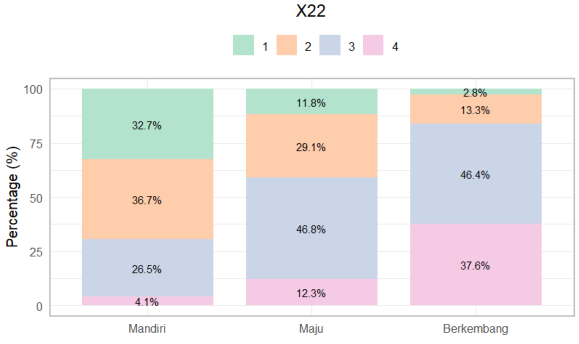
(i) k-prototype



(ii) IDM status



(i) k-prototype



(ii) IDM status

The primary source of drinking water for most families (X6), the primary source of water for bathing/washing for most families (X7), the occurrence of extraordinary events (KLB): COVID-19 (X20), and the availability of sports facilities (X22).

Figure 6. Bar plot of categorical variables distinguishing the clusters

X6 reveals that the primary drinking water sources vary, with bottled water being more dominant in Cluster 1 or “Mandiri” (Self-reliant) villages. At the same time, the use of wells is higher in Cluster 3 or the

"Berkembang" (Developing) cluster. Most village clusters from the k-prototype and villages categorized by IDM status use metered or non-metered piped water as their source for bathing and washing. Variable X20 shows that COVID-19 is recorded in all clusters from both k-prototype and IDM status. However, Cluster 3 and villages with the "Berkembang" (Developing) status have a higher percentage of villages that did not experience extraordinary events such as COVID-19.

CONCLUSION

This research compares the performance of clustering algorithms with separated imputation processes (k-prototype and simple k-medoids) to a clustering algorithm integrated with the imputation process (k-CMM) to map villages in Bogor Regency. IDM status and algorithm-based ground truths were embedded into the dataset to measure the algorithms' performance based on external criteria. The results show that the k-prototype is superior, producing more homogeneous, complete, and representative clusters.

The k-prototype successfully identified village conditions into three clusters that correspond to the Village Development Index (IDM) statuses: "Mandiri" (Self-Reliant), "Maju" (Advanced), and "Berkembang" (Developing). Cluster 1 has better access to infrastructure and services, while Clusters 2 and 3 face challenges in infrastructure and transportation. Integrating PODES data with IDM status provides valuable insights for improving policies and resource allocation for further village development. Unlike the k-prototype, simple k-medoids and k-CMM show slight changes in ground truth and a different number of clusters. Based on external validation, simple k-medoids ranks second. Simple k-medoids can adjust weights for numerical and categorical variables; using an improved Gower distance in this research minimizes the impact of outliers, enhancing clustering results.

A limitation of this research is the unavailability of ground truth in the dataset, requiring the construction of new ground truth, which is complex and time-consuming. Metrics such as computational time and memory usage could not be compared due to the different software used. Future research should use data with wider coverage, available labels, or more missing values in a mixed dataset. Standardizing software and tools would enable the use of additional metrics, providing a more comprehensive understanding of the algorithms.

REFERENCES

[1] P. Giordani, M. B. Ferraro, and F. Martella, *An Introduction to Clustering with R*, vol. 1. Singapore: Springer Singapore, 2020. doi: 10.1007/978-981-13-0553-5.

[2] O. Nasraoui, N. Chir, and C.-E. Ben, "Clustering Methods for Big Data Analytics," *Cham Springer Int. Publ.*, 2019, doi: 10.1007/978-3-319-97864-2.

[3] S. S. J and S. Pandya, "An Overview of Partitioning Algorithms in Clustering Techniques," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 5, no. 6, pp. 1943–1946, 2016.

[4] K. S. Dorman and R. Maitra, "An efficient k-modes algorithm for clustering categorical datasets," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 15, no. 1, pp. 83–97, 2022, doi: https://doi.org/10.1002/sam.11546.

[5] G. Gan, C. Ma, and J. Wu, "Data Clustering: Theory, Algorithms, and Applications," *Soc. Ind. Appl. Math.*, 2007, doi: 10.1137/1.9780898718348.

[6] W. Budiaji and F. Leisch, "Simple K-Medoids Partitioning Algorithm for Mixed Variable Data," *Algorithms*, vol. 12, no. 9, p. 177, 2019, doi: 10.3390/a12090177.

[7] A.-F. Karam, A. Helmy, and A. Mohammed, "An approach to enhance KNN based on data clustering using K-medoid," *Int. Mobile, Intelligent, Ubiquitous Comput. Conf.*, 2022, doi: 10.1109/MIUCC55081.2022.9781724.

[8] H. Liu, J. Li, Y. Wu, and Y. Fu, "Clustering with Outlier Removal," *IEEE Trans Knowl Data Eng*, vol. 33, no. 6, pp. 2369–2379, 2021, doi: : 10.1109/TKDE.2019.2954317.

[9] L. Zhang, L. Wang, and L. Zhang, "A novel fuzzy clustering algorithm with human-computer cooperation for incomplete data," *IEEE 5th Int. Conf. Electron. Inf. Emerg. Commun.*, 2015, doi: 10.1109/ICEIEC.2015.7284491.

[10] C. Lu, S. Song, and C. Wu, "Affinity Propagation Clustering with Incomplete Data," *Comput. Intell. Networked Syst. Their Appl.*, vol. 462, 2014, doi: 10.1007/978-3-662-45261-5_25.

[11] S. Alam, M. S. Ayub, S. Arora, and M. A. Khan, "An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity," *Decis. Anal. J.*, vol. 9, 2023, doi: 10.1016/j.dajour.2023.100341.

[12] M. D'Orazio, "Distances with mixed type variables some modified Gower's coefficients," *arXiv*,

2021, doi: https://doi.org/10.48550/arXiv.2101.02481.
[13] D.-T. Dinh, V.-N. Huynh, and S. Sriboonchitta, “Clustering mixed numerical and categorical data with missing values,” *Inf. Sci. (Ny).*, vol. 571, pp. 418–442, 2021, doi: https://doi.org/10.1016/j.ins.2021.04.076.
[14] D. N. Ramadhani, H. S. Rukmi, F. Arif, and A. U. Afifah, “Determining priority location for tourist village development using k-means clustering,” *AIP Conf. Proc.*, vol. 2772, no. 1, 2023, doi: https://doi.org/10.1063/5.0115121.
[15] F. A. Husna, D. Purwitasari, B. A. Sidharta, D. A. Sihombing, A. Fahmi, and M. H. Purnomo, “A Clustering Approach for Mapping Dengue Contingency Plan,” *Sci. J. Informatics*, vol. 9, no. 2, pp. 149–160, Nov. 2022, doi: 10.15294/sji.v9i2.36885.
[16] M. A. Sembiring, “Penerapan Metode Algoritma k-means Clustering untuk Penyebaran Penyakit Demam Berdarah Dengue (DBD),” *J. Sci. Soc. Res.*, vol. 4, no. 3, 2021, doi: 10.54314/jssr.v4i3.712.
[17] S. Oktarian and S. Defit, “Klasterisasi Penentuan Minat Siswa dalam Pemilihan Sekolah Menggunakan Metode Algoritma K-Means Clustering,” *J. Inf. dan Teknol. Vol.*, vol. 2, 2020, doi: 10.37034/jidt.v2i3.65.
[18] Y. Filki, “Jurnal Informatika Ekonomi Bisnis Algoritma K-Means Clustering dalam Memprediksi Penerima Bantuan Langsung Tunai (BLT) Dana Desa,” *J. Inform. Ekon. Bisnis Vol.*, vol. 4, no. 4, pp. 166–171, 2022, doi: 10.37034/infeb.v4i4.166.
[19] I. Rinjani, S. Anwar, and R. Herdiana, “Pengelompokan Daerah Bencana Alam Menggunakan Algoritma K-Means Clustering,” *J. Ilm. Sist. Inf. DAN ILMU Komput.*, vol. 3, no. 1, pp. 35–51, 2023, doi: 10.55606/juisik.v3i1.417.
[20] D. Istiawan and R. Wulandari, “Mapping of Social Vulnerability to Natural Hazards in Geodemographic Analysis Using Fuzzy Geographically Weighted Clustering,” *Sci. J. Informatics*, vol. 10, no. 4, pp. 413–422, 2023, doi: 10.15294/sji.v10i4.47418.
[21] and I. I. D. Chrisinta, I. M. Sumertajaya, “Evaluasi Kinerja Metode Cluster Ensemble Dan Latent Class Clustering Pada Peubah Campuran,” *Indones. J. Stat. Its Appl.*, vol. 4, no. 3, pp. 448–461, 2020, doi: 10.29244/ijsa.v4i3.630.
[22] M. K. A. Azrahwati, M. Nusrang and Z. Rais, “No TitleK-Means Cluster Analysis for Grouping Districts in South Sulawesi Province Based on Village Potential,” *ARRUS J. Math. Appl. Sci.*, vol. 2, no. 2, pp. 73–82, 2022, doi: 10.35877/mathscience739.
[23] BadanPusatStatistik, “Statistik Potensi Desa Indonesia 2021,” *2022*.
[24] Z. Huang, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,” *Data Min Knowl Discov*, vol. 2, no. 3, pp. 283–304, 1998, doi: 10.1023/A:1009769707641.
[25] and B. R. A. H. Foss, M. Markatou, “Distance Metrics and Clustering Methods for Mixed-type Data,” *Int. Stat. Rev.*, vol. 87, no. 1, pp. 80–109, 2019, doi: 10.1111/insr.12274.
[26] L. U. S. Sulastri and U. D. Syafitri, “K-prototypes Algorithm for Clustering Schools Based on The Student Admission Data in IPB University,” *Indones. J. Stat. Its Appl.*, vol. 5, no. 2, pp. 228–242, 2021, doi: 10.29244/ijsa.v5i2p228-242.
[27] Qomariyah and M. U. Siregar, “Comparative Study of K-Means Clustering Algorithm and K-Medoids Clustering in Student Data Clustering,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 7, no. 2, pp. 91–99, 2022, doi: 10.14421/jiska.2022.7.2.91-99.
[28] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate Imputation by Chained Equations in R,” *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011, doi: https://doi.org/10.18637/jss.v045.i03.
[29] A. Kassambara, “Practical Guide To Cluster Analysis in R,” STHDA, 2017.
[30] F. Kara, “Comparison of tree diameter distributions in managed and unmanaged Kazdağı fir forests,” *Silva Balc.*, vol. 22, no. 1, pp. 31–43, 2021, doi: 10.3897/silvabalcanica.22.e58020.
[31] A. Fatikhurrizqi and A. W. Wijayanto, “Comparison Of Regional Cluster Analysis According To Inclusive Development Indicators In Java Island 2018 Between Hierarchical And Partitioning Clustering Strategies,” *JITK (Journal Comput. Sci. Technol.*, vol. 6, no. 2, pp. 143–150, 2021, doi: https://doi.org/10.33480/jitk.v6i2.
[32] N. Bhatia, J. M. Sojan, S. Simonovic, and R. Srivastav, “Role of Cluster Validity Indices in Delineation of Precipitation Regions,” *Water*, vol. 12, no. 5, 2020, doi: 10.3390/w12051372.
[33] A. R. da Silva and C. T. dos S. Dias, “A cophenetic correlation coefficient for Tocher’s method,” *Pesqui Agropecu Bras*, vol. 48, no. 6, pp. 589–596, 2013, doi: 10.1590/S0100-204X2013000600003.
[34] S. D. G. Brock, V. Pihur and S. Datta, “clValid: An R Package for Cluster Validation,” *J Stat Softw*,

vol. 25, no. 4, 2008, doi: 10.18637/jss.v025.i04.
[35] Z. Xu, Y. Yin, H. Chen, H. Xu, and P. Li, "Algorithm for Determining Number of Clusters based on Dichotomy," *2020 Int. Conf. Internet Things IEEE Green Comput. Commun. IEEE Cyber, Phys. Soc. Comput. IEEE Smart Data IEEE Congr. Cybermatics*, 2020, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics50389.2020.00045.
[36] R. Aschenbruck and G. Szepannek, "Cluster Validation for Mixed-Type Data," *Arch. Data Sci. Ser. A*, vol. 6, no. 1, 2020, doi: 10.5445/KSP/1000098011/02.
[37] M. Rezaei and P. Franti, "Set Matching Measures for External Cluster Validity," *IEEE Trans Knowl Data Eng*, vol. 28, no. 8, pp. 2173–2186, 2016, doi: 10.1109/TKDE.2016.2551240.
[38] A. S. K. N. Ismail and K. A. F. A. Samah, "A Comparison Between External and Internal Cluster Validity Indices," *2021 IEEE 11th Int. Conf. Syst. Eng. Technol.*, pp. 229–233, 2021, doi: 10.1109/ICSET53708.2021.9612525.