Iranian Journal of Fuzzy Systems

Volume *, Number *, (****), pp. **

University of Sistan and Baluchestan

A generalization of the k-means algorithm for clustering data based on fuzzy population

M. Rezaei [1], A. Parchami [2] and A. Sheikhi [3]

[1,2,3] *Department of Statistics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran*

Rezaei.Marzieh2021@math.uk.ac.ir, parchami@uk.ac.ir, sheikhy.a@uk.ac.ir

Abstract

Clustering is one of the data mining tools frequently used in multivariate methods to group observations into clusters based on the similarity of variables. While many clustering algorithms have been developed for data extracted from well-defined (crisp) populations, numerous real-world problems involve inherently fuzzy populations such as high-consumption households or high-quality products. In such cases, each data point is described not only by its observed values but also by its degree of membership to the fuzzy population. This study proposes a novel clustering algorithm specifically designed for data extracted from fuzzy populations. The proposed method generalizes the classical k-means algorithm by considering membership degrees in the clustering process. The effectiveness of the proposed algorithm was evaluated on both synthetic and real-world datasets using several validity indices. According to these indices, the algorithm demonstrates satisfactory clustering quality and computational efficiency. The analysis of the proposed algorithm in the presence of outliers shows that, although outliers can slightly shift cluster centers, the overall clustering structure remains stable when clusters are well-separated.

Keywords: k-means algorithm, membership function, fuzzy population, weighting.

1 Introduction and preliminaries

Clustering is a fundamental approach aiming to partition a dataset into distinct groups, or clusters, such that the similarity among objects within cluster is significantly higher than that between clusters. Among most popular clustering techniques, the k-means algorithm is regarded as one of the most well-established methods. Despite its simplicity, k-means provides a conceptual foundation for numerous advanced clustering algorithms and can also be effectively applied to qualitative data. In this approach, observations are initially grouped into a predetermined number of clusters; often through an initial hierarchical step; and are iteratively reassigned to minimize within-cluster distances while maximizing between-cluster separation. Recent studies have focused on improving clustering and pattern recognition methods. Borlea et al. [11] enhanced the efficiency of the k-means algorithm by introducing a new centroid update approach. Hafs et al. [26] used the fusion of signature and fingerprint data in multimodal biometric systems to improve recognition accuracy. Ning [13] employed neural networks within the framework of edge computing to enhance the speed and efficiency of pattern recognition. For a comprehensive overview of clustering methodologies and their applications, see [8, 20, 18, 5, 1, 23].

In multidimensional datasets, variables contribute unequally to clustering. Feature weighting emphasizes informative variables, with variable selection as a special case assigning zero weight to excluded features [24, 28]. Beyond feature weighting, several studies have emphasized that the importance of observations (samples) may also vary. Observation weighting involves assigning a positive numerical weight to each observation to represent its influence within the clustering process. This mechanism allows algorithms to prioritize more reliable or informative data points. Observation weighting can be implemented through fixed or predetermined values reflecting the perceived significance of each

Corresponding Author: M. Rezaei
Received: **; Revised: **; Accepted: **.

observation, or through adaptive schemes where the weights are iteratively updated during the learning phase [2, 14]. For a more comprehensive discussion on this topic, the reader is referred to [9, 21].

Weighted k-means extends standard k-means by assigning weights to data points or features to reflect their importance. Point-weighted k-means assigns weights to data points and computes centroids as weighted means [16]. Recent studies have focused on developing weighting schemes and modifying clustering algorithms to improve performance on complex datasets. Gondeau et al. [3] proposed an object-weighted approach in which each data point is assigned a weight based on its influence on cluster overlap and outliers, enabling more effective computation of cluster centroids. Ning et al. [29] introduced the WeDIV algorithm, which employs a weighted distance measure; feature weights are determined according to their importance in minimizing within-cluster variance and a novel internal validation index. Borgwardt et al. [17] presented Balanced k-means for clustering with size constraints, where weights and centroids are computed to maintain balanced cluster memberships. So far, all studies on the weighted k-means algorithm, which assign weights to samples, have been conducted for data extracted from a classical population. Various methods for selecting appropriate weights have also been proposed by researchers from the perspective of classical population theory. Moreover, in some cases, data extracted from an imprecise (uncertain) population have been clustered using classical population-based approaches [15, 27].

In many practical applications, statistical analyses are conducted based on a census or a sample drawn from a fuzzy population [6]. Conventional methods [15, 10, 27] typically approximate a fuzzy population using crisp intervals prior to data collection and analysis, which can lead to information loss. Representative applications include assessing osteoporosis among elderly individuals, evaluating blood pressure in obese populations, analyzing the nutritional content of high-calorie foods, estimating yield from low-productivity trees, measuring average dust concentration during stormy days, and studying consumption patterns in low-income households. Within classical statistical frameworks, clustering is generally performed without considering the degree of membership of each observation in the fuzzy population. As a result, the intrinsic uncertainty inherent in such populations is ignored, potentially leading to a loss of valuable information. For instance, if the goal is to cluster data extracted from high-income households, the analyst may not know precisely which households and with what income levels should be considered in the clustering analysis. A more suitable approach involves modeling the fuzzy population through appropriate membership functions, thereby assigning specific membership degrees to each observation. Consequently, conventional clustering algorithms should be generalized to account for the varying influence of observations according to their membership degrees. In other words, clustering and related statistical analyses should be designed to assign greater weight to data points with higher membership degrees, while reducing the influence of those with lower degrees [6]. The practical perspectives of the proposed method presented in this article are considerably broad, as the approach is grounded in the concept of fuzzy populations, for which numerous real-world examples can be found across various scientific domains. In this article, we propose a novel clustering algorithm specifically designed for data extracted from a fuzzy population. The proposed method represents a generalization of the classical k-means algorithm, where observation weights are assigned based on membership functions that characterize the underlying fuzzy (or imprecise) population. The novelty of the proposed method lies in the approach used to determine the weights of samples (observations), which is based on examining the problem from the perspective of a fuzzy population. This approach enables a more comprehensive and appropriate assessment for clustering the data from a fuzzy population.

The paper is structured as follows. Section 2 provides a review of the definition of a fuzzy population and introduces the concept of the mean for data extracted from such a population. In Section 3, we present the proposed clustering algorithm, which is developed based on data extracted from a fuzzy population. Section 4 introduces several internal clustering quality evaluation criteria applied to the proposed algorithm. In Section 5, the benefits of analyzing clustering problems from a fuzzy population perspective, as opposed to a classical population viewpoint, are illustrated through a simple example. Section 6 evaluate the performance of the proposed method using a real-world dataset. Finally, Section 7 provides a detailed examination of the proposed method's behavior in the presence of outliers.

2 Fuzzy population

In statistical analysis, two fundamental sets are typically considered: the target population and the sample set. Classical statistics assumes that both of these sets - that is, the population and the sample - are precisely defined and well-delineated, implying that the membership or non-membership of each individual in the population or sample is unequivocal and unambiguous. However, in practical applications, two alternative scenarios are often encountered [6]:

(I) In cases when the members of the sample are not clearly defined, or the sample observations lack precise reporting, such samples are referred to as fuzzy or imprecise samples. For instance, when measuring soil salinity at various points within a region, the resulting data may exhibit some degree of ambiguity [12]. Similarly, in assessing

pain levels among patient samples, responses are often expressed using qualitative terms such as "severe pain", "moderate pain", "tolerable pain", and "mild pain" among others [22].

(II) There are situations when the target population is not precisely and clearly defined, resulting in what is known as an imprecise target population. For instance, if the objective is to study humidity levels on rainy days, the target population is defined as "rainy days". It is clear that different rainy days, characterized by varying amounts of rainfall, are not equally represented within this set. Instead, days with higher rainfall possess a greater degree of membership in the "rainy days" set, and vice versa. Another example is the study of income levels among "youth in Kerman". In this case, the target population of "youth in Kerman" is not a well-defined and precise set [6].

Definition 2.1. [6] *A Fuzzy Statistical Population – or simply fuzzy population refers to a fuzzy set of population under study, that is, all individuals, objects or items under study for which relevant data are collected. Similar to any fuzzy set, a fuzzy population can be determined by a membership function which can assigned a degree of membership to the fuzzy population between 0 and 1 for each unit.*

Remark 2.2. *The definition of "fuzzy population" reduces to the usual concept of statistical population, by replacement of the indicator function of usual population instead of the fuzzy population membership function.*

Example 2.3. *Suppose we aim to calculate the average income of young employees in Shiraz. Since the concept of "youth" is inherently fuzzy and imprecise, the members of the young population in Shiraz can not precisely defined. In fact, each individual in Shiraz belongs to the fuzzy set of young people with a membership degree ranging between 0 and 1. In this context, the fuzzy population of young individuals in Shiraz is first defined by a membership function, for example, the membership function illustrated in Figure 1 as follows:*

$$\mu(x) = \begin{cases} \dfrac{x-10}{10} & 10 \le x < 20 \\ 1 & 20 \le x < 30 \\ \dfrac{40-x}{10} & 30 \le x < 40 \\ 0 & otherwise \end{cases}$$

For instance, an individual aged 36 *has a membership degree of* $\mu_{youth}(36) = 0.4$ *in the fuzzy population of Shiraz youth.*

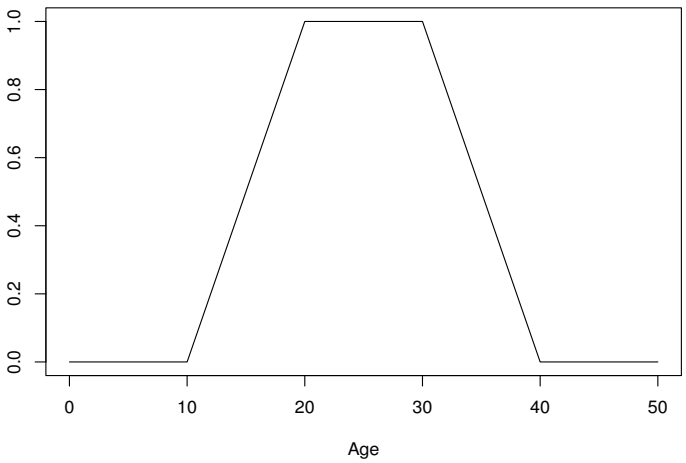


Figure 1: The membership function of youth based on age

After obtaining a random sample of individuals from this city (with a sample size of n), the sample data are recorded as $(x_1, \mu_1), \ldots, (x_n, \mu_n)$, where x_i denotes the income of the i-th individual and μ_i represents its degree of membership in the fuzzy set of young people in Shiraz. In this paper, such data are referred to as "data extracted from a fuzzy population" [6].

Definition 2.4. *Suppose the data* $(x_1, \mu_1), \ldots, (x_n, \mu_n)$ *are collected from a random sample drawn from a fuzzy population, where* x_i *corresponds to the data for the* i*-th individual, and* μ_i *denotes the degree of membership of the* i*-th individual in the fuzzy population of interest. Based on the data* $(x_1, \mu_1), \ldots, (x_n, \mu_n)$*, the mean of the data extracted from a fuzzy population using the following relationship is obtained.*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \mu_i}{\sum_{i=1}^{n} \mu_i}. \tag{1}$$

Remark 2.5. *If, instead of the membership function μ of the fuzzy set as defined in Eq. (1), the indicator function of the population, I_Ω, is utilized, it follows that $\sum_{i=1}^{n} \mu_i = n$. Consequently, the mean computed according to Eq. (1) simplifies to the classical arithmetic mean of the observations $x_1, \ldots, x_n$ as commonly employed in traditional statistics.*

From an implementation perspective, Table 1 illustrates key practical patterns in clustering based on fuzzy population. These characteristics make the table a valuable reference for designing innovative clustering systems. Three fundamental patterns emerge from this interdisciplinary analysis: first, the continuous and consistent transformation of real-world continuous phenomena into computable membership functions applicable across diverse domains; second, a unified conceptual framework that transcends disciplinary boundaries while preserving domain-specific implementations; and third, the demonstrated ability of fuzzy logic to bridge the gap between experts qualitative knowledge and quantitative data analysis. Collectively, these examples support Zadeh's original hypothesis that fuzzy sets can effectively model the inherent ambiguity present in natural systems and human reasoning processes. The table especially highlights how various scientific communities have adapted core fuzzy principles in a mathematically coherent way, developing membership functions tailored to address their unique clustering challenges.

Table 1: Applications of Clustering based on Fuzzy Population Across Various Scientific Domains

Scientific Domain	**Clustering Application**	**Potential fuzzy Population**	**Membership Criteria**
Cognitive Psychology	Talent identification programs	Gifted individuals	IQ scores, academic records
Medicine	Personalized treatment plans	Acute diabetic patients	Blood sugar level, body mass index (BMI)
Marketing	Customer segmentation	Loyal customers	Purchase frequency, customer feedback
Environmental Science	Urban planning policy clustering	Polluted areas	particulate levels, traffic data
Education	Student clustering	Talented students	Exam scores, participation rate
Food Science	Product grouping	High-quality products	Color, texture, defects
Dentistry	Treatment prioritization	Acute dental patients	Tooth decay measurements
Veterinary Medicine	Epidemic clustering	Large livestock	Animal disease symptoms
Financial Management	Loan applicant clustering	High credit risk applicants	Credit history, income
Sociology	Social welfare program design	Underprivileged neighborhoods	Income, access to Welfare services
Artificial Intelligence	Emotional state identification	Frustrated users	user behavior patterns
Agriculture	Disease detection systems	Healthy crops	Color, texture, spots
Bibliometrics	Research article classification	Highly cited research articles	Citation count, citation indices
Economics	Company classification for investment groups	Companies with relatively high returns	Returns, risk, trading volume
Geology	Earthquake preparedness	Seismic zones	Number, magnitude, depth of earthquakes
Environmental Health	Pollution monitoring	Contaminated water sources	Microbial contamination, heavy metals
Mechanical Engineering	Predictive maintenance scheduling	Worn machinery	Vibration sensor data
Ophthalmology	Disease screening for prevention	Acute eye patients	Retinal thickness, tissue changes, color patterns
Climatology	Agricultural planning	Favorable climatic regions	Temperature, rainfall, humidity, wind
Education Management	Resource allocation	Top educational institutions	Grades, parental satisfaction, facilities
Speech Therapy	Rehabilitation program design	Acute speech disorder patients	Stuttering, voice disorders, speech delays
Market Research	Inventory management	Best-selling products	weekly sales rate, product returns, peak purchase times
Pharmacology	Drug dosage optimization	Acute cancer patients	Genes, proteins, other biological markers
Soil Science	Soil conservation planning	Fertile agricultural lands	Slope, soil type, vegetation cover, rainfall
Automotive Engineering	More precise insurance premium assignment	Drivers with high-risk behavior	Speed, sudden braking, accident count
Epidemiology	Vaccination strategies	High-risk individuals for disease spread	Infection rates, population density, age
Acoustics	Clustering in audio equipment design	High-power sound systems	Sound clarity, noise level, frequency response
Energy Technology	Battery management	Degraded electric vehicles	Voltage, capacity, charge/discharge cycles
Clinical Psychology	Identification of groups in critical	Individuals in severe psychological crisis	Stress level, anxiety severity, depression symptoms

In the following section, a novel method for clustering data extracted from a fuzzy population is presented.

3 Clustering data extracted from fuzzy population

In this section, we introduce the Fuzzy Weighted k-means (FWk-means) algorithm to cluster data extracted from a fuzzy population. The FWk-means algorithm can be regarded as a weighted generalization of the classical k-means technique, specifically designed to incorporate fuzzy characteristics into the clustering process. Consider the observation matrix $\mathbf{X}$, comprising n observations and m variables, defined as follows:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} = [x_{ij}]_{n\times m}.$$

The matrix $\mathbf{X}$ can alternatively be represented by its columns and rows as follows

$$\mathbf{X} = \left[\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}\right] = \left[\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_n^T\right]^T,$$

where $\mathbf{x}^{(j)} = \left[x_{1j}, \dots, x_{nj}\right]^T$ and $\mathbf{x}_i = \left[x_{i1}, \dots, x_{im}\right]$ denote the vectors of the j-th variable/feature and the i-th observation from the matrix $\mathbf{X}$, respectively.

The main goal is to cluster the rows of the matrix $\mathbf{X}$, representing a set of observed data points, into k distinct groups or clusters. Let $P_1, P_2, \dots, P_p$ denote p initial fuzzy populations with corresponding membership functions $\mu_1, \mu_2, \dots, \mu_p$, respectively. Each membership function μ_i is determined by skilled professionals based on a subset of variables. The goal is to define the membership function of the interested/final fuzzy population, denoted by P_{final}, $P_{\text{final}} = P_1 \cap P_2 \cap .... \cap P_p$. To compute the membership function of the fuzzy population P_{final}, different t-norms (such as the minimum t-norm, the product t-norm, and others) can be employed. In this study, the minimum t-norm is primarily adopted, as the generalization of the intersection concept and its definition through the minimum operator provide a natural and intuitive formulation.

Assume $= [\mu_1, \mu_2, \dots, \mu_n]^T$ be the vector of membership degrees for n observations into fuzzy population P_{final}, in which

$$\mu_i = \min\{\mu_1(\mathbf{x}_i), \dots, \mu_p(\mathbf{x}_i)\}; i = 1, 2, \dots, n. \tag{2}$$

is the degree of belonging $\mathbf{x}_i$ to fuzzy population P_{final}, for $i = 1, 2, \dots, n$. Also $\mu_s(\mathbf{x}_i)$ for $i = 1, 2, \dots, n$ and $s = 1, 2, \dots, p$ denotes the membership degree of the i-th observation in the initial s-th fuzzy population.

For example, suppose we are going to perform clustering based on the fuzzy population of high-quality and high-consumption products produced by a factory, denoted as P_{final}. In this case, we deal with two initial fuzzy populations, "high-quality products" and "high-consumption products", such that $P_{\text{final}} = P_1 \cap P_2$. Assume that the i-th product has a membership degree of $\mu_1(\mathbf{x}_i) = 0.7$ in the initial fuzzy population of high-quality products, and also the membership degree of $\mu_2(\mathbf{x}_i) = 0.5$ in the initial fuzzy population of high-consumption products. Therefore, the membership degree of the i-th product in the fuzzy population of "high-quality and high-consumption products" is equqal to $\mu_i = \min\{\mu_1(\mathbf{x}_i), \mu_2(\mathbf{x}_i)\} = \min\{0.7, 0.5\} = 0.5$.

The Euclidean distance matrix is defined by

$$\mathbf{D} = d(\mathbf{X}, \mathbf{C}) = [d_{il}]_{n\times k}, \tag{3}$$

where $\mathbf{C} = [c_{lj}]_{k\times m}$ represents the matrix of cluster centers. The quantity d_{il} denotes the Euclidean distance between the i-th observation and the center of the l-th cluster for $i = 1, 2, \dots, n$ and $l = 1, 2, \dots, k$, and equals to

$$d_{il} = d(\mathbf{x}_i, \mathbf{c}_l) = \|\mathbf{x}_i - \mathbf{c}_l\|_2 = \sqrt{(x_{i1} - c_{l1})^2 + \dots + (x_{im} - c_{lm})^2}. \tag{4}$$

here, $\|\cdot\|_2$ denotes the Euclidean (L_2) norm of a vector [7].

The objective function of the proposed method is defined as

$$T(\mathbf{U}, \mathbf{C}, \boldsymbol{\mu}) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il}\ \mu_i\ d_{il}^2. \tag{5}$$

The i-th observation is assigned to the l-th cluster whose center minimizes the objective function $T(\mathbf{U}, \mathbf{C}, \boldsymbol{\mu})$, i.e.,

$$l_i = \arg\min_{1\le l\le k} T(\mathbf{U}, \mathbf{C}, \boldsymbol{\mu}),$$

where l_i represents the index of the nearest cluster center to the i-th observation. Here, $\mathbf{U} = [u_{il}]_{n\times k}$ denotes the membership matrix in which

$$u_{il} = \begin{cases} 1 & \mathbf{x}_i \text{ is belong to } l\text{-th cluster} \\ 0 & \mathbf{x}_i \text{ is not belong to } l\text{-th cluster,} \end{cases} \tag{6}$$

represents the degree of membership of the i-th observation to the l-th cluster. In this context, u_{il} acts as an indicator function representing the membership of the i-th observation into the l-th cluster, for $i = 1, 2, \ldots, n$ and $l = 1, 2, \ldots, k$ such that the constraint $\sum_{l=1}^{k} u_{il} = 1$ is satisfied for each observation.

Theorem 3.1. *In each iteration of the* FWk-means *algorithm the coordinates of the l-th cluster center is equal to $\mathbf{c}_l = [c_{l1}, \ldots, c_{lm}]^T$ for $l = 1, 2, \ldots, k$, where*

$$c_{lj} = \frac{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} \mu_i x_{ij}}{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} \mu_i}, \quad j = 1, 2, \ldots, m. \tag{7}$$

Proof. The objective of the FWk-means algorithm is to minimize the following function with respect to the center of l-th cluster ($\mathbf{c}_l$):

$$T(\mathbf{U}, \mathbf{C}, \boldsymbol{\mu}) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il}\ \mu_i\ d_{il}^2 = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il}\ \mu_i \times \|\mathbf{x}_i - \mathbf{c}_l\|_2^2 . \tag{8}$$

To minimize Eq. (8) with respect to $\mathbf{c}_l$, we take the derivative of the objective function and set it to zero as

$$\frac{\partial}{\partial \mathbf{c}_l} T(\mathbf{U}, \mathbf{C}, \boldsymbol{\mu}) = -2 \sum_{i=1}^{n} u_{il} \mu_i (\mathbf{x}_i - \mathbf{c}_l) = 0. \tag{9}$$

$$\mathbf{c}_l = \frac{\sum_{i=1}^{n} u_{il} \mu_i \mathbf{x}_i}{\sum_{i=1}^{n} u_{il} \mu_i} = \frac{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} u_{il} \mu_i \mathbf{x}_i + \sum_{i:\mathbf{x}_i \notin \{l\text{-th cluster}\}} u_{il} \mu_i \mathbf{x}_i}{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} u_{il} \mu_i + \sum_{i:\mathbf{x}_i \notin \{l\text{-th cluster}\}} u_{il} \mu_i}. \tag{10}$$

Based on the indicator function in (6), Eq. (10) simplifies to

$$\mathbf{c}_l = \frac{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} \mu_i \mathbf{x}_i}{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} \mu_i}, \quad l = 1, 2, \ldots, k. \tag{11}$$

Consequently, the coordinates of the l-th cluster center are determined as follows

$$c_{lj} = \frac{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} \mu_i x_{ij}}{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} \mu_i}, \quad j = 1, 2, \ldots, m, \quad l = 1, 2, \ldots, k. \tag{12}$$

which is equivalent to Eq. (7). □

The Eq. (7) is a generalized version of the proposed mean of the data extracted from fuzzy population in Eq. (1). The FWk-means algorithm is not merely a numerical procedure; rather, it provides a solution to an optimization problem that formalizes the clustering task based on a fuzzy population. In other words, the clustering problem defined over a fuzzy population can be formulated as a mathematical optimization problem, which is defined by the objective function of the FWk-means algorithm in Eq. (5). FWk-means effectively solves this optimization problem by minimizing the objective function in Eq. (5) at each iteration, either approximately or iteratively. The FWk-means algorithm is introduced to partition n observations $\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T$ into k clusters, with the objective of minimizing the function defined in Eq. (5) using the procedure outlined in Algorithm 1.

Algorithm 1 FWk-means clustering procedure for partitioning n data points into l clusters.

Require:
(1) Dataset $\mathbf{X} = \left[\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T\right]^T$, where $\mathbf{x}_i^T \in \mathbb{R}^m$
(2) Number of clusters k, where $2 \leq k < n$
(3) Membership degrees vector $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_n]^T$, where $\mu_i \in [0, 1]$
Ensure: Cluster centroids $\mathbf{c}_1, \ldots, \mathbf{c}_k$ and assignment of each $\mathbf{x}_i^T$ to a cluster

`Randomly select` k `initial centroids` $\mathbf{c}_1, \ldots, \mathbf{c}_k$ `from the data points` $\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T$
repeat
  for $i = 1$ `to` n **do**
    `Assign` $\mathbf{x}_i^T$ `to` l`-th cluster such that:`

$$l_i = \arg \min_{1 \leq l \leq k} T(\mathbf{U}, \mathbf{C}, \boldsymbol{\mu})$$

  where l_i denotes the index of the cluster center closest to observation i.
  end for
  for $l = 1$ `to` k **do**
    `Update centroid` $\mathbf{c}_l = [c_{l1}, \ldots, c_{lm}]^T$ `using the mean of assigned points:`

$$c_{lj} = \frac{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} \mu_i x_{ij}}{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} \mu_i}, \quad j = 1, 2, \ldots, m.$$

  end for
until `Centroids do not change`
return `Final cluster centroids and assignments`

Remark 3.2. *The main goal of the* FWk-means *clustering algorithm at each iteration is to update the cluster centers in a manner that leads to the minimization of the objective function given in (5). In other words, the algorithm aims to determine the set of cluster centers* $\mathbf{c}_l = [c_{l1}, \ldots, c_{lm}]^T$ *for* $l = 1, \ldots, k$ *that minimizes the objective function. According to Theorem 3.1, this minimum is achieved at the point where each component* c_{lj}*, for* $j = 1, \ldots, m$*, is computed using Eq. (7). In the case when the considered population is classical (crisp), and according to Remark 2.2 since* $\mu_i = 1$ *for all* $i = 1, \ldots, n$*, the objective function (5) reduces to* $T(\mathbf{U}, \mathbf{C}, \boldsymbol{\mu}) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il}\, d_{il}^2$*, which coincides with the objective function of the conventional* k-means *algorithm. The objective function of the* k-means *algorithm is minimized when the cluster centers are set to the arithmetic mean of the data points belonging to each cluster. Therefore, it can be concluded that the* FWk-means *algorithm, designed for clustering data extracted from a fuzzy population, is a generalization of the classical* k-means *algorithm.*

Remark 3.3. *In Eq. (5),* u_{il} *and* μ_i *denote, respectively, the membership degree of the ith observation in the lth cluster and its membership degree in the corresponding fuzzy population. Since the clustering structure considered in this study is hard (non-fuzzy),* u_{il} *takes the value of one if the ith observation is assigned to cluster l, and zero otherwise. Furthermore, the values of* μ_i *lie within the interval* $[0, 1]$*, with larger values indicating a greater relative importance of the observation within the fuzzy population. Accordingly, based on Eq. (5), at each iteration of the* FWk-means *algorithm, the weighted sum of squared distances between each observation and the center of its assigned cluster is minimized. This minimization process continues until the algorithm converges to a local minimum.*

4 Internal evaluation criteria for FWk-means algorithm

In this section, several indices are introduced to evaluate the quality of the proposed clustering algorithm. Internal clustering evaluation indices are tools that assess the quality of cluster structures without requiring the true class labels of the data. These indices are defined based on intrinsic properties of the dataset, such as cluster compactness and separation. By quantifying the balance between the closeness of data points within a cluster and the separation between clusters, these indices provide a measure of the overall cohesion and distinctness of the clustering results. The computation of these indices in the context of a fuzzy population differs somewhat from their computation in a classical (crisp) population setting.

Fuzzy Weighted Sum of Squared Errors (FWSSE): The Fuzzy Weighted Sum of Squared Errors is introduced to evaluate clustering quality, which measures the dispersion of data points within each cluster and is defined as the sum of squared distances between each data point and the weighted centroid of the cluster it belongs to. Let $\mathbf{X} = \left[\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T\right]^T$ be a dataset partitioned into k clusters, and let $\mathbf{c}_l = [c_{l1}, \ldots, c_{lm}]^T$ denote the centroid of l-th cluster where c_{lj} is obtained from the Eq. (7) for $l = 1, \ldots, k$. Then, Fuzzy Weighted Sum of Squared Errors is given by:

$$FWSSE = \sum_{l=1}^{k} \sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} \mu_i \; d_{il}^2.$$

A lower FWSSE indicates that data points are closer to their respective cluster centroids, implying more compact and higher-quality clustering.

Fuzzy Weighted Davies Bouldin Index (FWDBI): The Fuzzy Weighted Davies Bouldin Index is considerd both cluster compactness and separation. Let S_l denote the average distance of the members of the l-th cluster to its centroid, and let $d_{ll'}$ represent the distance between the centroids of the l-th and l'-th clusters. The coordinates of all cluster centers used in the computation of this index are obtained according to Eq. (7). The similarity between clusters l and $l^{'}$ is defined as

$$R_{ll'} = \frac{S_l + S_{l'}}{d_{ll'}}, \quad l, l^{'} = 1, 2, \ldots, k, \; l \neq l^{'},$$

where S_l and $d_{ll'}$ are computed as follow

$$S_l = \frac{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} d_{il}}{\sum_{i:\mathbf{x}_i \in \{l\text{-th cluster}\}} \mu_i}, \quad l = 1, \ldots, k, \quad i = 1, \ldots, n.$$

where d_{il} is computed according to Eq. (4), and

$$d_{ll'} = d(\mathbf{c}_l, \mathbf{c}_{l'}) = \|\mathbf{c}_l - \mathbf{c}_{l'}\|_2, \quad l, l' = 1, 2, \ldots, k, \; l \neq l'.$$

The maximum similarity for cluster l is then defined as

$$R_l = \max_{l \neq l'} R_{ll'}, \quad l, l' = 1, 2, \ldots, k.$$

Finally, the Fuzzy Weighted Davies-Bouldin Index is computed as the average over all k clusters:

$$FWDBI = \frac{1}{k} \sum_{l=1}^{k} R_l.$$

lower FWDBI values indicate more compact and well-separated clusters, which corresponds to better clustering performance.

Fuzzy Weighted Dunn Index (FWDI) : The Fuzzy Weighted Dunn Index evaluates clustering quality by emphasizing clusters that are both compact and well-separated. Let $d_{ll'}$, for $l, l' = 1, 2, \ldots, k$, denote the distance between clusters l and $l^{'}$, and let δ_l represent the diameter of the l-th cluster, defined as the largest distance between any two points within the cluster, which is given by

$$\delta_l = \max_{\mathbf{x}_i, \, \mathbf{x}_{i'} \in \{l\text{-th cluster}\}} d_{ii'}, \quad l = 1, 2, \ldots, k,$$

where the distance between the i-th and $i^{'}$-th observations, $d_{ii'}$ is computed as follows

$$d_{ii'} = d(\mathbf{x}_i, \mathbf{x}_{i'}) = \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2 = \sqrt{(x_{i1} - x_{i'1})^2 + \cdots + (x_{im} - x_{i'm})^2}, \quad , i, i' = 1, 2, \ldots, n.$$

Also, the coordinates of all cluster's centers that have been used to compute this index are obtained according to Eq. (7). The Fuzzy Weighted Dunn Index is then defined as

$$FWDI = \frac{\min_{1 \leq l < l' \leq k} d_{ll'}}{\max_{1 \leq l < l' \leq k} \delta_l},$$

where k is the number of clusters. Higher FWDI values indicate better clustering performance, corresponding to clusters that are both more compact and better separated.

Remark 4.1. *The clustering criteria which were introduced in this section constitute generalizations of the conventional indices employed in the classical setting. When the underlying population is crisp, according to Remark 2.2, the membership degrees of all observations are $\mu_i = 1$ for $i = 1, \ldots, n$. Moreover, as noted in Remark 2.5, the cluster centers computed in Theorem 3.1 reduce to the arithmetic mean of the observations within each cluster. Therefore, the indices $FWSSE$, $FWDBI$, and $FEDI$, originally are proposed for evaluating clustering quality in a fuzzy population, correspond respectively to the Sum of Squared Errors (SSE), the Davies-Bouldin Index (DBI), and the Dunn Index (DI) in a classical population.*

5 A comparative study of clustering in classical and fuzzy populations

This section focuses on the clustering of data extracted from a fuzzy population representing "intelligent youth", analyzed from both classical and fuzzy perspectives. The dataset employed for this purpose has been synthetically generated via simulation. The aim of this clustering exercise is to facilitate a deeper understanding of fuzzy populations and to explore practical strategies for managing and analyzing data extracted from such populations in the context of clustering techniques.

5.1 Clustering of simulated data using the classical k-means algorithm

If we restrict ourselves to a formal definition of the problem, then, given that the concepts of youth and intelligence are inherently ambiguous and imprecise, multiple definitions of the population of "intelligent youth" can be proposed. Some of these assumed definitions are outlined below:
Population Ω_1: individuals aged between 10 and 45 years, with an IQ scores ranging from 85 to 180;
Population Ω_2: individuals aged between 15 and 35 years, with an IQ scores ranging from 110 to 175;
Population Ω_3: individuals aged between 18 and 30 years, with an IQ scores ranging from 130 to 170.

In this example, 100 simulated data points are generated from population Ω_1. A summary of these data points is presented in Table 2. Column 1 of Table 2 corresponds to the observation number, columns 2, 3, 4 and 5 represent the age, IQ of the individuals, average daily sleep duration (in minutes) and average daily duration use of technology (mobile and computer) measured in minutes, respectively. Columns 6 to 8 indicate the membership status of each individual within the classical population Ω_1, Ω_2 and Ω_3 defined above.

Table 2: Simulated data for individuals and their membership degrees into three populations Ω_1, Ω_2, Ω_3

Sample # (i)	*age* (x_{i1})	*IQ* (x_{i2})	*Sleep* (x_{i3})	*TechUsage* (x_{i4})	$I_{\Omega 1}$	$I_{\Omega 2}$	$I_{\Omega 3}$
1	40	111	418	241	1	0	0
2	24	126	253	339	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
29	32	173	571	49	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
46	24	139	302	410	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
99	41	157	453	418	1	0	0
100	16	107	366	146	1	0	0

Table 3 presents the descriptive statistics for the variables in the study for the classical population. For each variable, the mean, variance, and standard deviation are reported.

Table 3: Descriptive statistics of the key variables for the classical population

	mean	variance	sd
Age	29.23	92.64	9.62
IQ	131.79	697.96	26.41
Sleep	411.49	11218.99	105.91
TechUsage	336.02	26934.32	164.11

In this example, the number of variables is $m = 4$. The variables *Age* and *IQ* were employed to identify the population of intelligent young individuals. After defining the interested population, clustering was performed based on the variables of *average daily sleep duration* and *average daily duration use of technology* using the k-means algorithm. We are going to cluster individuals in the population of "intelligent youth" into $k = 2$ clusters on *Sleep* and *TechUsage* as basic features. Given the precise definitions of the populations Ω_1, Ω_2, and Ω_3, it is evident that some of the data points extracted from population Ω_1 do not belong to populations Ω_2 or Ω_3. Consequently, the sample sizes are $n_{\Omega_1} = 100, n_{\Omega_2} = 44$ and $n_{\Omega_3} = 14$. The scatter plot illustrating the 100 simulated data points and the three populations Ω_1, Ω_2 and Ω_3 is presented in Figure 2.

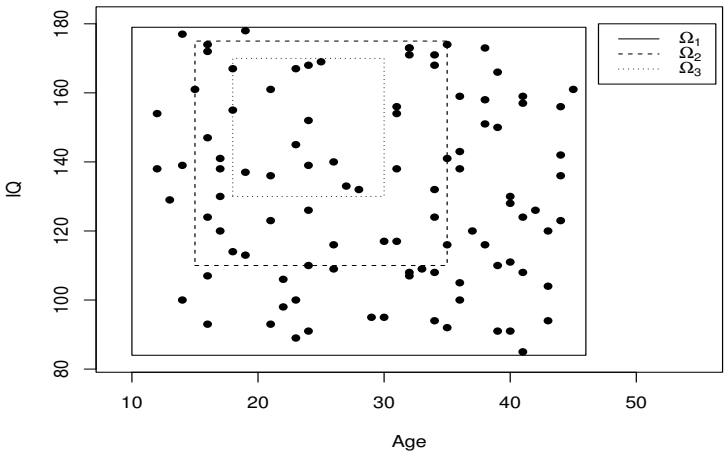


Figure 2: Scatter plot of the 100 simulated data points and three considered classic populations

The indicator function I_{Ω_r} in Table 2 denotes the membership status of each observation in the classical population Ω_r, for $r = 1, 2, 3$. The clustering results for the simulated data generated from the populations Ω_1, Ω_2 and Ω_3 using the k-means algorithm, partitioned into two clusters, are illustrated in Figure 3. In the final iteration of the k-means algorithm, the centroids of clusters 1 and 2 are located at $(413.28, 200.40)$, $(409.54, 482.93)$, respectively. The number of observations assigned to clusters 1 and 2 in this iteration are 52 and 48 respectively, as summarized in Table 7.

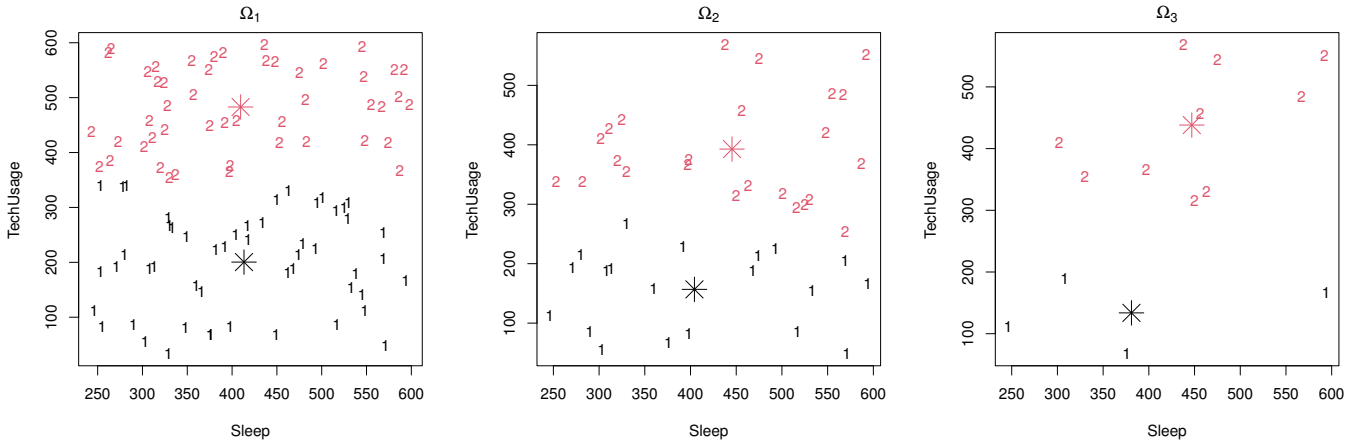


Figure 3: Clustering of the 100 data points extracted from populations Ω_1, Ω_2 and Ω_3 using the k-means algorithm

5.2 Clustering of data extracted from the fuzzy population of intelligent youth using the FWk-means algorithm

Since the concepts of youth and intelligence are inherently ambiguous and imprecise, it is preferable to initially define membership functions for each fuzzy variable − youth and intelligence − based on expert knowledge. By adopting this approach, there is no need to consider multiple separate definitions of the interested population. Furthermore, the population of "intelligent youth", being composed of two vague concepts, is modeled in an imprecise and fuzzy manner. The fuzzy membership functions corresponding to fuzzy concepts of youth and intelligence are respectively defined by follows:

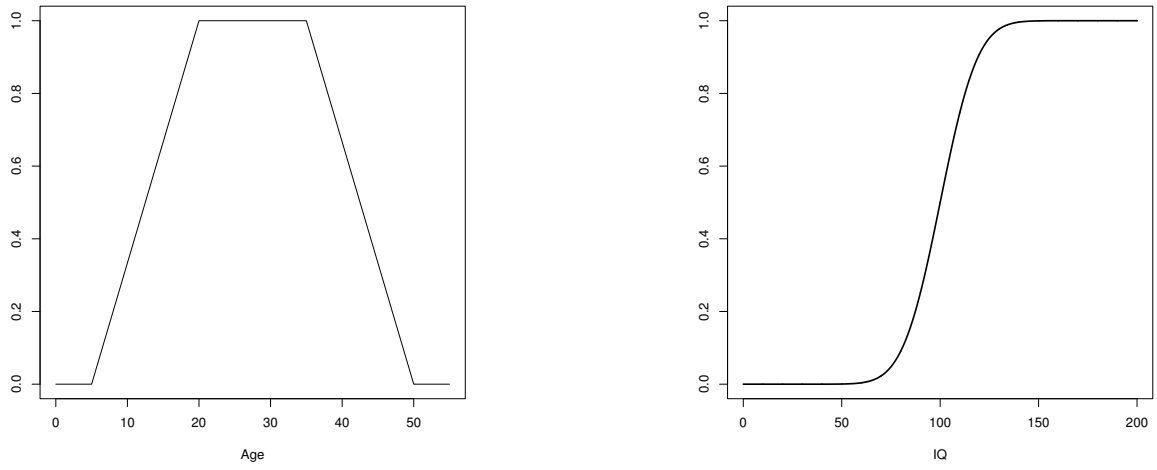
$$\mu_1(x) = \begin{cases} \dfrac{x-5}{15} & 5 \leq x < 20 \\ 1 & 20 \leq x < 35 \\ \dfrac{50-x}{15} & 35 \leq x < 50 \\ 0 & otherwise \end{cases}$$

and

$$\mu_2(x) = \Phi\big(\frac{x-100}{15}\big),$$

where Φ denotes the cumulative distribution function of the standard normal distribution [6]. The graphs of the membership functions of youth and intelligence are drawn in Fgure 4.

In Fuzzy logic studies, the selection of the membership function type and the determination of its parameters are among the critical steps in the design process. Commonly used membership functions include triangular, trapezoidal, Gaussian, and sigmoidal functions. The choice of the membership function depends on the complexity of the problem, the clarity of the fuzzy boundaries, and the need for smoothing. The parameters of membership functions, such as the center and width, are typically determined based on data, expert knowledge, and relevant literature. For instance, age ranges reported in scientific studies can be used to define the membership function representing the concept of youth. Similarly, a Gaussian membership function for IQ is often defined based on the fact that IQ ranges are typically determined using standardized intelligence tests such as the Wechsler and Stanford–Binet scales. In these tests, the mean score is set to 100 and the standard deviation to 15. For a more comprehensive review of methods for selecting membership functions and determining their parameters, the reader is referred to [4, 20, 19, 25].



(a) (b)

Figure 4: The membership functions of (a) Youth based on age (b) Intelligence based on IQ

Table 4 presents the descriptive statistics for the variables in the study for the fuzzy population. For details on the computation of mean, variance, and standard deviation for crisp data extracted from a fuzzy population, refer to [6].

Table 4: Descriptive statistics of the key variables for the fuzzy population

	mean	variance	sd
Age	28.62	78.65	8.86
IQ	136.95	600.58	24.50
Sleep	419.14	11283.44	106.22
TechUsage	322.28	25925.09	161.01

The scatter plot illustrating the 100 simulated data points from the fuzzy population of "intelligent youth" with their membership degrees is presented in Figure 5. In this graph, each observation is depicted according to its degree of membership in the corresponding fuzzy population.

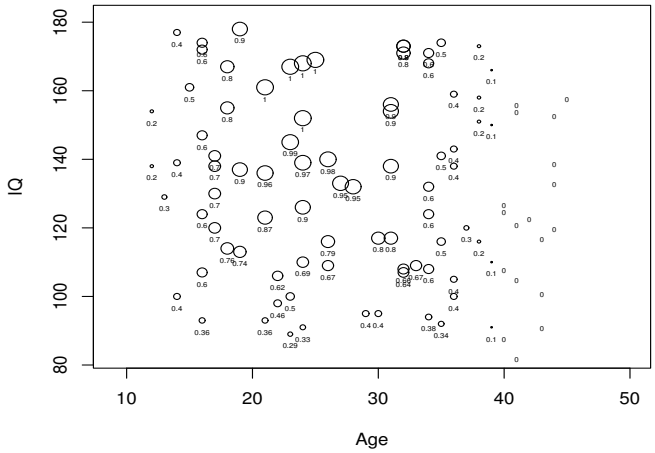


Figure 5: Scatter plot of the 100 simulated data points from the fuzzy population of intelligent youth with their membership degrees

According to Table 5, for instance, the 46^{th} individual with the age of 24 and the intelligence quotient of 139 belongs to the initial fuzzy populations of *young individuals* and *intelligent individuals* with the membership degrees $\mu_1(24) = 1$ and $\mu_2(139) = 0.97$, respectively. Moreover, this individual is a member of the final fuzzy population *intelligent youth* with the membership degree $\mu_{46} = min\{1, 0.97\} = 0.97$. Columns 6, 7, and 8 in Table 5 represent the degrees of membership of each observation in the initial fuzzy population of youth, the initial fuzzy population of intelligent individuals, and the final fuzzy population of intelligent youth ($p = 2$), respectively. The clustering results obtained by applying the proposed algorithm to the simulated data from the fuzzy population of "intelligent youth" are presented in Figure 6.

Table 5: Membership degrees of the observations to the fuzzy population of intelligent youth

Sample # (i)	*age* (x_{i1})	*IQ* (x_{i2})	*Sleep* (x_{i3})	*TechUsage* (x_{i4})	$\mu_1(x_{i1})$	$\mu_2(x_{i2})$	μ_i
1	40	111	418	241	0.66	0.70	0.66
2	24	126	253	339	1	0.90	0.90
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
29	32	173	571	49	1	0.99	0.99
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
46	24	139	302	410	1	0.97	0.97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
99	41	157	453	418	0.60	0.99	0.60
100	16	107	366	146	0.73	0.63	0.63

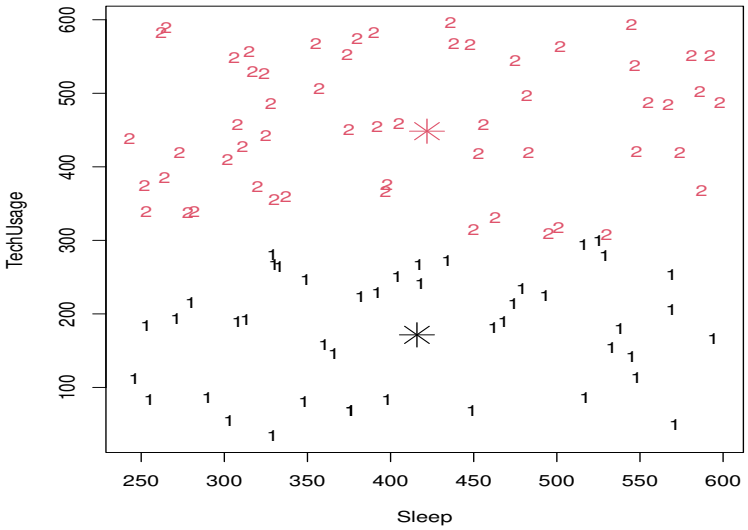


Figure 6: Clustering of data extracted from the fuzzy population of intelligent youth using the FWk-means algorithm

In the final iteration of the FWk-means algorithm, the centroids of clusters 1 and 2 with min operator are located at $(415.84, 171.49)$ and $(421.88, 448.46)$, respectively. The number of observations assigned to clusters 1 and cluster 2 in this iteration with min operator are respectively 44 and 56 (see Table 7).

5.3 Problem analysis: classical and fuzzy populations

Table 6 provides a comparative overview of FWk-means and several widely used clustering algorithms, emphasizing the type of data they operate on (sample status), the assumed population status, and the clustering approach (hard or fuzzy). As illustrated, algorithms such as k-means, k-medoids, and DBSCAN are designed for crisp data and perform hard clustering, whereas fuzzy c-means and fuzzy possibilistic c-means are intended for fuzzy data and implement fuzzy clustering. In contrast, FWk-means is specifically developed to handle crisp samples extracted from a fuzzy population while preserving hard clustering. This summary highlights the suitability of each algorithm for different types of data and clustering scenarios, providing a clear guide for selecting appropriate methods in practical applications.

Table 6: Comparison of FWk-means and other clustering algorithms by data and clustering type

Algorithm	Sample status	Population status	Clustering type
FWk-means	crisp	fuzzy	hard
k-means	crisp	crisp	hard
fuzzy c-means	fuzzy	crisp	fuzzy
fuzzy possibilistic c-means	fuzzy	crisp	fuzzy
k-medoids	crisp	crisp	hard
DBSCAN	crisp	crisp	hard

The intersection defined by the minimum operator is not the only admissible formulation; alternative t-norms may also be employed. Table 7 reports the cluster centroids, the number of observations assigned to each cluster, and the objective function values across all iterations for both the classical k-means and the FWk-means algorithms. The results for the FWk-means algorithm are provided separately for the minimum operator and the algebraic product operator. Tables 7 and 8 present the results produced by the two clustering algorithms, thereby facilitating an independent evaluation of their performance. Since the k-means and FWk-means algorithms are implemented within classical and fuzzy frameworks, respectively, a direct comparison of their results is not meaningful.

Table 7: Clustering results of 100 observations using the k-means and FWK-means algorithms

# Iteration	center of cluster 1			center of cluster 2		
	k-means	FWk-means		k-means	FWk-means	
		min	Alg. product		min	Alg. product
1	(402.59, 324.65)	(409.15, 319.29)	(401.81, 318.00)	(420.03, 346.94)	(428.53, 325.20)	(430.23, 325.59)
2	(377.13, 212.35)	(327.78, 276.66)	(331.19, 264.00)	(447.24, 464.73)	(507.01, 366.27)	(503.46, 375.35)
3	(402.23, 202.94)	(332.98, 243.12)	(350.39, 221.53)	(421.52, 480.18)	(491.01, 388.43)	(482.83, 410.46)
4	(413.28, 200.40)	(365.98, 200.86)	(385.27, 189.23)	(409.54, 482.93)	(466.33, 430.32)	(453.67, 445.06)
5	–	(391.02, 173.47)	(401.78, 173.74)	–	(441.49, 440.75)	(436.57, 445.58)
6	–	(409.49, 163.89)	(414.38, 166.80)	–	(426.33, 441.50)	(425.74, 443.95)
7	–	(412.85, 167.99)	(417.45, 170.47)	–	(424.13, 445.20)	(423.46, 447.28)
8	–	(415.84, 171.49)	–	–	(421.88, 448.46)	–
	# observations in cluster 1			# observations in cluster 2		
	k-means	FWk-means		k-means	FWk-means	
		min	Alg. product		min	Alg. product
1	51	52	52	49	48	48
2	52	47	47	48	53	53
3	52	48	48	48	52	52
4	52	44	45	48	56	55
5	–	42	43	–	58	57
6	–	43	44	–	57	56
7	–	44	44	–	56	56
8	–	44	–	–	56	–

# Iteration	objective function value		
	k-means	FWk-means	
		min	Alg. product
1	3454954	2527612	2438116
2	1929917	1912123	1789034
3	1798334	1731173	1511924
4	1784390	1429486	1313029
5	–	1329732	1273386
6	–	1305567	1261430
7	–	1302353	1260097
8	–	1301050	–

In evaluating the performance of both clustering algorithms, the mean execution time, the mean number of iterations to convergence, and the mean convergence rate were considered. Execution time, representing the total time required for the algorithm to reach convergence and reflecting its overall computational efficiency, was measured as the difference between the start and end times of each run in seconds. The mean number of iterations indicates the number of iterative steps needed to achieve stability or satisfy the stopping criterion, while the convergence rate reflects the speed at which the objective function approaches its optimal value during the iterative process, calculated based on the relative change of the objective function between two consecutive iterations. Since both the k-means and FWk-means algorithms are sensitive to random initializations, each algorithm was executed 30 times using different sets of initial conditions. The average execution time, number of iterations to convergence, and convergence rate over these 30 runs are reported in Table 8. Together, these three indicators provide a comprehensive assessment of the algorithms' computational efficiency, stability, and convergence behavior.

According to Table 8, the mean execution time of the FWk-means algorithm over 30 independent runs with different random initializations is 0.102 seconds for the minimum operator and 0.094 seconds for the algebraic product operator. These results indicate that, on average, the algorithm requires approximately 0.102 seconds per run with the minimum operator and 0.094 seconds per run with the algebraic product operator to reach convergence, demonstrating its satisfactory computational efficiency. The mean number of iterations to convergence is reported as 5 and 5.5 for the minimum and algebraic product operators, respectively. Thus, the algorithm typically stabilizes after approximately 5 iterations with the minimum operator and 5.5 iterations with the algebraic product operator, indicating a fast convergence speed. The mean convergence rate is 0.1732 and 0.1624 for the minimum and algebraic product operators, respectively, indicating that, on average, the algorithm approaches the optimal value by 17.32% per iteration with the minimum operator and by 16.24% per iteration with the algebraic product operator. Overall, These results highlight the favorable computational performance of the FWk-means algorithm in terms of execution speed, convergence rate, and the stability of its iterative process. In this example, it appears that the FWk-means algorithm using the minimum operator performs slightly better than the product operator in terms of convergence rate, even though its execution time is slightly higher and the number of iterations shows only minor differences.

Table 8: Performance evaluation metrics for the k-means and FWk-means algorithms

	k-means	FWk-means	
		min	Alg. product
mean execution time (in seconds)	0.077	0.102	0.094
mean number of iterations	3.43	5	5.5
mean convergence rate	0.2026	0.1732	0.1624

In this case study, a dataset of 100 observations was randomly generated from the population Ω_1 using a discrete uniform distribution in the R programming environment. Thereafter, k-means and FWk-means clustering algorithms were employed to partition the extracted data from population Ω_1 into two distinct clusters. The goal of this study is to highlight the benefits of problem formulation based on the fuzzy population theory, aiming for more accurate and fair clustering outcomes. Given the inherent vagueness of the interested population, its determination is inevitably imprecise and, to some extent, subjective. This may cause confusion for the user and potentially lead to misuse of the clustering results in the study. In the case of clustering data pertaining to the population of "intelligent youth" from a classical (crisp) perspective, the primary challenge lies in establishing a membership criterion, that is, determining which individuals, based on their age and intelligence quotient (IQ), should be regarded as members of this population within the clustering process. If we adopt a more balanced perspective on the problem, it can be stated that each individual possesses a degree of membership, ranging between 0 and 1, to the concept of being young, as well as a degree of membership, also within the same range, to the concept of being intelligent. These membership degrees play a crucial role in defining the population under consideration (interested population) and, ultimately, may influence the outcomes of the clustering process. Due to the existence of multiple and sometimes conflicting definitions of this population, classical clustering approaches, as depicted in Figure 2, may occasionally result in the exclusion of several observations, leading to information loss. On the other hand, the fuzzy approach facilitates modeling of the "intelligent youth" population by defining suitable membership functions for the variables of age and IQ. This ensures that all individuals participate in the clustering analysis to some degree. Of course the selection of appropriate membership functions may involve a degree of subjectivity. Furthermore, the degree of membership and the associated importance of each observation in the fuzzy population contribute meaningfully to the clustering computations.

6 Real-world case study: Estimation of obesity levels based on eating habits and physical condition dataset

In this section, clustering is performed on a real dataset extracted from a fuzzy population by employing the FWk-means algorithm. Real-world dataset has been used to evaluate the proposed algorithm in this study were sourced from the UCI Machine Learning Repository.

The dataset is designed to support the estimation of obesity levels among individuals from Mexico, Peru, and Colombia. It contains 2,111 instances, each characterized by 17 features related to dietary habits and physical health status. The target variable, NObesity (Obesity Level), categorizes the instances into seven classes: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. Notably, 77 percent of the dataset was synthetically generated using the SMOTE technique implemented in Weka software, whereas the remaining 23 percent was collected directly from users through an online platform.

The Body Mass Index (BMI) is an important measure used to assess the proportionality of an individual's height and weight. Using this index, one can determine whether a person is overweight or underweight. Therefore, other physical characteristics such as age, gender (female, male), ethnicity, bone structure, and so forth do not affect the calculation of BMI. The body mass index is commonly calculated as

$$BMI = \frac{weight}{(height)^2},$$

where weight should be measured in kilograms (kg) and the height in meters (m). Although the physical unit of BMI is kg/m^2, it is commonly reported as a dimensionless index. For adults aged 19 years and older, regardless of gender, the BMI classification is as follows:

Table 9: Determination of weight status based on body mass index range

body mass index range	weight status
less than 18.5	underweight
$18.5 - 24.9$	ideal weight (Normal weight)
$25 - 29.9$	overweight
$30 - 34.9$	obesity class I
$35 - 39.9$	obesity class II
40 and above	obesity class III (morbid obesity)

Body Mass Index is a simple measure that reflects only our weight status. Although it provides an approximate indication of body fat, it neither captures fat distribution nor distinguishes between fat and muscle mass. For instance, a professional athlete with substantial muscle mass and low body fat may have a BMI of 30, which would incorrectly classify the individual as obese. In reality, athletes typically have a higher proportion of muscle mass, and such a classification would be misleading, as they are not actually obese. Moreover, BMI-based classification is not reliable for certain groups, including pregnant women, professional athletes, individuals with severe illnesses, and those with physical disabilities.

In this section, the fuzzy population representing "obese individuals" is modeled based on the body mass index, with the membership function defined as follows (see Figure 7)

$$\mu(BMI) = \begin{cases} 0 & BMI < 18.5, \\ \dfrac{BMI - 18.5}{21.5} & 18.5 \leq BMI < 40, \\ 1 & BMI \geq 40. \end{cases}$$

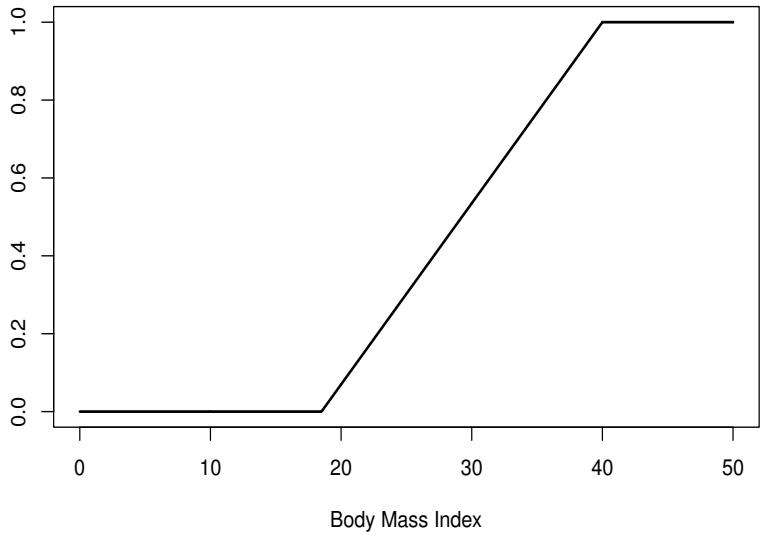


Figure 7: The membership function of fat based on BMI

Table 10 presents the values of the validity indices for the FWk-means algorithm across different numbers of clusters. In a "fuzzy population of obese individuals", each observation is associated with a membership degree between 0 and 1. As shown in Section 4, the membership degree of each observation influences the computation of the clustering quality indices. Therefore, in a fuzzy population, a relative comparison of the indices across different numbers of clusters provides a more appropriate and reliable assessment of clustering performance. According to Table 10, the FWSSE index decreases monotonically with increasing k, indicating a reduction in within-cluster variance and an improvement in cluster compactness. Although $k = 3$ corresponds to the lowest $FWDBI$ and a near-maximum $FWDI$, but the $FWSSE$ has not yet reached relative stability and increasing k up to 5 results in a notable reduction in $FWSSE$. Moreover, the corresponding $FWDBI$ and $FWDI$ values at $k = 5$ are relatively favorable, representing, respectively, one of the lower and one of the higher values among the tested cluster numbers.

Table 10: FWk-means clustering quality indices for varying numbers of clusters (k)

k	$FWSSE$	$FWDBI$	$FWDI$
2	154652.15	1.513	0.500
3	75883.82	1.488	0.432
4	52690.74	2.433	0.346
5	30638.30	2.560	0.286
6	24877.51	2.588	0.266
7	19427.93	2.590	0.214
8	17699.19	2.551	0.168
9	17208.62	2.502	0.106
10	16107.46	2.705	0.100

In this example, the optimal value of k is determined using the Elbow method. In Figure 8, the FWSSE values are plotted against different numbers of clusters. According to the figure the FWSSE values decreases sharply up to $k = 5$ and then becomes almost stable. Beyond $k = 5$, the rate of decrease becomes significantly slower, corresponding to the elbow point. This point indicates that adding more clusters does not result in a substantial improvement in data compactness. Therefore, $k = 5$ is selected as the optimal number of clusters.

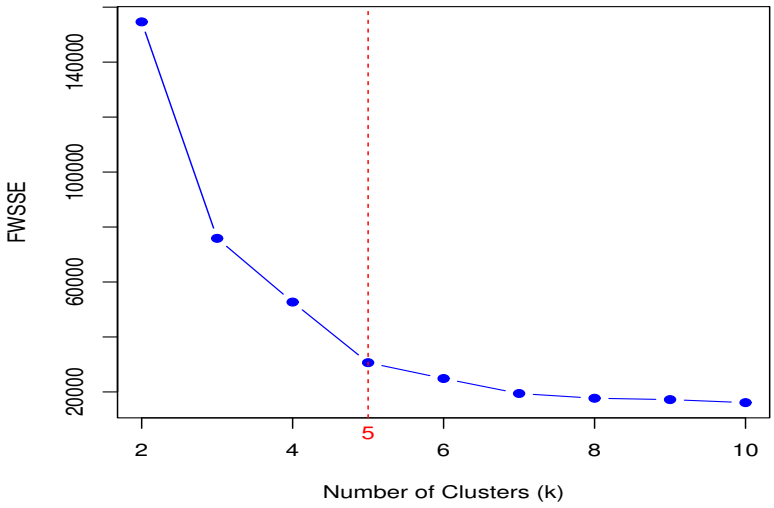


Figure 8: Elbow method for determining optimal k

After modeling the interested fuzzy population representing "obese individuals", and selecting the optimal number of clusters, k, using the Elbow method, $n = 2111$ extracted data from this fuzzy population will be clustered based on two features height and weight into $k = 5$ clusters. Here $p = 1$ denote the number of the initial fuzzy populations. The clustering results of the obesity dataset using the FWK-means algorithm into five clusters are shown in Figure 9.

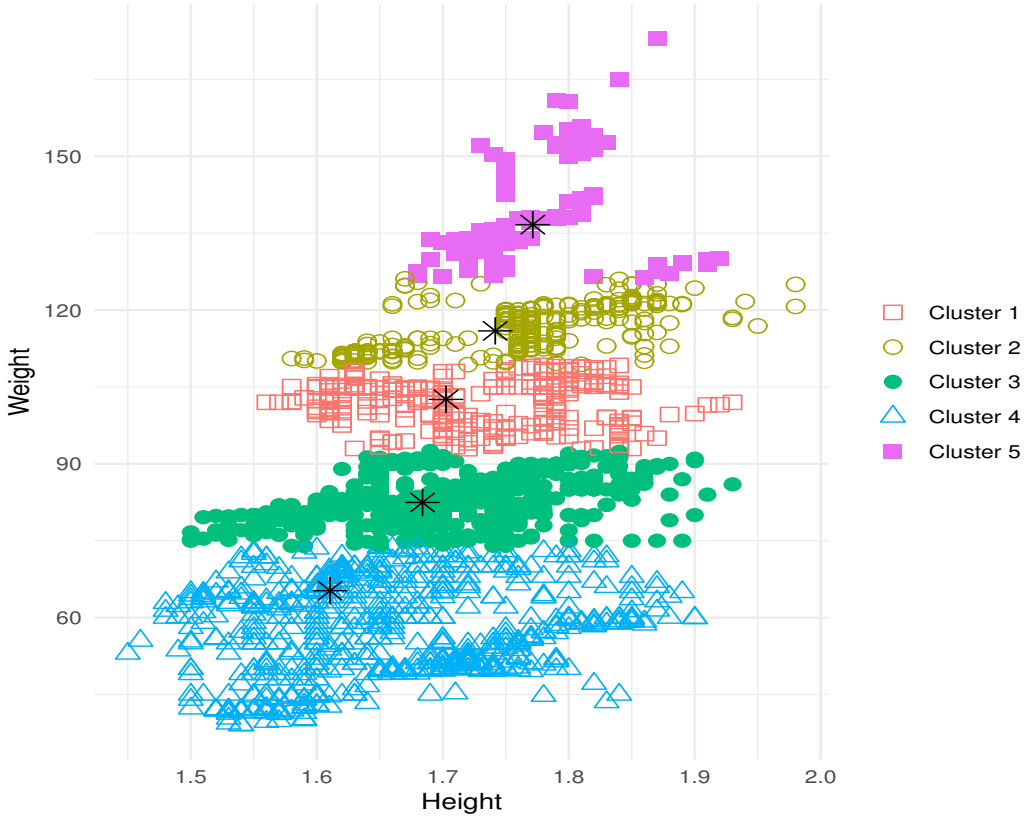


Figure 9: Clustering of the obesity dataset using the FWK-means algorithm

Table 11: Clustering results of the obesity dataset using the FWK-means algorithm

Cluster #	Cluster centers	The number of observations in last iteration
1	(1.70, 102.56)	322
2	(1.74, 115.92)	344
3	(1.68, 82.46)	595
4	(1.61, 65.24)	697
5	(1.77, 136.65)	153

The cluster centers and the corresponding number of observations in each cluster, obtained at the final iteration of the FWK-means algorithm, are reported in Table 11. In this example, the implementation of the FWk-means algorithm on a real-world dataset related to obesity is investigated, and the clustering performance of the proposed FWk-means algorithm is analyzed based on the internal evaluation indicis described in Section 4. First, the "fuzzy population of obese individuals" was modeled by defining an appropriate membership function for the interested fuzzy population. Then, clustering of the data extracted from this fuzzy population was performed using the proposed algorithm, resulting in five distinct clusters. The dataset included labeled data, with individuals categorized into classes such as Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. Since the concept of obesity is inherently vague and fuzzy, categorical labels such as "Overweight Level I" or "Obesity Type III" do not provide a precise or intuitive understanding of individuals' conditions. Modeling the fuzzy population of obesity through a suitable membership function provides a more nuanced and meaningful interpretation. For example, expressing that an individual belongs to the "fuzzy population of obese individuals" with a membership degree of 0.3, while another has a membership degree of 1, is considerably more significant and fairer than simply classifying one individual as underweight and another as having Type III obesity. Even a professional athlete with significant muscle mass and very low body fat may have a body mass index of 30, the BMI formula categorizes this individual as obese, while, the membership degree $\mu(30) = 0.53$ indicates a moderate degree of membership in the "fuzzy population of obese individuals". Since membership degrees vary among individuals, with each person belonging to the "fuzzy population of obese individuals" to a degree between 0 and 1, considering these membership degrees into

the clustering process and the computation of clustering validity indices is both reasonable and appropriate. Although classical clustering validity indices have well-defined interpretive ranges, such ranges are not meaningful in this study due to the influence of membership degrees on their computation. Therefore, the clustering quality indices were computed for various values of k, and their interpretation was performed relatively, depending on the specific goals of the clustering analysis.

7 Stability analysis of the FWK-means algorithm in the presence of outliers

In this section, a numerical example is presented to assess the stability of the FWK-means algorithm in the presence of outliers. The aim is to investigate how the cluster centers estimated by FWK-means are influenced by both the spatial positions of outliers (i.e., whether they are close to or far from other observations) and their membership degrees within the fuzzy population. To evaluate the algorithm's sensitivity to outliers, a two-dimensional synthetic dataset is generated, comprising three well-separated clusters with distinct dispersion characteristics:

- Cluster 1: normally distributed with mean 5 and variance 1.2;
- Cluster 2: normally distributed with mean 10 and variance 1;
- Cluster 3: normally distributed with mean 0 and variance 1.

The sample sizes for Clusters 1, 2, and 3 are set to 120, 110, and 100, respectively. This setup ensures that the clusters are well-separated, providing a clear basis for evaluating the effect of outliers on the stability of cluster centers and the overall clustering structure.

Three distinct scenarios were designed to investigate the effects of outlier positions and their membership degrees within the fuzzy population. Here, the membership degree of an observation indicates its degree of belonging in the fuzzy population rather than to a specific cluster. In all scenarios, data were generated in a two-dimensional space and comprised the three primary clusters described previously.

Scenario 1: Outliers far from the clusters with equal and high membership degrees. All observations in the main clusters were assigned a membership degree of 1, effectively reducing the fuzzy population to a classical (crisp) population. Twelve outliers were generated well away from the main clusters, drawn from two normal distributions with means -10 and 20 and unit variance in both dimensions. These outliers were assigned the same membership degree of 1 as the main-cluster observations, ensuring consistency with the crisp cluster structure.

Scenario 2: Outliers far from the clusters with small and varying membership degrees. The membership degrees of observations in the three main clusters were independently sampled from a uniform distribution over [0.2, 1], reflecting their membership in the fuzzy population. As in Scenario 1, twelve outliers were generated well away from the main clusters, drawn from normal distributions with means -10 and 20 and unit variance in both dimensions. Their membership degrees were independently sampled from a uniform distribution over [0.3, 1], also reflecting their membership in the fuzzy population.

Scenario 3: Outliers near the clusters with small and varying membership degrees. The membership degrees of observations in the three main clusters were independently sampled from a uniform distribution over [0.2, 1], reflecting their membership in the fuzzy population. Unlike the previous scenarios, twelve outliers were placed close to Clusters 2 and 3. These outliers were generated from normal distributions with means 10.5 and 0.5 and variances 0.3 and 0.2, respectively, with their membership degrees independently sampled from a uniform distribution over [0.3, 1] within the fuzzy population.

First, the FWK-means algorithm was applied to the observations in the main clusters of the fuzzy population to determine the initial cluster centers. Then, after adding the outliers, the algorithm was rerun to compute the updated cluster centers. To enable consistent comparisons across scenarios, the following terms are defined and evaluated:

Percentage of Points Shift per Cluster. This measure describes the stability of each cluster by indicating the proportion of its original points that reassigned cluster assignments after the inclusion of outliers. Higher percentages indicate greater sensitivity to outliers, whereas lower percentages suggest more robust cluster structures. It is calculated as follows:

$$\text{Points Shift per Cluster } (\%) = \frac{\text{Reassigned points after outlier inclusion}}{\text{Total points in the cluster}} \times 100.$$

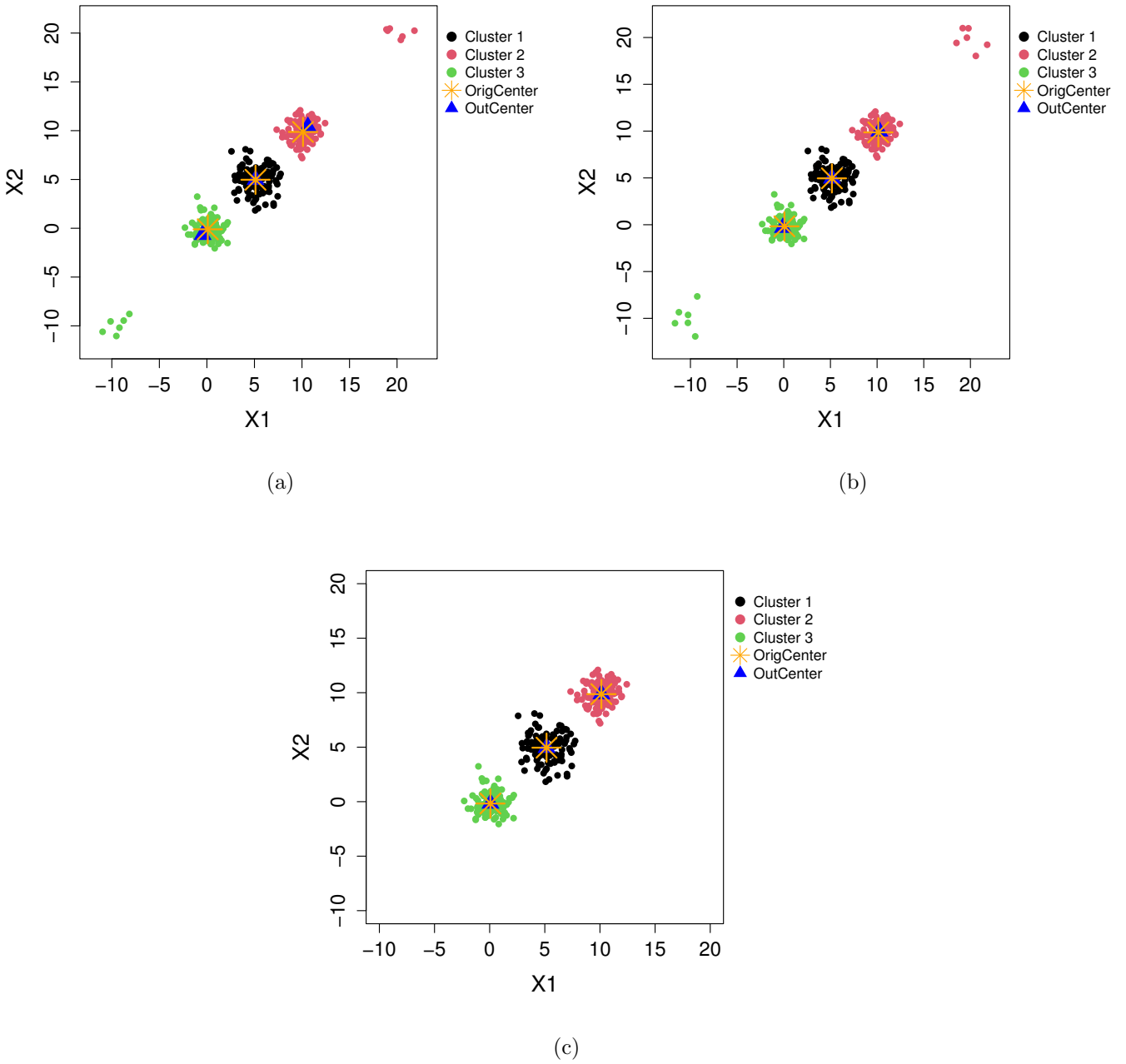
Cluster Radius without Outliers. This is calculated as the average Euclidean distance between each point and the center of its assigned cluster before outliers are included.
Cluster Radius with Outliers. This is calculated as the average Euclidean distance between each point and the center of its assigned cluster after outliers are included.
Percentage of Cluster Center Shift. This measures the impact of outliers on clusters with varying dispersion and is calculated as follows:

$$\text{Cluster Center Shift (\%)} = \frac{\text{Euclidean distance between cluster centers}}{\text{Cluster radius without outliers}} \times 100.$$

Table 12: Performance results of the FWK-means algorithm in the presence of outliers across three scenarios

		Cluster 1	Cluster 2	Cluster 3
Scenario 1	centers without outliers	(5.11, 4.97)	(10.10, 9.87)	(0.09, -0.10)
	centers with outliers	(5.11, 4.97)	(10.62, 10.39)	(-0.44, -0.66)
	cluster size without outliers	120	110	100
	cluster size with outliers	120	116	106
	points shift per cluster (%)	0	0	0
	cluster radius without outliers	1.486	1.235	1.178
	cluster radius with outliers	1.486	1.979	2.009
	Euclidean distance between centers	0	0.73	0.77
	cluster center shift (%)	0	59.38	65.85
Scenario 2	centers without outliers	(5.15, 4.95)	(10.13, 9.86)	(0.06, -0.17)
	centers with outliers	(5.15, 4.95)	(10.24, 9.98)	(-0.09, -0.31)
	cluster size without outliers	120	110	100
	cluster size with outliers	120	116	106
	points shift per cluster (%)	0	0	0
	cluster radius without outliers	1.488	1.235	1.181
	cluster radius with outliers	1.488	1.885	1.931
	Euclidean distance between centers	0	0.16	0.21
	cluster center shift (%)	0	13.08	18.37
Scenario 3	centers without outliers	(5.15, 4.95)	(10.13, 9.86)	(0.06, -0.17)
	centers with outliers	(5.15, 4.95)	(10.13, 9.87)	(0.07, -0.15)
	cluster size without outliers	120	110	100
	cluster size with outliers	120	116	106
	points shift per cluster (%)	0	0	0
	cluster radius without outliers	1.488	1.235	1.181
	cluster radius with outliers	1.488	1.209	1.157
	Euclidean distance between centers	0	0.007	0.011
	cluster center shift (%)	0	0.60	0.97

Table 12 summarizes the performance of the FWK-means algorithm in the presence of outliers across three scenarios. In this table, *centers before outliers* refers to the initial positions of the cluster centers prior to the inclusion of outliers, while *centers after outliers* indicates their positions after the outliers have been added. *Cluster size before outliers* and *cluster size after outliers* denote the number of points in each cluster before and after the inclusion of outliers, respectively. *Points shift per cluster (%)* shows the percentage of points that changed clusters. *Cluster radius without outliers* and *cluster radius with outliers* correspond to the radius of each cluster before and after the outliers, respectively. *Euclidean distance between centers* reports the Euclidean distance between the cluster centers following the outlier perturbation, and *cluster center shift (%)* represents the percentage displacement of the cluster centers before and after the inclusion of outliers.



(a)

(b)

(c)

Figure 10: Clustering results of the FWK-means algorithm in the presence of outliers for the three scenarios: (a) Scenario 1, (b) Scenario 2, and (c) Scenario 3.

According to Table 12, despite the presence of outliers, the overall clustering structure remained largely intact across all three scenarios, with no data points reassigned between the original clusters. Outliers were assigned to the nearest cluster center. The radius of Cluster 1 remained unchanged in all scenarios; however, after the addition of outliers, the radii of Clusters 2 and 3 increased in Scenarios 1 and 2, but decreased in Scenario 3. Cluster centers shifted to varying degrees under the influence of outliers. In Scenario 1, outliers with high and equal membership degrees (1) were added at locations far from the cluster centers, resulting in substantial shifts for Clusters 2 and 3, by 59.38% and 65.85%, respectively, while Cluster 1 remained unaffected. In Scenario 2, outliers with low and varying membership degrees were added at the same locations as in Scenario 1, leading to smaller shifts of 13.08% and 18.37% for Clusters 2 and 3, indicating greater stability. In Scenario 3, outliers with similarly low membership degrees as in Scenario 2 were added near the cluster centers, producing minimal displacements of 0%, 0.60%, and 0.97% for Clusters 1, 2, and 3, respectively.

These results demonstrate that both the positions of outliers and their membership degrees within the fuzzy population significantly affect the stability of FWK-means cluster centers in the presence of outliers. The algorithm shows greater robustness when outliers are located near cluster centers and have low membership values.

Figure 10 illustrates the clustering results of the FWK-means algorithm in the presence of outliers across three scenarios. In all scenarios, the overall clustering structure remains unchanged, with all points staying within their original clusters; however, the positions of the cluster centers are affected by the outliers. In Scenario 1 (Subfigure a), outliers with high membership degrees positioned far from the main clusters cause substantial shifts in the centers of Clusters 2 and 3, while Cluster 1 remains largely unaffected. In Scenario 2 (Subfigure b), although the outliers are still distant, their lower membership degrees result in smaller shifts of the cluster centers. In Scenario 3 (Subfigure c), outliers with low membership degrees located near the main clusters produce minimal displacement of the centers. Since the cluster centers are displaced in the presence of outliers, if the centers are close to each other, the overall clustering structure may no longer be preserved. Overall, these results indicate that the spatial positions of outliers and their membership degrees in the fuzzy population critically influence the displacement of FWK-means cluster centers, while the clustering structure itself remains stable.

This numerical example demonstrates that the stability of the FWK-means algorithm in the presence of outliers depends on both the membership degrees of observations in the fuzzy population and the positions of the outliers. Outliers with high membership degrees located far from the main observations have the most pronounced negative impact on cluster centers. In contrast, distant outliers with low membership degrees cause smaller displacements of the centers, while outliers positioned near the main observations with low membership degrees further reduce these shifts. Therefore, in practical applications, considering both the membership degrees and the positions of outliers is essential for effectively identifying outliers within fuzzy data.

8 Conclusions and future works

This paper presents an extension of the classical k-means algorithm specifically designed for clustering data extracted from fuzzy populations. We first reviewed the concept of a fuzzy population and provided several examples to facilitate a clearer understanding of a fuzzy population. In a classical population, each observation has a membership degree of either zero or one to the population. In contrast, within a fuzzy population, every observation is associated with a membership degree ranging continuously between zero and one to the interested fuzzy population. Practical examples of fuzzy population include blood pressure measurements within "fuzzy population of obese individuals", blood sugar levels among "elderly fuzzy population", and energy consumption in "fuzzy population of high-usage households". One of the fundamental characteristics of the fuzzy population is the absence of well-defined boundaries in some of the defining variables. Consequently, during the clustering process, analysts often select subsets of observations based on their subjective interpretation of the fuzzy population and their personal criteria. On the other hand, assigning equal importance to all observations is also inappropriate and problematic due to the inherently fuzzy nature of the population. To address these challenges, a new clustering approach, called the FWk-means method, has been introduced, in which the membership degree of each observation in the fuzzy population is considered into the clustering computations. The proposed algorithm is a generalized version of the classical k-means algorithm. To evaluate its performance, the clustering validity indices were extended from the classical population framework to a fuzzy population, and these indices account for the membership degree of each observation in their computation. Consequently, the classical interpretive ranges of these indices are no longer meaningful for assessing clustering quality in a fuzzy population. Therefore, the performance of the proposed algorithm was evaluated across different values of k in a relative manner by comparing lower or higher index values to assess clustering quality within the fuzzy population framework. These findings demonstrate that the proposed method effectively handles outliers when clusters are well-separated, preserving the overall clustering structure and confirming its robustness. A notable limitation is that outliers can cause slight shifts in cluster centers, which may compromise the algorithm's stability when the centers are close to each other.

For future research, the proposed algorithm can be extended to fuzzy clustering of data extracted from a fuzzy population. According to the method presented in this paper, the assignment of observations to clusters is performed deterministically using binary values (zero or one), meaning that each observation belongs to exactly one cluster. By generalizing the FWk-means algorithm to a fuzzy clustering framework, it becomes possible to assign observations to clusters in a soft manner, with membership degrees ranging between zero and one, allowing each observation to simultaneously belong to multiple clusters with different membership degrees.

Acknowledgements

The authors would like to express their gratitude to the anonymous referees for their valuable comments, which greatly improved the quality of this paper.

References

[1] A. A. Ross, K. Nandakumar, A. K. Jain, *Handbook of multibiometrics*, International Series on Biometrics, Springer, Boston, MA, USA, (2019). https://doi.org/10.1007/978-0-387-71041-9

[2] A. Gondeau, B. Gosselin, *Object weighting: A new clustering approach to deal with heterogeneous data*, Scientific Reports, **11**(1) (2021), 1-13. https://doi.org/10.1038/s41598-021-86561-8

[3] A. Gondeau, Z. Aouabed, M. Hijri, P. Peres-Neto, V. Makarenkov, *Object weighting: A new clustering approach to deal with outliers and cluster overlap in computational biology*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, **18**(3) (2021), 633-641. https://doi.org/10.1109/TCBB.2019.2930177

[4] A. Jain, A. Sharma, *Membership function formulation methods for fuzzy logic systems: A comprehensive review*, Journal of Critical Reviews, **7**(19) (2020), 8717-8733. https://doi.org/10.31838/jcr.07.19.999

[5] A. K. Jain, R. P. W. Duin, J. Mao, *Statistical pattern recognition*, Wiley, New York, USA, (2000). https://doi.org/10.1109/34.824819

[6] A. Parchami, P. D'Urso, S. M. Taheri, *On some descriptive statistics based on fuzzy population*, Soft Computing, Revised, (2025).

[7] B. Mirkin, *Clustering for data mining: A data recovery approach*, Chapman and Hall/CRC, Boca Raton, (2005). https://doi.org/10.1201/9781420034899

[8] C. C. Aggarwal, C. K. Reddy, *Data clustering: Algorithms and applications*, Data Mining and Knowledge Discovery Series, Chapman and Hall/CRC, Boca Raton, (2014). https://doi.org/10.1201/9781315373515

[9] D. Zhang, L. Wang, H. Wang, C. Shi, *Weighted fuzzy clustering with sample and feature weighting for high-dimensional data*, Knowledge-Based Systems, **196** (2020), Article 105789. https://doi.org/10.1016/j.knosys.2020.105789

[10] H. J. Zimmermann, *Fuzzy set theory and its applications*, 4th ed., Springer, Heidelberg, 2001. https://doi.org/10.1007/978-94-010-0646-0

[11] I. D. Borlea, R. E. Precup, F. Drčgan, A. B. Borlea, *Centroid update approach to k-means clustering*, Advances in Electrical and Computer Engineering, **17**(4) (2017), 3-10. https://doi.org/10.4316/AECE.2017.04001

[12] J. Mohammadi, S. M. Taheri, *Pedomodels fitting with fuzzy least squares regression*, Iranian Journal of Fuzzy Systems, **1**(1) (2004), 45-61. https://doi.org/10.22111/ijfs.2004.28

[13] J. Ning, *Neural network-based pattern recognition in the framework of edge computing*, Romanian Journal of Information Science and Technology, **27**(1) (2024), 106-119. https://doi.org/10.24189/ijis.2024.011

[14] J. R. Zubizarreta, E. A. Stuart, D. S. Small, P. R. Rosenbaum, *Handbook of matching and weighting adjustments for causal inference*, CRC Press, (2023).

[15] K. A. Linderman, J. B. Sweeney, M. C. Ulrich, W. Li, Q. Song, X. Yang, *K-means clustering of overweight and obese population using quantile-transformed metabolic data*, Journal of Biomedical Informatics, **109** (2020), Article 103511. https://doi.org/10.1016/j.jbi.2020.103511

[16] K. Kerdprasop, N. Kerdprasop, P. Sattayatham, *Weighted K-means for density-biased clustering*, Data Warehousing and Knowledge Discovery, Lecture Notes in Computer Science, Springer, BerlinHeidelberg, **3589** (2005), 488-497. https://doi.org/10.1007/11563983-54

[17] K. M. Borgwardt, H. P. Kriegel, M. Pfeifle, *Balanced k-means for clustering with size constraints*, Proceedings of the 2013 SIAM International Conference on Data Mining, SIAM, (2013), 34-42. https://doi.org/10.1137/1.9781611972832.4

[18] K. P. Murphy, *Probabilistic machine learning: An introduction*, MIT Press, Cambridge, MA, (2022). https://doi.org/10.7551/mitpress/13467.001.0001

[19] M. A. Mohammadi, M. Taheri, *Membership function formulation methods in fuzzy systems: A systematic review*, Artificial Intelligence Review, **57** (2024), 145. https://doi.org/10.1007/s10462-023-10574-3

[20] M. Eftekhari, A. Mehrpooya, F. Saberi-Movahed, V. Torra, *How fuzzy concepts contribute to machine learning*, Studies in Fuzziness and Soft Computing, Springer, **416**, (2022). https://doi.org/10.1007/978-3-030-74073-5

[21] M. Huang, S. Wang, L. Zhang, *Feature-weighted k-means clustering with adaptive local structure learning*, Pattern Recognition, **112** (2021), Article 107728. https://doi.org/10.1016/j.patcog.2020.107728

[22] M. Namdari, J. H. Yoon, A. R. Abadi, S. M. Taheri, S. H. Choi, *Fuzzy logistic regression with least absolute deviations estimators*, Soft Computing, **19** (2014), 909-917. https://doi.org/10.1007/s00500-014-1261-9

[23] P. N. Tan, M. Steinbach, A. Karpatne, V. Kumar, *Introduction to data mining*, 2nd ed., Pearson, (2018).

[24] R. C. Amorim, *A survey on feature weighting based k-means algorithms*, Journal of Classification, **33**(2) (2016), 210-242. https://doi.org/10.1007/s00357-016-9208-4

[25] S. R. Mahdavi, M. Mousakhani, N. Yaghoobi, *Automatic membership function generation using clustering methods for fuzzy systems*, Applied Soft Computing, **90** (2020), Article 106176. https://doi.org/10.1016/j.asoc.2020.106176

[26] T. Hafs, H. Zehir, A. Hafs, H. Brahmia, A. Nait-Ali, *Enhancing recognition in multimodal biometric systems: Score normalization and fusion of online signatures and fingerprints*, Romanian Journal of Information Science and Technology, **27**(1) (2024), 37-49. https://doi.org/10.24189/ijis.2024.003

[27] V. P. Singh, A. K. Mishra, *Clustering diabetic patients based on their healthcare service utilization patterns*, ResearchGate, (2021).

[28] Y. Xu, C. Wang, H. Zhao, *Adaptive feature weighting for high-dimensional clustering via sparse regularization*, IEEE Transactions on Neural Networks and Learning Systems, **33**(9) (2022), 4713-4726. https://doi.org/10.1109/TNNLS.2021.3066063

[29] Z. Ning, J. H. Chen, J. Huang, U. J. Sabo, Z. Yuan, Z. Dai, *WeDIV – An improved k-means clustering algorithm with a weighted distance and a novel internal validation index*, Information Sciences, **603** (2022), 242-259. https://doi.org/10.1016/j.ins.2022.04.078