DOI: https://doi.org/10.34123/jurnalasks.v16i2.789

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 2, December 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

Development of a Hybrid Fuzzy Geographically Weighted K-Prototype Clustering and Genetic Algorithm for Enhanced Spatial Analysis: Application to Rural Development Mapping

Agung Budi Santoso[1*], Arya Candra Kusuma[2], Rani Nooraeni[3]

[1]BPS-Statistics Bima City, Indonesia, [2]BPS-Statistics Indonesia, Jakarta, Indonesia, [3]Politeknik Statistika STIS, Jakarta, Indonesia
**Corresponding Author: E-mail address:* agung.budi@bps.go.id

ARTICLE INFO

Article history:
Received 16 August, 2024
Revised 29 October, 2024
Accepted 9 November, 2024
Published 31 December, 2024

Keywords:
Clustering; Geographically Weighted Cluster; Mixed-type data; Village Development; Genetic Algorithm

Abstract

Introduction/Main Objectives: Clustering methods are crucial for geodemographic analysis (GDA) as they enable a more accurate and distinct characterization of a region. This process facilitates the creation of socio-economic policies and contributes to the overall advancement of the region. **Background Problems:** The fuzzy geographically weighted clustering (FGWC) method, which is a GDA technique, primarily handles numerical data and is prone to being stuck in local optima. **Novelty:** This study proposed two novel clustering methodologies: fuzzy geographically weighted k-prototypes (FKP-GW) and its hybrid clustering model, which combines genetic algorithm-based optimization (GA-FKP-GW). **Research Methods:** This research conduct simulation study comparing two of the proposed clustering method. For the empirical application, this study applied clustering technique using the official Village Potential Survey of Temanggung, Indonesia. **Finding/Results:** The evaluation results of experiments conducted on simulated data and study cases indicate that the proposed method yields distinct clustering results compared to the previous method while being comparably efficient. The empirical application identifies four distinct groups from the clustered villages, each displaying unique characteristics. The results of our research have the potential to benefit the development of the GDA method and assist the local government in formulating more effective development policies.

1. Introduction

Clustering methods are currently crucial for analyzing an area. Employing appropriate clustering techniques enables a more precise and distinctive description of the characteristics of an area and distinguishes between different groups. Hence, it is imperative to conduct a thorough investigation into selecting appropriate techniques to achieve optimal clustering outcomes. Geo-demographic analysis (GDA) is a commonly employed method for analyzing a specific area. Geodemography is the study and examination of the attributes or trends of individuals or inhabitants, specifically in relation to their geographical location [1]. The fuzzy geographically weighted clustering (FGWC) method is an appropriate clustering technique for geo-demographic analysis (GDA) as it takes into account the spatial variations in population size and distance between regions [2]. However, the exclusive restriction of the current FGWC method to numerical data presents a challenge when clustering mixed-type geographic data.

The primary goal of GDA is to create clusters or groups based on socioeconomic status within a particular region. This process aids in the formulation of socio-economic policies and the overall development of the region [3]. Geographic data analysis uses Fuzzy Geographically Weighted Clustering (FGWC), a highly effective clustering method. Mason and Jacobson [4] developed this method by combining the fuzzy C-means (FCM) method with the neighborhood effect (NE). This integration allows the method to give greater consideration to the geographical impact of each data element [4]. However, it is similar to the previous approach, known as FCM, which is exclusively suitable for numerical data [5], [6], [7], [8]. This approach typically employs the Euclidean distance as its cost function, which is suitable for measuring distances between numerical data but not for categorical data [9], [10], [11].

The K-prototype algorithm, developed by Huang [12], manages mixed-type data clustering efficiently. This algorithm integrates the K-means algorithm with K-modes [12]. The K-Prototype is a clustering technique that makes use of partial clustering [13]. For clustering mixed data types, the K-prototype algorithm is an efficient and scalable algorithm [12]. Nevertheless, the K-Prototype method still has limitations when it comes to determining the centroid for categorical data [14]. The centroid determination for categorical attributes in the K-Prototype (KP) algorithm still depends on the mode value of those attributes. Because the mode value is used as the centroid, it is unable to provide an accurate description of objects [11], [13].

In 2012, Ji, et al. [14] introduced a technique called fuzzy K-prototype (FKP) that combines fuzzy clustering or soft clustering with the K-prototype algorithm. Unlike the previous method, FKP does not use the mode value as the centroid. Instead, it incorporates a fuzzy centroid into the algorithm [14].The Fuzzy K-Prototype method, a partition-based clustering technique, also encounters issues with the randomly selected initialization centroid. This will result in the method being stuck in the local optimum solution [12], [15].

To address the issue of a local optimum solution, one effective approach is to employ the metaheuristic method [15], [16], [17], [18], [19], [20], [21]. A metaheuristic method that can be utilized is the genetic algorithm, which operates based on Charles Darwin's theory of natural selection [16], [22], [23]. A genetic algorithm is a metaheuristic algorithm that has undergone extensive development and modification to optimize its performance by effectively balancing the trade-off between exploitation and exploration [16]. This algorithm is commonly referred to as a search algorithm that involves fewer mathematical computations compared to other algorithms [24].

Several previous studies have improved the methodology for geodemographic clustering techniques and clustering mixed data. From the FGWC development that consider spatial aspect but limited to numerical data, prior study have improved optimization by using numerous metaheuristic optimization, namely Artificial Bee Colony (ABC) [25], Ant Colony Optimization (ACO) [26], Gravitation Search Algorithm (GSA) [27], Intelligent Firefly Optimization (IFO) [19], and Chaotic Flower Pollination Algorithm (CFPA) [28]. From the clustering mixed data, previous research also developed the K-Protype method. Because the determination of the initial centroid has a great influence on the clustering solution, Nooraeni et al. were conducted to optimize the cluster center initialization using the K-Prototype method with the GA algorithm (GA-KP) [29], [30]. Nooraeni et al. have also enhanced the GA-KP algorithm by resolving the issue of using the mode value as a centroid for categorical data, as well as addressing the problem of finding the local optimum solution by utilizing a fuzzy centroid [31]. However, the development of a hybrid between FGWC and FKP methods using genetic algorithm optimization remains unexplored.

Therefore, this research will develop a hybrid method between FGWC and FKP to overcome the weakness of mixed-type geographic data and optimizing them using a Genetic Algorithm. This integrated approach enables the analysis of mixed data clusters with spatial considerations, effectively overcoming local optima. Furthermore, this study is organized as follows: The Methodology section explains the proposed clustering methods, namely FKP-GW and GA-FKP-GW. To assess the efficiency of the proposed clustering, the methodology section also explains the evaluation through simulation data and empirical data. The Results section explains the application of the FKP-GW and GA-FKP-GW clustering methods. Finally, the Conclusion section summarizes the results of this research, including the development of geodemographic clustering methods and their application for regional development.

2. Material and Methods

2.1 Data

The study utilizes both simulated and empirical data to evaluate the hybrid fuzzy geographically weighted K-prototype clustering and genetic algorithm method. Simulated data are generated to assess the algorithm's performance in a controlled environment, allowing for a clear examination of clustering behavior. Real-world data, on the other hand, is sourced from Indonesia's Village Development Index (VDI) [32] and provides practical context by applying the method to rural development mapping.

2.1.1 Simulated Data

Two sets of simulated data are generated to test the clustering algorithm. The first set consists of random data with seven attributes, including four categorical and three numerical variables. The second set expands on this with eight attributes (four categorical and four numerical). Both of these sets of simulated data are generated with 50 observations and 100 observations, resulting in 4 scenarios of simulation. These simulated datasets allow for an exploration of the algorithm's ability to manage and cluster mixed data types, testing its robustness in assigning clusters effectively under varied attribute structures and sample sizes.

2.1.2 Empirical Data

For implementing the algorithm in the real-world case, the proposed method will be applied to clustering the villages of Temanggung Regency based on indicators of the village development index (VDI). The BPS-Statistic Indonesia aims to find out the potential of each village based on several indicators of socio-economic development. In this study, the term "village" includes the nagari, Transmigration Settlement Unit, and Entity of Transmigration Settlement, which are still fostered by the relevant ministries and government agencies of Indonesia. The Village Development Index (VDI), also known as Indeks Pembangunan Desa (IPD), describes the development progress of a village at a specific time [32]. The VDI calculation is obtained from the results of the Village Potential Statistics (PODES) data collection. The calculation of VDI from PODES 2018 data involves 5 dimensions and 42 indicators that document the availability of infrastructure and service accessibility [33].

The BPS publication on the Village Development Index 2018 reveals that there are 5,606 independent villages, 55,369 developing villages, and 14,461 underdeveloped villages [32]. The publication also explained that the island of Java—Bali became the island with the highest VDI. Despite being the island with the highest VDI, Central Java, one of the provinces, has an VDI that is lower than the average VDI for the island of Java-Bali. Temanggung Regency, one of Central Java's regencies, is actively involved in the design of the Village Development Work Plan (RKPD). It is proven by his achievement of winning a second-place award in National District Level Development 2019. However, the VDI value of Temanggung Regency in 2018 was 66.05, indicating that it remains below the average VDI of Central Java Province. This indicates that Temanggung Regency has the potential to enhance its development plan by identifying the potential of each village within its territory. Thus, clustering was carried out for mapping villages in the Temanggung Regency area to determine the characteristics of each village so that development planning could be more targeted.

The indicators of VDI were collected from PODES 2018, which consists of 289 villages with 88 numerical attributes and 94 categorical attributes. The study then uses the population of each village as the geographic effect variable, and the centroid point of the shapefile for each village as the distance between village areas.

2.1.3 Data Preprocessing

The beginning step of the pre-processing. For the numerical attributes $\boldsymbol{x}^r$, standardized using the min–max method. For the function min-max that used in this research can be written in equation (1).

$$x_{ij}'^{\,r} = \frac{x_{ij}^r - \min(x_j^r)}{\max(x_j^r) - \min(x_j^r)} \tag{1}$$

Where $x_{ij}'^{\,r}$ and x_{ij}^r are the new numerical attribute data standardized and the real old numerical attribute data, min and max x_j^r are the minimum and maximum value in j-th numerical attributes. Then, change the numeric value to be categorical or called discretized as much as the specified number of T intervals. This approach also used by Ji., et al. [14] to develop fuzzy k-prototype algorithm.

This process pre-processing is carried out for all data used in the FKP, FKP-GW, and GA-FKP-GW. For variables distance in geographically weighted compute the distance of each area observation are using function spDists() from package "sp" in Rstudio [34], which can return the result as matrix distance every area observation.

2.2 Clustering Methods

The clustering methods are divided into two types, hierarchical clustering and partitional clustering. In hierarchical clustering, data is grouped hierarchical or stratified. While the partitional clustering methods data is grouped into several clusters without any hierarchical structure [35], [36]. So, the number of clusters must be determined from the beginning step of processing. Determining the number of clusters can be done in various ways, drawing a scree plot to see the significant decrease of the total cost function for each cluster can be one alternative of determination [37], [38].

2.2.1 Fuzzy Geographically Weighted Clustering

FGWC is a method developed by Mason & Jacobson [4] that integrates fuzzy clustering with Neighborhood Effects by taking into account the distance between regions and their populations which makes them more sensitive to neighboring effects and the results clusters are more geographically aware. The FGWC method performs a new calculation for its membership value in each iteration using the equation (2) [4].

$$\boldsymbol{\mu}_i' = \alpha \boldsymbol{\mu}_i + \beta \frac{1}{A} \sum_{j=1}^{n} w_{ij} \boldsymbol{\mu}_j \tag{2}$$

Where $\boldsymbol{\mu}_i'$ is the new membership degree value of the i-th region, while $\boldsymbol{\mu}_i$ is the old membership degree value of the i-th region and $\boldsymbol{\mu}_j$ is the value of the j-th region old membership degree. α and β are the weights of the old membership value and new membership of other objects whose sums are equal by 1. A is the value that ensures the weighting of the membership value in the range of zero and one [0,1]. w_{ij} is the weights between the two geographic areas can be written in equation (3).

$$w_{ij} = \frac{(m_i m_j)^b}{d_{ij}^a} \tag{3}$$

With the m_i and m_j is the number of population in the region of the i and j, while d_{ij} is the distance between regions i and j. a is the magnitude of the interaction effect of the distance between regions and b is the magnitude of the population interaction between the two regions. Both parameters can be determined by the researcher.

2.2.2 K-Prototype

This method is a partitional clustering developed from the K-means method, which previously could only be used for numeric data types but is still maintaining its efficiency. The KP algorithm is still classified as hard clustering with each object can be grouped into one cluster only. The cost function for mixed-type data can be written in equation (4) [14].

$$\begin{aligned} E &= \sum_{l=1}^{k} (E_l^r + E_l^c) \\ &= \sum_{l=1}^{k} \left(\sum_{i=1}^{n} \mu_{il} \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{i=1}^{n} \mu_{il} \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c) \right) \end{aligned} \tag{4}$$

Where E_l^r and E_l^c is total cost from every numeric and categorical attribute from cluster l, respectively. x_{ij}^r is the value of object i in j-th numeric attributes set, x_{ij}^c is value of object i in j-th categorical attributes set, and γ_l is a weight for categorical attribute in the cluster l. q_{lj}^r and q_{lj}^c is the prototype or

centroid of every numeric and categorical attribute, respectively. This algorithm cover in function kproto() from package "clustMixType" [39] in Rstudio.

2.2.3 Fuzzy K-Prototype

Ji et al. [14] developed the Fuzzy K-Prototype clustering method for mixed data, utilizing the KP algorithm as their foundation. This method uses fuzzy clustering to find the centroid of the numeric and categorical attributes of the cluster. It also uses new dissimilarity measures that look at the significance level of each numeric attribute and the distance between the values of the categorical attributes. The cost function equation for Fuzzy K-Prototype can be written in equation (5) [14].

$$E(\boldsymbol{\mu}, \mathbf{Q}) = \sum_{l=1}^{k}\left(\sum_{i=1}^{n}\mu_{il}^{\alpha}\left(\sum_{j=1}^{m_r}\left(\omega_l\left(x_{ij}^{r} - v_{jl}^{r}\right)\right)^2 + \sum_{j=1}^{m_c}\varphi\left(x_{ij}^{c}, \tilde{v}_{jl}^{c}\right)^2\right)\right) \tag{5}$$

Notation $\boldsymbol{\mu}$ represents the matrix of cluster degree membership and $\mathbf{Q}$ represents the matrix of cluster centroid. Where centroid for numeric value v_{jl}^{r} and fuzzy centroid categorical value $\tilde{v}_{jl}^{c}$ can written in equation (6) and equation (7), respectively [14].

$$v_{jl}^{r} = \frac{\sum_{i=1}^{n}(\mu_{il})^{\alpha}x_{ij}^{r}}{\sum_{i=1}^{n}(\mu_{il})^{\alpha}} \tag{6}$$

$$\tilde{v}_{jl}^{c} = \frac{a_{jl}^{1}}{\omega_{jl}^{1}} + \frac{a_{jl}^{2}}{\omega_{jl}^{2}} + \cdots + \frac{a_{jl}^{\kappa}}{\omega_{jl}^{\kappa}} + \cdots + \frac{a_{jl}^{t}}{\omega_{jl}^{t}} \tag{7}$$

Fuzzy matrix partition then updated using equation (8).

$$\mu_{il} = \frac{1}{\sum_{z=1}^{k}\left(\frac{d(x_i, q_l)}{d(x_i, q_z)}\right)^{\frac{1}{\alpha-1}}} \tag{8}$$

Where $\sum_{j=1}^{m_r}(\omega_l(x_{ij}^{r} - v_{jl}^{r}))^2$ is the calculation of the distance i-th observation to the l-th cluster centroid for numeric attributes, while $\sum_{j=1}^{m_c}\varphi(x_{ij}^{c}, \tilde{v}_{jl}^{c})^2$ is the calculation of the distance i-th observation to the l-th cluster centroid for the categorical attribute [14]. μ_{il}^{α} is a matrix membership value for i-th observation in the l-th cluster, with parameter fuzziness α whose value is determined by the researcher, while ω_l is the significance value of each numeric attribute in the l-th cluster. The FKP method used in this study employs the following algorithms [14]:

1. Determine the maximum iteration, number of clusters k, values of fuzziness parameter (α), and threshold (ε).
2. Initialization value of partition matrix membership by generating random value.
3. Calculate the centroid for numeric attributes using equation (6) and centroid for categorical attributes using equation (7).
4. Calculate the distance between observation to the centroid.
5. Update the fuzzy matrix partition membership using equation (8).
6. If the difference between the previous cost function is less than the threshold or has reached the maximum iteration, then stop process clustering. If not, return to step 3.

2.2.4 The Proposed Method: Fuzzy K-Prototype Geographically Weighted (FKP-GW)

This research develops a hybrid method to cluster mixed-type geographic data while considering spatial effects. The proposed method, Fuzzy K-Prototype Geographically Weighted (FKP-GW), uses the following algorithm:

1. The algorithm requires the input of several parameters, including data, the number of clusters k, the fuzziness coefficient m, the maximum iteration, and the threshold ε. Additionally, it requires the input of several geographically weighted parameters, including the matrix population, the matrix distance for each observation, α, β, a, and b.
2. Generate partition matrix membership according to the predetermined number of clusters k with random values.
3. Calculate the centroid for numerical and categorical attributes using equations (6) and (7).
4. Determine the observation's distance to the obtained centroid, then add the distances for both numeric and categorical attributes.
5. Using equation (8), update the fuzzy matrix partition membership based on the previously obtained matrix distance.
6. Modify the previous matrix membership with geographical weights according to equation (2).
7. If the difference between the previous cost function is less than the threshold or has reached the maximum iteration, then stop process clustering. If not, return to step 3.

2.2.5 The Proposed Method: Genetic Fuzzy K-Prototype Geographically Weighted (GA-FKP-GW)

This research proposes the Genetic Fuzzy K-Prototype Geographically Weighted (GA-FKP-GW) as the next hybrid method. The Genetic Algorithm is a search algorithm that incorporates Charles Darwin's theory of natural selection [16], [22], [23]. It appears to operate similarly to a natural selection process, where the individuals that survive are those that can endure throughout the evolutionary process. In this method, there are several terms, such as chromosome, which consists of the candidate best solution for matrix membership; genes, which is the value of matrix membership; individual, population, generation, selection, crossover, mutation, and elitism. The proposed method, known as GA-FKP-GW, uses the following algorithm:

1. Input Parameters: Parameters that need to be inputted for GA-FKP-GW are mutation rate, maximum generation, and numbers of population, in addition to several parameters that are used in the FKP-GW algorithm,
2. Initialization individual as much as the numbers of population that contain candidate chromosomes; this chromosome contains candidate matrix membership. The pseudocode used in initialization is:

> **For** i = 1 to n **do**
>
> Generate random numbers from [0,1];
>
> For the i-th point of chromosome;
>
> Calculate
>
> **End for**

Source: Gan, et al. [17]

3. Evaluation of the fitness value of each individual by counting the cost function and evaluating it using equation (9).

$$f = \frac{1}{(h + a)} \tag{9}$$

The lowest value of the cost function shows this individual has a good potential solution.

4. Create new populations by repeating a few steps for as many as maximum generations until the new population produces the most optimal solution:
   a. Selection: The selection process involves the selection of chromosomes to facilitate the subsequent processes of crossovers and mutations. The mutation method used by the researcher is the roulette wheel. This method is done by calculating the probability value of each chromosome based on its fitness value with equation (10).

$$P_i = \frac{f_i}{f_{\text{total}}} \; i = 1,2, \ldots n_{\text{population}} \tag{10}$$

After obtaining the probability value, compute the cumulative probability value using the obtained value. To select individuals, generate a random number R in interval [0,1]. Individuals will be selected if the cumulative probability $\geq R$.

b. Crossover: This crossover process aims to enhance the diversity of chromosomes within a population. In this study, the researcher employed a one-step fuzzy K-prototype, which was weighted geographically. The pseudocode used in Crossover is:

For i=1 to N **do**

Let $\boldsymbol{\mu}_i$ be the fuzzy membership represented by s_i;

Obtain the new set of cluster centers $\boldsymbol{q}_i$ given $\boldsymbol{\mu}_i$ according to formula no 7 & 8;

Obtain the fuzzy membership $\boldsymbol{\mu}_i$ given $\boldsymbol{q}_i$ according to formula no 5;

Replace s_t with the chromosome representing $\boldsymbol{\mu}_i$;

End for

Source: Gan, et al. [17]

c. Mutation: This process involves altering one or more genes within a chromosome, resulting in the creation of a new chromosome and avoiding the issue of a local optimum. Mutations can occur when the probability of a gene being mutated is less than the probability of a mutation rate in a chromosome, and then the gene will be replaced or mutated. The pseudocode used in Mutation is:

For t=1 to N **do**

Let $(a_1, a_2, \cdots, a_{n,k})$ denote the chromosome s_t;

For i=1 to N **do**

Generate k random real number $v \in [0, 1]$;

If $v \leq pm$ **then**

Generate k random numbers $v_{i1}, v_{i2}, \ldots, v_{ik}$ from [0,1] for the i-th point of chromosome;

Replace

End if

End for

Source: Gan, et al. [17]

d. Elitism: This process stores the chromosomes with the best fitness value, with the aim of causing these chromosomes to decrease in their fitness values during the crossover and mutation processes.

5. The best candidate individual or chromosomes that give the best solution will be used as a matrix membership in algorithm FKP-GW for providing a good cluster result.

2.2.6. Indicators Evaluation Clustering Methods

To evaluate the clustering result by using some clustering methods, there are some indicators that can be used to evaluate it. The indicators that can be used are the cost function of its methods, index PC, index SC, and index CVC. By comparing these indicators, it can be used to know how good clustering methods provide clustering results. Several indicators were employed in this study.

1. Partition Coefficient (PC)
   The PC index is an index that measures the amount of overlap between clusters. The PC index value is measured using equation (11).

$$PC = \frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n}\mu_{il}^{2} \tag{11}$$

   The greater the PC index value, the better the cluster result.

2. Classification Entropy (CE)
   The CE index is an index that measures the fuzziness between clusters. The CE index value is measured using equation (12).

$$CE = -\frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n}\mu_{il}\log(\mu_{il}) \tag{12}$$

   The smaller the CE index value, the better the cluster results.

3. Categorical Variance Criterion (CVC)
   CVC is a combination method Category Utility (CU) for categorical attributes and measurement variance σ^2 for numeric attributes. The CVC formula can be written in equation (13) as follows [40].

$$CVC = \frac{CU}{1+\sigma^2} \tag{13}$$

   Where:

$$CU = \sum_{l=1}^{k}\left(\frac{|C_l|}{N}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[P(X_j = V_{ij}|C_l)^2 - P(X_j = V_{ij})^2\right]\right)$$

$$\sigma^2 = \sum_{l=1}^{k}\frac{1}{|C_l|}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(V_{i,j}^{l} - V_{i,avg}^{l}\right)^2$$

   The higher the CVC value, the better the cluster results

3. Results and Discussion

3.1 Comparison Previous Clustering Method with Proposed Clustering Method (FKP-GW and GA-FKP-GW)

The comparison between the proposed clustering method and previous clustering methods is done using simulated data. In this research, there are two types of simulation data, namely data type 1 and data type 2. Data type 1 comprises seven attributes, while data type 2 comprises eight. Each dataset

contains 50 and 100 observations. Some parameters were determined to evaluate the results of the proposed method and compare them with the previous method. Table 1 provides details on the number cluster, parameter fuzziness, parameter alpha, and beta for each data set used in this research.

Table 1. Number of cluster and Parameter Fuzziness for Simulated Data

Data		Number of Cluster (k)	Fuzziness Parameter	Geographically Weighted Parameters	
				Alpha (α)	Beta (β)
(1)		(2)	(3)	(4)	(5)
Type 1	$N = 50$	3	1.3	0.9	0.8
(7 attributes)	$N = 100$	3	1.9	0.9	0.8
Type 2	$N = 50$	3	1.3	0.9	0.8
(8 attributes)	$N = 100$	4	1.4	0.9	0.8

To determine the number of clusters on each dataset, this research is considering a scree plot of decreasing the value of the total cost function for each cluster. In this study, the maximum iteration is 100, and the threshold is 0.00005. The genetic parameters used in this research include a maximum generation of 20, a population of 20, and a mutation rate of 0.00005. Table 2 displays the value of the cost function as the basis for the clustering evaluation, with the smallest value yielding the best result.

Table 2. Cost Function Values of Each Clustering Methods for Simulated Data

Data		Cost Function Value of each Clustering Methods				
		KP	FKP	GA-FKP	FKP-GW	GA-FKP-GW
(1)		(2)	(3)	(4)	(5)	(6)
Type 1	$N = 50$	10.86	0.915	0.915	0.990	0.984
(7 attributes)	$N = 100$	17.18	1.403	1.403	1.531	1.531
Type 2	$N = 50$	13.30	0.944	0.944	1.011	1.006
(8 attributes)	$N = 100$	20.19	1.556	1.556	1.675	1.672

Table 2 compares the cost function values of five clustering methods applied to simulated data with varying sample sizes and attributes. It is clear that the proposed methods, FKP-GW and GA-FKP-GW, consistently produce lower cost function values than the standard KP method. This shows the benefit of adding geographical weighting and fuzziness to the clustering process. Although GA-FKP achieves the smallest cost function values across all data types, the proposed methods perform competitively, particularly for larger datasets. The cost function values show minimal differences, indicating that while GA-FKP excels in cost reduction, FKP-GW and GA-FKP-GW still identify significant spatial features that could aid in the clustering of geographically dispersed data.

For smaller datasets ($N = 50$), the proposed methods show slight improvements in cost function compared to the traditional methods, and for larger datasets ($N = 100$), all fuzzy-based methods tend to converge in performance. Despite the marginally higher cost values in some cases, the spatially weighted methods offer the potential for more nuanced insights due to their ability to incorporate geographical variability. This highlights the trade-off between optimizing a cost function and obtaining clusters that reflect the spatial context of the data, making FKP-GW and GA-FKP-GW valuable for applications where spatial relationships are critical.

3.2 Implementing GA-FKP and GA-FKP-GW For Village Development Data

Before proceeding with the clustering process, it is essential to first determine the appropriate number of clusters and the level of fuzziness. Choosing the correct number of clusters guarantees the model captures the data's structure without overfitting or underfitting. Additionally, defining the degree of fuzziness is crucial in a fuzzy clustering approach, as it allows for flexible membership of data points

across clusters, better reflecting the complexity and ambiguity in real-world data. The determination of the number of clusters used is done using the scree plot shown in the Figure 1.

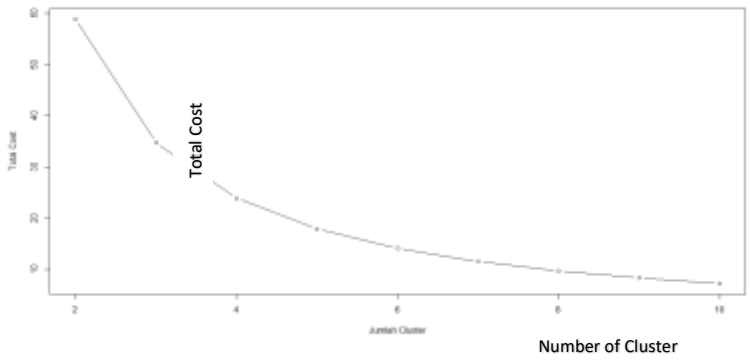

Figure 1. Scatter plot total cost function for each number of clusters

The scree plot suggests that four clusters are a suitable choice for the analysis using the GA-FKP-GW method. The "elbow" at 4 clusters indicates a significant reduction in the cost function, after which the improvements become marginal. This gives four clusters a reasonable balance between complexity and capturing meaningful distinctions in the data. Given that the Village Development Index (VDI) traditionally used 3 categories, expanding to 4 clusters allows for greater nuance in representing village development patterns while maintaining interpretability within the cluster formed.

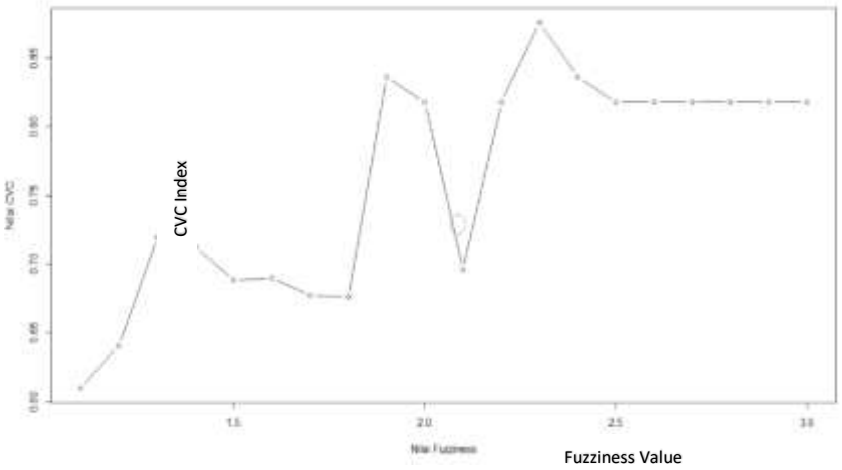

Figure 2. Plot index CVC for every parameter fuzziness

Next, determination of the fuzziness hyperparameter is conducted using the CVC plot in Figure 2. The CVC plot reveals the optimal value of fuzziness for the fuzzy K-prototype clustering method. Observing the graph, it is clear that as the fuzziness parameter increases, the CVC values fluctuate, indicating varying cluster validity across different levels of fuzziness. Approximately 2.4 is the optimal value for producing the best balance between cluster compactness and separation, resulting in the highest CVC value. Consequently, a fuzziness value of 2.4 provides the most appropriate configuration for capturing the characteristics of the data set under analysis.

The hybrid clustering method proposed, GA-FKP-GW, has been successfully built. This proposed method can be used for clustering mixed-type data and adding the spatial effect to it. Table 3 displays the results of the membership matrix for both the previous clustering method (GA-FKP) and the proposed method (GA-FKP-GW).

Table 3. Matrix Membership Degree from Previous Method Clustering Method (GA-FKP) and Proposed Method (GA-FKP-GW)

Clustering Method	Observation	Membership Degree of Each Cluster			
		Cluster 1	Cluster 2	Cluster 3	Cluster 4
(1)	(2)	(3)	(4)	(5)	(6)
GA-FKP	1	0.2408	0.2404	0.2537	0.2648
	2	0.2456	0.2454	0.2518	0.2570
	3	0.2443	0.2439	0.2524	0.2592
	4	0.2444	0.2441	0.2524	0.2588
	…	…	…	…	…
GA-FKP-GW	1	0.2435	0.2432	0.2526	0.2606
	2	0.2471	0.2470	0.2511	0.2546
	3	0.2460	0.2458	0.2516	0.2564
	4	0.2462	0.2460	0.2516	0.2560
	…	…	…	…	…

Table 3 presents the performance of the GA-FKP and GA-FKP-GW methods in clustering mixed-type data, demonstrating the incorporation of spatial effects in the latter approach. The differences in membership degrees between the two methods highlight the influence of the geographically weighted component in GA-FKP-GW. By applying this spatial weighting, the geographic location of each observation influences its likelihood of belonging to a specific cluster, leading to shifts in membership values compared to the traditional GA-FKP method.

These subtle differences in membership degrees reflect the intended design of GA-FKP-GW, which adjusts for spatial heterogeneity across the data. This effect is especially important when mapping rural development because the results of the clustering now take into account local geographic factors. This could lead to clusters that better reflect the data's underlying spatial structure and characteristics. To sum up, Table 3 shows how well GA-FKP-GW works at improving the clustering process by adding spatial factors. This makes the analysis more relevant for data that is spread out geographically.

3.3 Profiling villages in Temanggung Regency Based on GA-FKP-GW Algorithm

The proposed method GA-FKP-GW is successfully implemented to analyze clustering data study cases using indicators of the Village Development Index (VDI) of Temanggung Regency 2018. This study cases VDI indicators were included in PODES. The previous section provides explanations for all the parameters used in this clustering. The result from the analysis cluster to data study case using GA-FKP-GW, where from 88 numeric and 94 categorical attributes with 289 villages included, resulted in 4 clusters. To see the detail of mapping distribution villages grouped by GA-FKP-GW and VDI, see Figure 3.

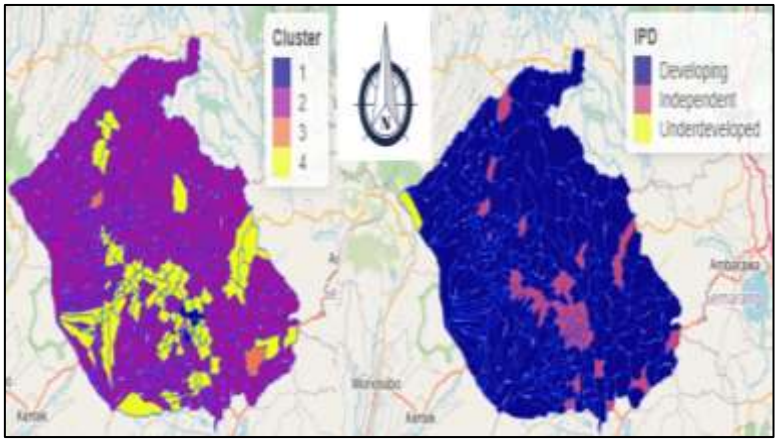

Figure 3. Mapping villages grouped by GA-FKP-GW and VDI

Figure 3 displays a mapping of the distribution of villages according to their cluster and VDI. The villages included in cluster 1 are located and gathered in the central area of Temanggung Regency. Meanwhile, the villages included in cluster 2 are seen scattered throughout the region. Cluster 3's villages appear dispersed, not concentrated in a single adjacent area. Despite the presence of several scattered villages on the area's edge, the villages included in cluster 4 appear to gather in the center of the Temanggung Regency area. The comparison of mapping villages by cluster and VDI original category reveals that certain developing villages share the same color with cluster 2, while independent villages align with cluster 4. Table 4 provides further details.

Table 4. Total villages grouped by Cluster and VDI

VDI	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
(1)	(2)	(3)	(4)	(5)	(6)
Independent	4	10	2	25	41
Developing	0	192	1	53	246
Underdeveloped	0	2	0	0	2
Total	4	204	3	78	289

In order to identify the distinct features of each cluster, researchers chose to examine the attributes that significantly differed between them. A statistical test is used to test whether there is a difference in average between clusters of each attribute [41], [42]. For numerical attributes, researchers used the statistic test one-way Anova. Every attribute was significant, then compute the average and interpret it using descriptive analysis. Some attributes are grouped into a few variables that were included in the calculation of VDI. For detailed profiling villages, each cluster in every variable VDI can be seen below.

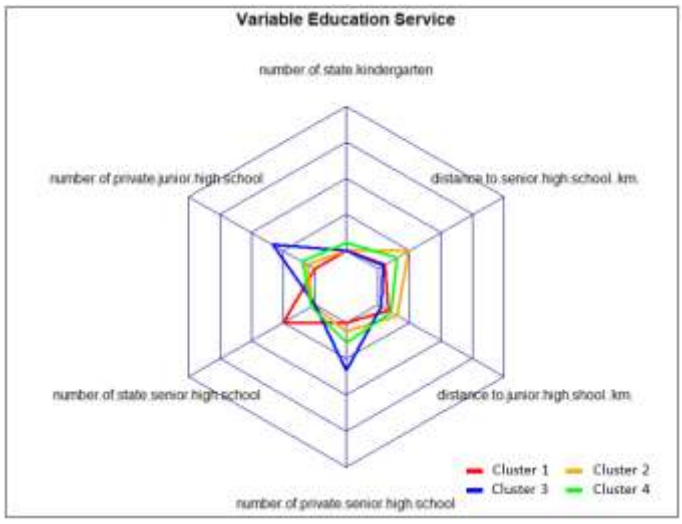

Figure 4. Spider plot variable education service

Figure 4 displays the characteristics of each cluster within the education service variable. Figure 4 shows the average indicators on the education service variable. In this variable, cluster 1 becomes the cluster with the highest average number of state senior high schools, the cluster 3 becomes the cluster with the highest average number of private junior high schools and state high schools, and the cluster 4 becomes the cluster with the highest average number of kindergartens. Meanwhile, Cluster 2 is the one with the farthest average distance to educational services among other clusters.

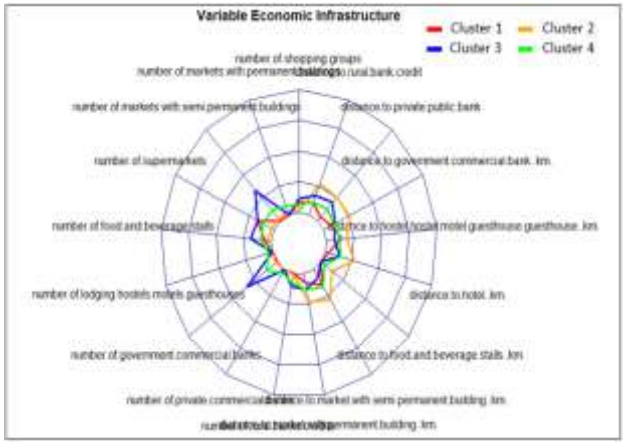

Figure 5. Spider plot variable economic infrastructure

Figure 5 displays the characteristics of each cluster in terms of variable economic infrastructure. Cluster 3 stands out as the cluster with the highest average number of infrastructures, with Cluster 4 and Cluster 1 following closely behind. In contrast to other clusters, Cluster 2 has the longest average distance to the closest economic infrastructure.

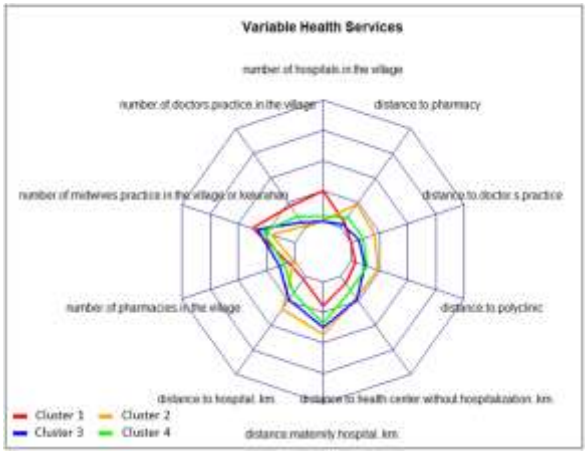

Figure 6. Spider plot variable health services

Figure 6 displays the characteristics of each cluster in terms of variable health services. Compared to other clusters, cluster 1 has the highest average number of health services and the closest average distance to isolated health services, while cluster 3 has the highest average number of pharmacies in the village. Meanwhile, cluster 2 is the cluster with the least and farthest average number and distance to health services compared to other clusters.

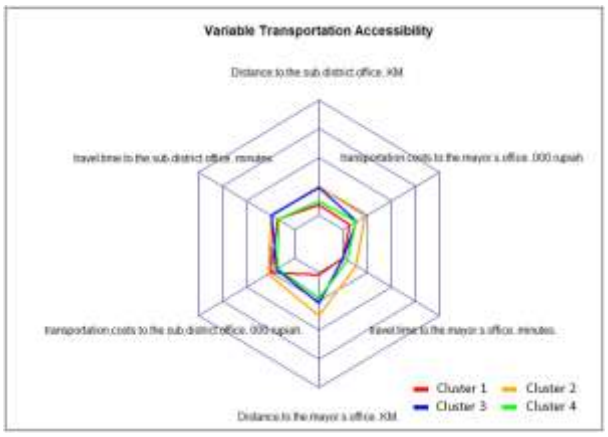

Figure 7. Spider plot variable transportation accessibility

Figure 7 illustrates the characteristics of each cluster in the transportation accessibility variable. Compared to other clusters with the least transportation time and cost, cluster 1 in the transportation accessibility variable has the closest average distance to the sub-district office and district head, while cluster 2 has the average distance and time and typically travels to the district and district offices with the highest costs.

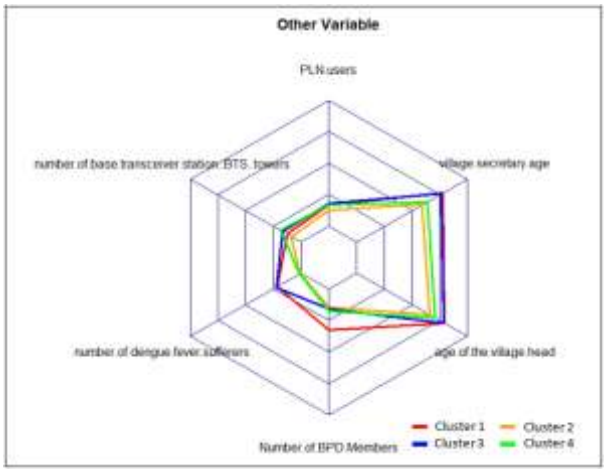

Figure 8. Spider plot of other variable

Figure 8 is a spider plot that displays the average values of other variables for each cluster. In this plot, some variables were combined, including energy infrastructure, communication and information infrastructure, public health, independence, and quality of human resources. The plot image reveals that cluster 3 has the highest average indicator of PLN users, while cluster 2 has the lowest. For the number of BTS towers, cluster 4 is the cluster with the highest average number of towers, while cluster 2 is the cluster with the fewest average BTS towers. For indicators of the number of patients with dengue fever, cluster 3 becomes the cluster with the most sufferers on average, while cluster 2 becomes the average patient with the least. For indicators of the number of BPD members, cluster 1 is the cluster with the highest average number of members, while cluster 2 is the cluster with the least average number of members. For the HR quality variable, cluster 1 is the cluster with the youngest average age of the village head and secretary and cluster 2 with the oldest average age.

For categorical attributes, researchers count the frequency of every categorical attribute within each cluster that is significant and has a different frequency of each cluster using the statistic test chi-square. By counting the frequency of each cluster, the researchers discovered that cluster 2 differs from other clusters in that it has access to the nearest infrastructure, which is easily accessible by road. In addition, cluster 2 was found to have fewer sports facilities than other clusters.

The descriptive analysis of numerical and categorical attributes, which is significant, shows that cluster 1 is the cluster with the best average health service but has the least average number of economic infrastructure. On the other hand, Cluster 3 has the best average economic infrastructure and educational services. Cluster 4 is the cluster with the most stable and above average infrastructure facilities and

infrastructure among other clusters. Meanwhile, cluster 2 is the cluster with the average number of facilities and infrastructure that is less than the other clusters, and the average distance to other infrastructure is the farthest. Cluster 2 also features easy road access to the nearest infrastructure, and its villages have fewer sports facilities than those in other clusters.

After the descriptive analysis, the researcher calculated the distribution of villages in each cluster based on the subdistrict.Table 5 displays the distribution table, which divides the number of villages in each cluster into several sub-districts in Temanggung Regency.

Table 5. Distribution villages each cluster by sub-district

Sub-district	Cluster 1	Cluster 2	Cluster 3	Cluster 4
(1)	(2)	(3)	(4)	(5)
Bansari	0	12 (92.3%)	0	1 (7.7%)
Bejen	0	12 (85.7%)	0	2 (14.3%)
Bulu	0	13 (68.4%)	0	6 (31.6%)
Candiroto	0	11 (78.6%)	1 (7.1%)	2 (14.3%)
Gemawang	0	9 (90%)	0	1 (10%)
Jumo	0	13 (100%)	0	0
Kaloran	0	9 (64.3%)	0	5 (35.7%)
Kandangan	0	16 (100%)	0	0
Kedu	0	7 (50%)	0	7 (50%)
Kledung	0	6 (46.2%)	0	7 (53.8%)
Kranggan	0	12 (92.3%)	0	1 (7.7%)
Ngadirejo	0	15 (75.0%)	0	5 (25%)
Parakan	0	3 (18.8%)	0	13 (81.3%)
Pringsurat	0	11 (78.6%)	1 (7.1%)	2 (14.3%)
Selopampang	0	7 (58.3%)	0	5 (41.7%)
Temanggung	4 (16%)	8 (32%)	1 (4%)	12 (48%)
Tembarak	0	10 (76.9%)	0	3 (23.1%)
Tlogomulyo	0	6 (50%)	0	6 (50%)
Tretep	0	11 (100%)	0	0
Wonoboyo	0	13 (100%)	0	0
Total	**4 (1.4%)**	**204 (70.6%)**	**3 (1%)**	**78 (27%)**

Table 5 is presented in order to find out how well the local government develops or ensures equal distribution of villages in each subdistrict. Almost all sub-districts group several villages into distinct clusters, indicating that the characteristics of the villages within a sub-district are not uniform or equal. However, there are several sub-districts, namely District Jumo, District Kandangan, District Tretep, and District Wonoboyo, of which all of the villages are included in Cluster 2.

4. Conclusion

This study contributes to the development of geodemographic clustering methodology. The study successfully develops and applies two methods, FKP-GW and GA-FKP-GW, for cluster analysis of mixed-type geographic data. Experiments comparing the proposed method with the previous method in clustering simulation and study case data have shown that the proposed method yields different clustering results. However, the evaluation results indicate that the proposed method is still less efficient than the previous method. This proposed method is highly effective and can be utilized to conduct clustering analysis of geographic data that contains a combination of different types.

The empirical implementation of the proposed method GA-FKP-GW in clustering analysis for data PODES, which serves as a composition indicator for calculating VDI of Temanggung Regency 2018, successfully grouped 289 villages into four clusters. Villages with an independent status dominate clusters 1 and 3, while villages with a developing status dominate clusters 2 and 4. The result of descriptive analysis for each cluster has its own characteristics: cluster 1 is good in health services, cluster 3 is good in economic infrastructure, cluster 4 has the most stable and adequate infrastructure among all clusters, and cluster 2 is the cluster with the least infrastructure with the distance farthest.

This study still has limitations that can be improved in further research. This study repeatedly tests several parameters, including alpha and beta, potentially leading to less than optimal results. Therefore, further research is required to identify suitable parameters. On the other hand, some genetic algorithm

operators used in this study only focus on reducing the processing time, so it is possible to apply other operators to get more optimal results. Nevertheless, this study can serve as a reference for the development of the geodemographic analysis method and clustering analysis.

Ethics approval

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Agung Budi Santoso: Conceptualization, Methodology, Software, Data Curatiom, Writing – Original Draft, Writing – Review and Editing, Visualization. **Arya Candra Kusuma**: Conceptualization, Writing – Original Draft, Writing – Review and Editing. **Rani Nooraeni:** Conceptualization, Methodology, Validation. **Arie Wahyu Wijayanto:** Validation, Writing – Review and Editing.

References

[1] P. Sleight, *Targeting customers : how to use geodemographic and lifestyle data in your business / Peter Sleight.*, Second edi. Henley-on-Thames: NTC, 1997.

[2] R. Harris, P. Sleight, and R. Webber, *Geodemographics, GIS and Neighbourhood Targeting*. in Mastering GIS: Technol, Applications & Mgmnt. Wiley, 2005. [Online]. Available: https://books.google.co.id/books?id=Z8K25AxTjDcC

[3] L. H. Son, B. C. Cuong, P. L. Lanzi, and N. T. Thong, "A novel intuitionistic fuzzy clustering method for geo-demographic analysis," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9848–9859, Aug. 2012, doi: 10.1016/j.eswa.2012.02.167.

[4] G. A. Mason and R. D. Jacobson, "Fuzzy Geographically Weighted Clustering," in *Proceedings of the 9th International Conference on Geocomputation*, Sep. 2007, pp. 1–7.

[5] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, no. 2–3, pp. 191–203, 1984, doi: 10.1016/0098-3004(84)90020-7.

[6] H. Izakian and A. Abraham, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1835–1838, Mar. 2011, doi: 10.1016/j.eswa.2010.07.112.

[7] J. Wu, H. Xiong, C. Liu, and J. Chen, "A generalization of distance functions for fuzzy c-means clustering with centroids of arithmetic means," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 3, pp. 557–571, 2012, doi: 10.1109/TFUZZ.2011.2179659.

[8] Z. Feng and R. Flowerdew, "Fuzzy geodemographics: a contribution from fuzzy clustering methods," in *Innovations In GIS 5*, CRC Press, 1998, pp. 141–149. doi: 10.1201/b16831-20.

[9] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. 1981.

[10] L. Hunt and M. Jorgensen, "Clustering mixed data," *WIREs Data Min. Knowl. Discov.*, vol. 1, no. 4, pp. 352–361, Jul. 2011, doi: 10.1002/widm.33.

[11] D.-W. Kim, K. H. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern Recognit. Lett.*, vol. 25, no. 11, pp. 1263–1271, Aug. 2004, doi: 10.1016/j.patrec.2004.04.004.

[12] Z. Huang, "Clustering Large Data Sets With Mixed Numeric And Categorical Values," Proceedings Of 1st Pacific-Asia Conference on Knowledge Discouvery And Data Mining," 1997, *Singapore*.

[13] X. Zhong, T. Yu, and H. Xia, "A new partition-based clustering algorithm for mixed data," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2017.

[14] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Knowledge-Based Syst.*, vol. 30, pp. 129–135, Jun. 2012, doi: 10.1016/j.knosys.2012.01.006.

[15] J. Ji, Y. Chen, G. Feng, X. Zhao, and F. He, "Clustering mixed numeric and categorical data with artificial bee colony strategy," *J. Intell. Fuzzy Syst.*, vol. 36, no. 2, pp. 1521–1530, Mar. 2019, doi: 10.3233/JIFS-18146.

[16] W. Alomoush and A. Alrosan, "Review: Metaheuristic Search-Based Fuzzy Clustering Algorithms," *CoRR*, vol. abs/1802.0, 2018, [Online]. Available: http://arxiv.org/abs/1802.08729

[17] G. Gan, J. Wu, and Z. Yang, "A genetic fuzzy <math altimg="si218.gif" overflow="scroll"> <mrow> <mi>k</mi> </mrow> </math> -Modes algorithm for clustering categorical data," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 1615–1620, Mar. 2009, doi: 10.1016/j.eswa.2007.11.045.

[18] W. Min and Y. Siqing, "Improved K-means clustering based on genetic algorithm," in *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, 2010, pp. V6-636-V6-639. doi: 10.1109/ICCASM.2010.5620383.

[19] B. I. Nasution, R. Kurniawan, T. H. Siagian, and A. Fudholi, "Revisiting social vulnerability analysis in Indonesia: An optimized spatial fuzzy clustering approach," *Int. J. Disaster Risk Reduct.*, vol. 51, Dec. 2020, doi: 10.1016/j.ijdrr.2020.101801.

[20] B. S. Hadi, "Pendekatan Modified Particle Swarm Optimization dan Artificial Bee Colony pada Fuzzy Geographically Weighted Clustering (Studi Kasus pada Faktor Stunting Balita di Provinsi Jawa Timur) [Modified Particle Swarm Optimization and Artificial Bee Colony Approach on Fuzzy Geographically Weighted Clustering (Case Study on Stunting Factors of Toddlers in East Java Province)]," *Inst. Teknol. Sepuluh Nop.*, 2017.

[21] R. Gupta, S. K. Muttoo, and S. K. Pal, "Meta-Heuristic Algorithms to Improve Fuzzy C-Means and K-Means Clustering for Location Allocation of Telecenters Under E-Governance in Developing Nations," *Int. J. FUZZY Log. Intell. Syst.*, vol. 19, no. 4, pp. 290–298, Dec. 2019, doi: 10.5391/IJFIS.2019.19.4.290.

[22] M. Gen and R. Cheng, "Genetic algorithms and engineering design, Canada," 1997, *John Wiley & Sons, Inc*.

[23] R. L. Haupt and S. E. Haupt, *Practical Genetic Algorithms*. in Wiley InterScience electronic collection. Wiley, 2004. [Online]. Available: https://books.google.co.id/books?id=k0jFfsmbtZIC

[24] E. Wirsansky, *Hands-On Genetic Algorithms with Python: Applying genetic algorithms to solve real-world deep learning and artificial intelligence problems*. Packt Publishing, 2020. [Online]. Available: https://books.google.co.id/books?id=A0vODwAAQBAJ

[25] A. W. Wijayanto, A. Purwarianti, and L. H. Son, "Fuzzy geographically weighted clustering using artificial bee colony: An efficient geo-demographic analysis algorithm and applications to the analysis of crime behavior in population," *Appl. Intell.*, vol. 44, no. 2, pp. 377–398, Mar. 2016, doi: 10.1007/s10489-015-0705-7.

[26] A. W. Wijayanto, S. Mariyah, and A. Purwarianti, "Enhancing clustering quality of fuzzy geographically weighted clustering using Ant Colony Optimization," in *4th International Conference on Data and Software Engineering (ICoDSE 2017)*, Palembang, Indonesia: institute of electrical and electronics engineers (IEEE), 2018. doi: 10.1109/ICODSE.2017.8285858.

[27] S. Pramana and I. H. Pamungkas, "Improvement Method of Fuzzy Geographically Weighted Clustering using Gravitational Search Algorithm," *J. Ilmu Komput. dan Inf.*, vol. 11, no. 1, p. 10, Feb. 2018, doi: 10.21609/jiki.v11i1.580.

[28] B. I. Nasution, F. M. Saputra, R. Kurniawan, A. N. Ridwan, A. Fudholi, and B. Sumargo, "Urban vulnerability to floods investigation in jakarta, Indonesia: A hybrid optimized fuzzy spatial clustering and news media analysis approach," *Int. J. Disaster Risk Reduct.*, vol. 83, Dec. 2022, doi: 10.1016/j.ijdrr.2022.103407.

[29] R. Nooraeni, "Cluster Method Using A Combination of Cluster K-Prototype Algorithm and Genetic Algorithm for Mixed Data," *J. Apl. Stat. Komputasi Stat.*, vol. 7, no. 2 SE-Articles, p. 17, Dec. 2015, doi: 10.34123/jurnalasks.v7i2.23.

[30] R. Nooraeni, N. P. Yudho, and S. Pramana, "Mapping the socio-economic vulnerability in Aceh to reduce the risk of natural disaster," 2018, p. 030012. doi: 10.1063/1.5062736.

[31] R. Nooraeni, M. I. Arsa, and N. W. Kusumo Projo, "Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering," *Procedia Comput. Sci.*, vol. 179, pp. 677–684, 2021, doi: 10.1016/j.procs.2021.01.055.
[32] BPS, "Indeks Pembangunan Desa 2018 [Village Development Index 2018]," Jakarta, 2019.
[33] BPS, "Statistik Potensi Desa Indonesia (Village Potential Statistics Of Indonesia) 2018," Jakarta, 2018.
[34] E. J. Pebesma and R. Bivand, "Classes and methods for spatial data in {R}," *R News*, vol. 5, no. 2, pp. 9–13, Nov. 2005, [Online]. Available: https://cran.r-project.org/doc/Rnews/
[35] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. in Pearson International Edition. Pearson Addison Wesley, 2006. [Online]. Available: https://books.google.co.id/books?id=_XdrQgAACAAJ
[36] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," 2011.
[37] S. Pramana, B. Yuniarto, I. Santoso, R. Nooraeni, and L. H. Suadaa, *Data Mining dengan R, Konsep dan Implementasi [Data Mining with R, Concepts and Implementation]*. 2023.
[38] S.-H. Jun, "An Optimal Clustering using Hybrid Self Organizing Map," *Int. J. Fuzzy Log. Intell. Syst.*, vol. 6, no. 1, pp. 10–14, Mar. 2006, doi: 10.5391/IJFIS.2006.6.1.010.
[39] G. Szepannek, "clustMixType: User-Friendly Clustering of Mixed-Type Data in R," *R J.*, vol. 10, no. 2, p. 200, 2019, doi: 10.32614/RJ-2018-048.
[40] C.-C. Hsu and Y.-P. Huang, "Incremental clustering of mixed data based on distance hierarchy," *Expert Syst. Appl.*, vol. 35, no. 3, pp. 1177–1185, Oct. 2008, doi: 10.1016/j.eswa.2007.08.049.
[41] W. Johnson and R. Wichern, "Applied Multivariate Statistical Analysis Sixth Edition," 2007.
[42] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and statistics for engineers and scientists*, vol. 5. Macmillan New York, 1993.