RESEARCH

Parameter Estimation Based on Validity-Aware Gustafson Kessel Clustering

Yeşim Akbaş · Serkan Akbaş · Türkan Erbay Dalkılıç

Received: 30 July 2025 / Revised: 20 October 2025 / Accepted: 26 December 2025
© The Author(s) 2026

Abstract
Regression analysis estimates model parameters to describe the relationship between a dependent variable and one or more independent variables. While classical regression assumes homogeneity among observations, real-world data often exhibit heterogeneous structures. In such cases, fuzzy logic–based approaches offer an alternative for estimating the unknown parameters of regression models. However, effective results for parameter estimation cannot be obtained for ellipsoidal scattering data in these approaches either. In this study, a method based on the Gustafson-Kessel (GK) clustering algorithm has been proposed to overcome these disadvantages. Since the proposed method aims to perform parameter estimation especially on ellipsoidal data scatters, Davies-Bouldin Index (DBI), one of the clustering validity indices that provide effective results in determining the optimal number of clusters, has been adapted using the Mahalanobis distance. GK clustering algorithm was applied to obtain the membership degrees that determine the contributions of the observations in the identified clusters to the regression model to be constructed. To evaluate the performance of the proposed algorithm, applications were conducted on various datasets, and the results obtained were compared with estimates based on the Fuzzy C-Means (FCM) and Type-2 FCM clustering methods.

Keywords Parameter estimation · Clustering validity · Fuzzy clustering · Gustafson Kessel clustering algorithm

1 Introduction

Regression analysis is an analysis method that allows model coefficients to be estimated using statistical methods in order to mathematically express the relationship between the dependent variable and one or more independent variables. In classical regression analysis, it is assumed that the data comes from a single class, that is, a homogeneous main mass, while in practice the data obtained often do not meet this assumption; they consist of subgroups with different characteristics, multiple structures and heterogeneous distributions. This situation can cause bias in the parameter estimates of classical regression analyses and therefore high errors. In such cases, the use of alternative methods becomes necessary. Fuzzy clustering-based analysis methods offer effective solutions in cases where the data are not homogeneous, the observations belong to different structures and the class boundaries cannot be determined with clear lines. Instead of assuming that each observation belongs to only

Published online: 18 January 2026

one cluster, the fuzzy logic approach assumes that each observation can belong to different clusters with certain membership degrees. One of the important steps in the process of parameter estimation based on fuzzy logic is to determine the clusters in the data set and to obtain the membership degrees of each observation to these clusters. The obtained membership degrees can be used as weights in the regression model to determine the extent to which each observation will contribute to the estimation process. Thus, unlike classical methods, instead of giving equal weight to all observations, the effect of each observation on the model can be dynamically changed according to the structure it belongs to.

Benatti et al., in their work, theoretically examined center-based fuzzy clustering methods and discussed how some constraints could be integrated into the clustering process in addition to classical models. They proposed a more general constrained fuzzy clustering model by assigning weights to each data point and predetermining the total weight of the clusters [1]. In his work, Wu used fuzzy input/output data to obtain alpha level least squares estimate from a linear regression model and then generated fuzzy estimates of the regression parameters [2]. Kula and Dalkılıç proposed an algorithm for parameter estimation in the case where Type-2 fuzzy sets characterized by Gaussian membership function are considered as independent variables [3]. Ahmadini presented his proposed Intuitionistic Fuzzy Logistic Regression model on the birth weight dataset by using the Tanaka model [4] in estimating uncertain parameters that contain uncertainty [5]. In their study, Chou et al., estimated the parameters of the effective variables for Taiwan's air cargo volume using a fuzzy linear regression model with triangular fuzzy numbers, based on the model proposed by Tanaka, assuming that the error terms are caused by uncertain parameters [6]. Jung et al., in their work, proposed computational formulas of the likelihood functions for parameter estimation based on both triangular and trapezoidal fuzzy data [7]. In their study, Dalkılıç and Sağırkaya estimated the parameters using Type-1 and Type-2 FCM algorithms and then compared them according to the difference between the estimated values and the observed values [8]. Enke et al. presented a hybrid model combining multiple regression analysis, fuzzy Type-2 clustering and fuzzy Type-2 neural network to predict stock market prices [9]. In their work, Zarandi et al. proposed a new Type-2 fuzzy c-regression clustering algorithm for the structure identification phase of Takagi–Sugeno systems [10]. Mokhtari et al. presented a novel fuzzy least squares estimation method based on extended Euclidean distance to develop regression models in full-interval Type-2 fuzzy environments [11]. In his research, Shi proposed an improved clustering algorithm based on the Type-2 fuzzy c-regression model for use in fuzzy model identification and demonstrated that this algorithm achieved high success in various simulations [12]. For cases in which the data designated for estimating regression model parameters are aggregated around multiple centroids, several cluster validity indices proposed in the literature are employed during the decision-making phase for the optimal number of clusters. DBI is a widely preferred index among cluster validity measures and allows the determination of the ideal number of clusters by taking into account the similarity between clusters and the dispersion within clusters [13–16].

In this study, DBI was used to determine the optimum number of clusters in the data set. In the proposed method, the DBI, which traditionally uses the Euclidean distance in the literature when determining the optimum number of clusters, was redesigned using the Mahalanobis distance for ellipsoid-scattered data and the obtained DBI value was used. To demonstrate the greater effectiveness of the DBI value calculated using Mahalanobis distance in determining the number of clusters, DBI values based on Euclidean distance were also computed. The optimal number of clusters was determined by comparing the DBI values based on both distance measures. Depending on the number of clusters determined by DBI, the GK clustering algorithm was applied to the data. The GK algorithm defines separate covariance matrices for each cluster, allowing clusters to form elliptical, unlike classical methods.

An algorithm capable of efficient parameter estimation has been designed, which can be used as an alternative to classical regression methods when the dataset is clustered in an ellipsoid scatter around multiple centers. In this algorithm, the membership degrees obtained from the GK clustering algorithm are used as weights in the estimation process of the unknown parameters of the regression model. This proposed method was tested on four different data sets. For each example, regression models were obtained for the clusters formed by operating the proposed method, the existing FCM algorithm available in the literature, and the Type-2 FCM method using code

written in the MATLAB R2023b. The significance of the models was tested with ANOVA, and the error amounts and coefficients of determination for all models were calculated and compared to demonstrate the effectiveness of the proposed model. The comparisons revealed that the proposed approach yielded more successful and low-error results, especially in heterogeneous data.

The organization of this article is as follows. In Sect. 2, the fuzzy clustering approach and clustering validity indexes are included. In Sect. 3, the algorithm proposed in this study is presented. Section 4 is devoted to the comparative analysis of the proposed algorithm with the FCM and Type-2 FCM approaches used in the literature on four different examples. Results and conclusions are presented in Sect. 5.

2 Fuzzy Clustering Methods

Clustering algorithms are used to create groups by bringing together data points with similar characteristics [17]. Additionally, these algorithms function as a data analysis method to identify similar data, leading to more meaningful results [18]. If the membership of the data to existing clusters is uncertain, fuzzy logic-based clustering algorithms may be preferred. Fuzzy clustering is one of the most used clustering methods owing to its flexible assignments of data to clusters [19, 20]. After the concept of fuzzy set was introduced by Zadeh in 1965 [21], explaining the quantification of features with the degree of functions in fuzzy sets, Bellman and Zadeh continued their studies on decision making in fuzzy environments [22, 23].

The FCM algorithm, proposed by Dunn [24] and developed by Bezdek [25], has taken its place as a widely used fuzzy clustering algorithm in literature. The FCM algorithm utilizes the Euclidean distance measurement to separate a p-dimensional dataset into clusters based on a predetermined number of centroids [26–28]. It's important to note that FCM relies on Euclidean distances and is confined to identifying clusters with uniform spherical shapes. However, the GK clustering algorithm, as proposed by Gustafson and Kessel, offers an alternative approach without such limitations [29]. The GK algorithm overcomes the shortcoming of the FCM algorithm by employing the Mahalanobis distance in the objective function while performing distance calculations and becomes effective in recognizing ellipsoidal clusters with different shape and orientation [30]. Furthermore, as indicated in the study by Abdullah et al., the GK algorithm outperforms Euclidean distance-based methods by considering the shape of the data [31]. Miller et al., highlights the superiority of GK clustering over FCM in handling various sub-cluster types [32]. Wang et al. demonstrates that the GK algorithm achieves higher classification accuracy compared to FCM [33]. Furthermore, Hadiloo et al. suggests that GK clustering surpasses FCM in delineating channel patterns [34]. Lastly, according to Ashish et al., all analysis results in their study indicate that GK clustering is a sensible choice for generating change detection maps compared to FCM clustering [35].

The GK-based clustering approach is frequently used in solving real-world problems in fields such as finance, biology, medicine and image processing. In the work of Peng et al., the GK method was successfully used to classify tissue denaturation with features extracted from improved multiscale permutation entropy [36]. Zheng et al. used the GK in the Noice Clustering GK algorithm developed to separate spectral data into clusters according to different pesticide levels and provided higher accuracy than other methods [37]. In their study, Dalkılıç et al. used GK as a pre-selection tool to determine the stocks to be included in the portfolio, allowing the investor to choose from more homogeneous and meaningful groups [38]. Markiewicz and Dziekiewicz achieved high accuracy by using modified GK to cluster tortuosity and lumen measurements of arterial segments and to classify patients into appropriate endovascular intervention groups [39]. In their study, Selamat and Ismail used GK together with user feedback in their content-based image retrieval system, resulting in better clustering of images and thus improving the quality of search results [40]. Podenok and Sadykhov used GK to distinguish terrain types in multispectral satellite images more accurately and clearly [41]. The objective function of the GK clustering algorithm is as given by;

$$J\left(\mu, V, \{A_i\}\right) = \sum_{j=1}^{C} \sum_{i=1}^{N} \left(\mu_{ij}\right)^m D^2\left(x_i, v_j\right), \tag{1}$$

where, m denotes the fuzziness index, x_i is the i^{th} observation value, v_j is the center of j^{th} cluster, $D\left(x_i, v_j\right)$ is the Mahalanobis distance between x_i andv_j. The partition matrix $\mu = [\mu_{ij}]_{C\times N}, j = 1, ..., C$ and $i = 1, ..., N$, where μ_{ij} is the membership of pattern x_i to cluster$j$.

Step 1 Initial membership degrees $\mu^0 = [\mu_{ij}]$ are determined randomly from a uniform distribution in the range [0,1].

Step 2 Fuzzy cluster centers are calculated for each cluster as in Eq. (2).

$$v_j = \frac{\sum_{i=1}^{N} \left(\mu_{ij}\right)^m x_i}{\sum_{i=1}^{N} \left(\mu_{ij}\right)^m}, j = 1, \ldots, C \tag{2}$$

Step 3 The fuzzy covariance matrix (F_j), is calculated for each cluster using Eq. (3).

$$F_j = \frac{\sum_{i=1}^{N} \left(\mu_{ij}\right)^m \left(x_i - v_j\right)\left(x_i - v_j\right)^T}{\sum_{i=1}^{N} \left(\mu_{ij}\right)^m}, j = 1, \ldots, C \tag{3}$$

Step 4 Mahalanobis distances for each data are calculated with Eq. (4).

$$D^2\left(x_i, v_j\right) = \left(x_i - v_j\right)^T A_j (x_i - v_j) \quad i = 1, ..., N, j = 1, \ldots, C \tag{4}$$

Here;

$$A_j = \sqrt{det\left(F_j\right)} * F_j^{-1}, j = 1, \ldots, C \tag{5}$$

Step 5 The membership matrix is updated with Eq. (6).

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{D(x_i, v_j)}{D(x_i, v_k)}\right)^{\frac{2}{m-1}}}, j = 1, \ldots, C i = 1, \ldots, N \tag{6}$$

Step 6 Updated membership degrees are compared with previous membership degrees and the process of obtaining optimal membership degrees is stopped when the condition $||\mu^t - \mu^{t-1}|| < \varepsilon$ is provided, where ε is the termination tolerance.

The primary goal of clustering analysis is to group similar elements together and classify dissimilar elements into separate clusters. Because reference class labels are often lacking, evaluating the results of clustering algorithms is a challenging task. To overcome this problem, various validity indices are used to objectively measure clustering success. The following section briefly discusses clustering validity indices frequently used in the literature. In this study, the DBI was chosen, as it provides successful results in determining the ideal number of clusters in datasets with an ellipsoidal structure.

2.1 Clustering Validity Indexes

Determining the optimum number of clusters is an important problem in terms of determining to what extent it can serve the purpose of the study after clustering. In order to objectively eliminate this problem, clustering validity indices are used to determine the optimum number of clusters. The basic working principles of clustering techniques are based on developing a weight matrix (μ_{ij}) for a given dataset and creating clusters based on these weights [42]. Most cluster validity indices detect intra-cluster harmony and inter-cluster separation [43].

Partition Coefficient (PC), first proposed by Bezdek for use in the field of fuzzy clustering, is a relatively easy index since only fuzzy membership values are used in the calculation of the clustering validity index. PC index measures the mean square value of the cluster membership matrix and its high value means the best clustering [27]. PC value is calculated as in Eq. (7).

$$V_{PC} = \frac{1}{N}\sum_{\mathrm{j}=1}^{C}\sum_{\mathrm{i}=1}^{N}\mu_{\mathrm{ij}}^{2} \tag{7}$$

Partition Entropy (PE) clustering validity index developed by Bezdek [44] measures the homogeneity of clusters. A lower entropy value means a better clustering result. As the entropy value increases, it can be said that the data contained in the clusters differ from each other. PE value is calculated as in Eq. (8).

$$V_{PE} = \frac{1}{N}\sum_{\mathrm{j}=1}^{C}\sum_{\mathrm{i}=1}^{N}\mu_{\mathrm{ij}}\log_{a}\mu_{ij} \tag{8}$$

The Fukuyama and Sugeno (FS) clustering validity index, given in Eq. (9) developed by Fukuyama and Sugeno for fuzzy clustering algorithms creates more homogeneous and better-defined clusters by considering both distribution and similarity elements [45]. When

$$\overline{v} = \sum_{j=1}^{C}\frac{v_j}{c}$$

then

$$V_{FS} = \sum_{\mathrm{j}=1}^{C}\sum_{\mathrm{i}=1}^{N}\mu_{\mathrm{ij}}^{m}||x_i - v_j||^2 - \sum_{\mathrm{j}=1}^{C}\sum_{\mathrm{i}=1}^{N}\mu_{\mathrm{ij}}^{m}||v_j - \overline{v}||^2 \tag{9}$$

The Xie-Beni (XB) clustering validity index given in Eq. (10), proposed by Xie and Beni [46] and developed by Pal and Bezdek [47], measures the clustering quality by considering both the internal homogeneity of clusters (compactness) and the separation between clusters. A lower XB value indicates more successful clustering [48].

$$V_{XB} = \frac{1}{N}\left(\frac{\sum_{\mathrm{j}=1}^{C}\sum_{\mathrm{i}=1}^{N}\mu_{\mathrm{ij}}^{m}||x_i - v_j||^2}{\min_{\mathrm{j}\neq\mathrm{k}}||v_k - v_j||^2}\right), k = 1, \ldots, C \tag{10}$$

The Silhoulette index, which was developed to determine the proximity of each data point to its own cluster and its distance to other clusters, in other words, its suitability, is calculated with Eq. (11).

$$V_{Sil_{x_i}} = \frac{(b(x_i) - a(x_i))}{\max(a(x_i), b(x_i))} \tag{11}$$

Here, $a(x_i)$ represents the average distance of the ith data point to other points in its own cluster, that is, the within-cluster harmony; $b(x_i)$ represents the average distance of the ith data point to the points in the nearest neighboring cluster to which it does not belong, that is, the out-of-cluster separation [49, 50].

Dunn Index (DI) is a measure of clustering validity that takes into account the degree of compactness of clusters and the degree of separation between clusters. It is a measure that maximizes the separation between clusters and minimizes the density within clusters. The implementation of DI is sensitive to noisy/outlier data and computationally expensive [51]. Compactness (Δ_j) for each cluster is calculated by using each pairwise distance $(d(x_i, x_l))$ with Eq. (12), and the separation value between each pair of clusters (C_k, C_j) is calculated by calculating the distance $d(v_k, v_j)$ between the centers of both clusters with Eq. (13). Using these compactness and separation values, the DI value is calculated as in Eq. (14) [52].

$$\Delta_j = \max_{x_i, x_l \epsilon C_j} d(x_i, x_l), i, l = 1, \ldots N_j \tag{12}$$

$$\delta_{kj} = d(v_k, v_j) \tag{13}$$

$$DI = \frac{\min_{i \le i \ne j < m} \delta(C_i, C_j)}{\max_{1 \le j \le m} \Delta_j} \tag{14}$$

2.1.1 Davies Bouldin Index

The DBI, one of the criteria used to evaluate how accurately the clusters resulting from cluster analysis are clustered, was developed by Davies and Bouldin in 1979 [53]. DBI aims to increase the distance between clusters while minimizing the intra-cluster variance [15].

Step 1 Centers (v_j) are determined for each cluster.

Step 2 For each cluster, the mean of intra-cluster distances (S_j) is calculated.

$$S_j = \frac{1}{N_j} \sum_{i=1}^{N_j} d(x_i, \overline{x_j}) \tag{15}$$

Here, S_j represents the average distance between the data points in the jth cluster, N_j is the number of observations in cluster j.

Step 3 The distances between the centers of two clusters are found as $d_{jk} = d(\overline{x}_j, \overline{x}_k)$.

Step 4 The DB similarity between two clusters is found.

$$R_{jk} = \frac{S_j + S_k}{d_{jk}} \tag{16}$$

Step 5 For each cluster, $R_j = \max_{j \ne k} R_{jk}$ is selected.

Step 6 The DBI is calculated.

$$DBI = \frac{1}{C}\sum_{i=1}^{C} R_j \tag{17}$$

Lower DBI value indicates better clustering performance. In this study, the distance measurements in Step 2 and Step 3 were calculated as the Mahalanobis distance using Eqs. (3–5), since the data sets to be modeled have ellipsoidal scattering.

3 An Algorithm for Parameter Estimation Based on Gustafson Kessel Algorithm

Regression analysis aims to estimate the relationship between the dependent variable (Y) and the independent variables and the regression coefficients (β). The most used method to estimate the parameters of the linear regression model is the least squares method [54].

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{18}$$

One of the critical assumptions for the application of this method is that the error terms follow a normal distribution with zero mean and constant variance. This is expressed as $\epsilon \sim N(0, \sigma^2)$.

$\widehat{Y}$ represents the estimators of the dependent variable Y and is determined by Eq. (19).

$$\widehat{Y} = X\widehat{B} \tag{19}$$

$\widehat{\beta}$ denotes the estimators of the regression coefficients represented by β, and these coefficients are calculated by the Eq. (20):

$$\widehat{B} = (X'X)^{-1}X'Y \tag{20}$$

Although it is assumed that the observations in the data set belong to only one class in classical regression analysis, classical methods may be insufficient in cases where the data has different distributions. In such cases, fuzzy analysis methods offer alternative solutions. In fuzzy regression analysis, clusters in the data set are defined and the membership degrees affecting the estimation are calculated for the existing data sets with fuzzy clustering-based parameter estimation.

The DBI, which shows strong performance, was used to determine the optimal number of clusters for the data sets. In this study, because the data sets produced for parameter estimation and taken from the literature consist of ellipse-shaped data, the DBI used to determine the number of clusters was adapted to use the Mahalanobis distance.

To facilitate rapid convergence to the optimal centroid throughout the algorithm's execution, the initial centroid values for the clusters, whose number is determined based on the DBI, were assigned as the midpoints of the value ranges spanned by the data points constituting each cluster. Using the obtained initial center values, the initial membership degrees were determined based on the distances of the observations to the cluster centers.

Using the GK clustering algorithm, the steps of parameter estimation can be explained as:

Step 1 The optimal number of clusters for the dataset is determined using the adapted DBI.

Step 2 The membership degree matrix μ_{ij} and the stopping criterion ε are determined.

Step 3 Cluster centers based on the fuzziness index m are calculated with Eq. (2).

Step 4 The variance covariance matrix is obtained with Eq. (3).

 https://doi.org/10.1007/s44196-025-01131-9

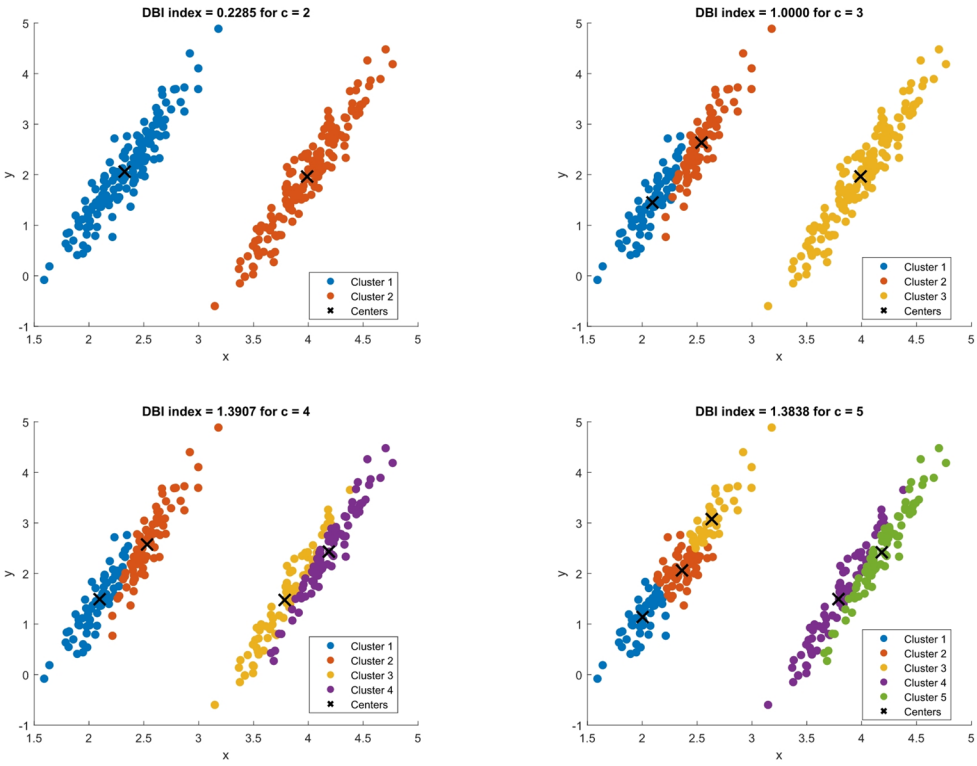


Fig. 1 DBI values for Dataset 1 corresponding to different numbers of clusters (2, 3, 4 and 5)

Table 1 DBI values based on Mahalanobis and Euclidean distances for different cluster numbers for Dataset 1

Methods	2 cluster	3 cluster	4 cluster	5 cluster
DBI (Mahalanobis)	0.2285	1.0000	1.3907	1.3838
DBI (Euclidean)	0.9650	0.8899	1.2088	1.0114

Step 5 The distances between each data point and the centers are calculated with Eq. (4).

Step 6 Membership degrees are obtained with Eq. (6).

Step 7 The centers of the clusters are updated with the new membership degrees.

Step 8 The process is stopped when the difference between the membership degrees in step t (μ_t) and the membership degrees in step t-1 (μ_{t-1}) for the existing clusters is smaller than the specified stopping criterion ε. Thus, the best membership degrees and centers are obtained.

Step 9 Based on the membership degrees obtained, data are divided into clusters.

Step 10 Estimation of the parameters of the linear regression model for the clusters is carried out by using the membership degrees obtained from GK fuzzy clustering as weights. Weighted independent variables $\left(X_{W_j(GK)}\right)$ are obtained by Eq. (21).

$$X_{W_j(GK)} = \mu_{ij(GK)} X_i, i = 1, ..., N, j = 1, \ldots C. \tag{21}$$

Using $X_{W_j(GK)}$, model parameters $\left(\widehat{\beta}_{j(GK)}\right)$ are obtained with Eq. (22).

$$\widehat{\beta}_{j(GK)} = \left(\left(X_{W_j(GK)}\right)' X_i\right)^{-1} \left(\left(X_{W_j(GK)}\right)' Y_i\right), i = 1, ..., N_j, j = 1, \ldots C. \tag{22}$$

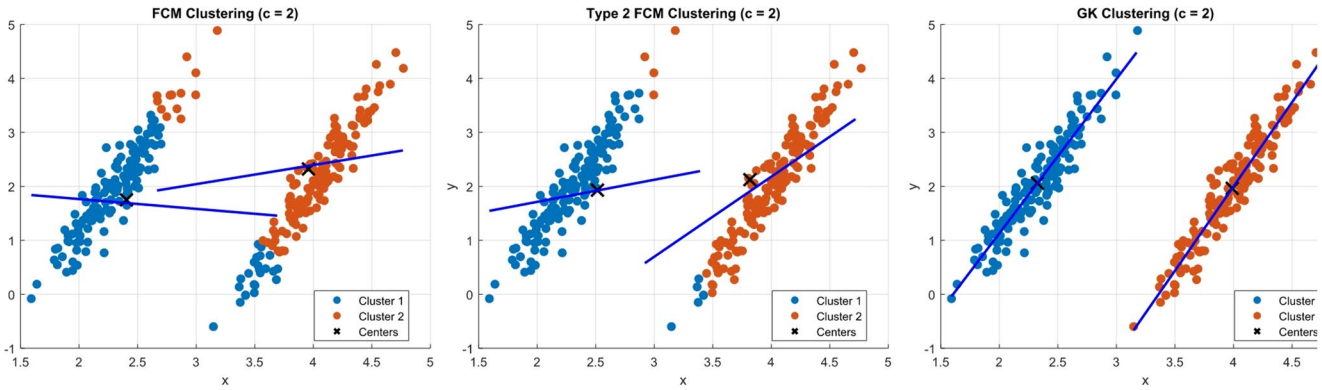


Fig. 2 Model estimates for clusters obtained with FCM, Type-2 FCM and GK for Dataset 1

Table 2 Linear regression models, ε, R^2 and p values for Dataset 1

Method	Cluster	n	Model	ε	R^2	p
FCM	Cluster 1	160	$\widehat{Y_1} = 2.1288 - 0.1820X_{1j}$	0.5100	0.010	0.214
	Cluster 2	140	$\widehat{Y_2} = 0.9815 + 0.3533X_{2j}$	0.5369	0.027	0.054
				$\varepsilon_{total} = 1.0469$		
Type-2 FCM ($m_1 = 5$ and $m_2 = 15$)	Cluster 1	152	$\widehat{Y_1} = -0.5485 + 1.0529X_{1j}$	0.3509	0.155	0.000
	Cluster 2	148	$\widehat{Y_2} = -3.7778 + 1.4889X_{2j}$	0.3375	0.371	0.000
				$\varepsilon_{total} = 0.6884$		
GK	Cluster 1	150	$\widehat{Y_1} = -4.5910 + 2.8603X_{1j}$	0.0965	0.874	0.000
	Cluster 2	150	$\widehat{Y_2} = -10.4880 + 3.1217X_{2j}$	0.0888	0.915	0.000
				$\varepsilon_{total} = 0.1852$		

The estimated value $\left(\widehat{Y}_{i(GK)}\right)$ are calculated by Eq. (23).

$$\widehat{Y}_{i(GK)} = X_{W_{j(GK)}}\widehat{\beta}_{j(GK)}, i = 1, ..., N. \tag{23}$$

Step 11 Error related to the models obtained using GK fuzzy clustering in parameter estimation is calculated with Eqs. (24) and (25).

$$\varepsilon_j = \sum_{i=1}^{N_j} \frac{\mu_{ij(GK)} \left(Y_i - \widehat{Y}_i\right)^2}{N_j} \tag{24}$$

$$\varepsilon = \sum_{j=1}^{C} \varepsilon_j \tag{25}$$

Step 12 According to the error criterion given by Eq. (25), the centers, membership degrees corresponding to these parameters, which give the smallest error, are determined.

 https://doi.org/10.1007/s44196-025-01131-9

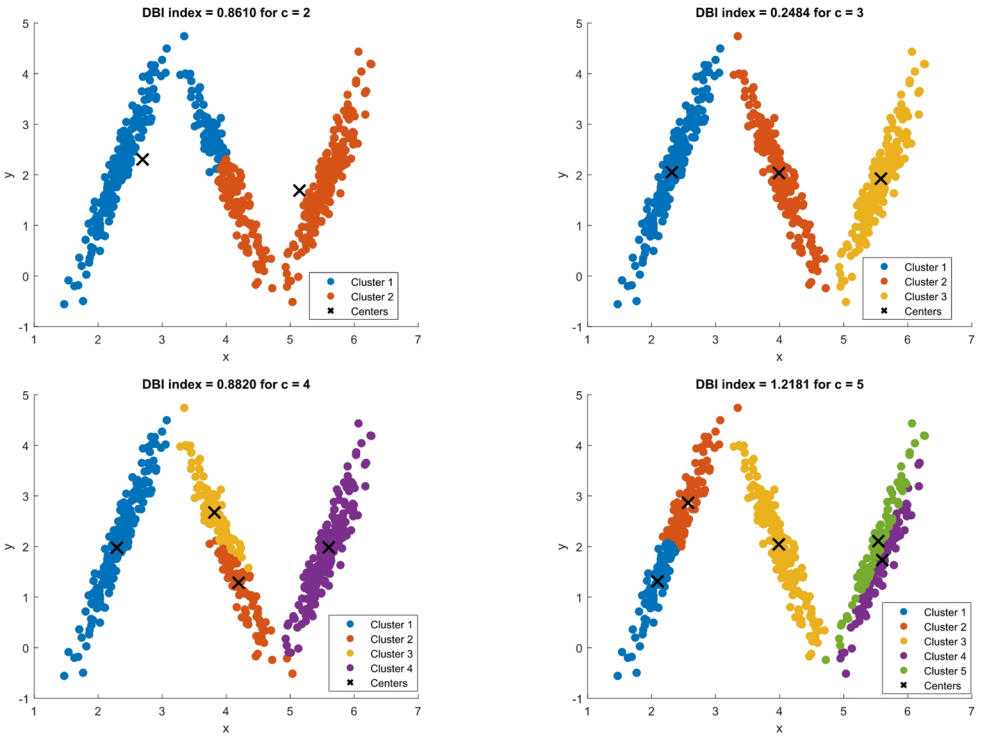


Fig. 3 DBI values for Dataset 2 corresponding to different numbers of clusters (2, 3, 4 and 5)

Table 3 DBI values based on Mahalanobis and Euclidean distances for different cluster numbers for Dataset 2

Methods	2 Cluster	3 Cluster	4 Cluster	5 Cluster
DBI (Mahalanobis)	0.8610	0.2484	0.8820	1.2181
DBI (Euclidean)	0.8077	0.9752	0.7902	1.9948

4 Applications

In this section, the weighted regression models constructed using the membership degrees of observations obtained through the GK algorithm are compared with those generated using the weights determined by the FCM and Type-2 FCM algorithms. In the Type-2 FCM method, coded using MATLAB R2023b, all possible combinations of values for the $[m_1, m_2]$ fuzzification parameter between 2 and 15 were tested, and the model was estimated using the $[m_1, m_2]$ fuzzification parameter that gave the best estimate. The significance of the models obtained from the methods was tested using the p-value determined by Analysis of Variance (ANOVA). To evaluate the effectiveness of the proposed method, four different datasets were analyzed, and the models obtained from each approach were compared in terms of the total error (ε) and the coefficient of determination (R^2). The data set used in the first three examples was artificially generated with the code written in the MATLAB R2023b, and in the fourth example, a ready-made data set available in the literature was used, and details about the data sets were included in each example.

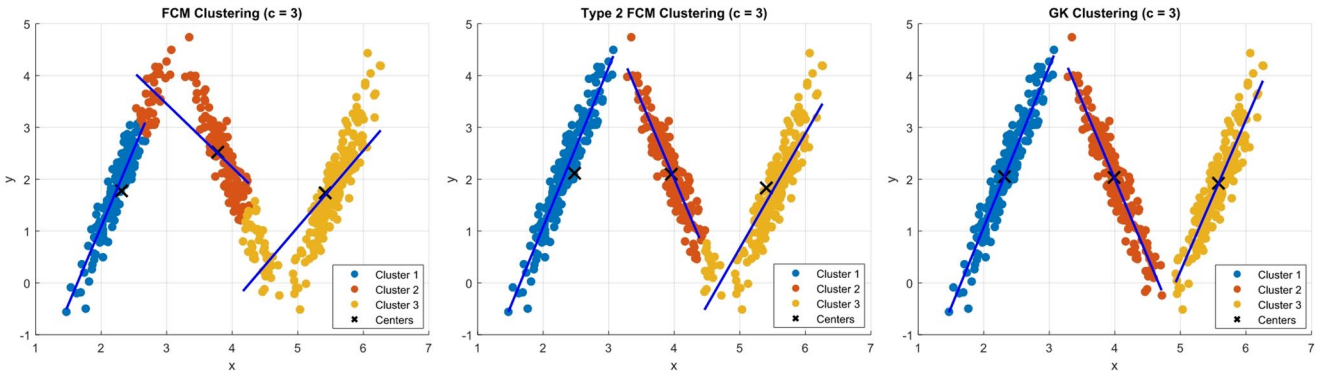


Fig. 4 Model estimates for clusters obtained with FCM, Type-2 FCM and GK for Dataset 2

Table 4 Linear regression models, ε, R^2 and p values for Dataset 2

Method	Cluster	n	Model	ε	R^2	p
FCM	Cluster 1	158	$\widehat{Y_1} = -4.9109 + 2.9976X_{1j}$	0.0640	0.847	0.000
	Cluster 2	198	$\widehat{Y_2} = 7.1182 - 1.2221X_{2j}$	0.1475	0.581	0.000
	Cluster 3	244	$\widehat{Y_3} = -6.3051 + 1.4763X_{3j}$	0.2105	0.601	0.000
				$\varepsilon_{total} = 0.4220$		
Type-2 FCM ($m_1 = 15$ and $m_2 = 15$)	Cluster 1	200	$\widehat{Y_1} = -5.0921 + 3.0843X_{1j}$	0.0846	0.912	0.000
	Cluster 2	182	$\widehat{Y_2} = 13.9127 - 2.9766X_{2j}$	0.0945	0.843	0.000
	Cluster 3	218	$\widehat{Y_3} = -10.3769 + 2.2096X_{3j}$	0.1919	0.782	0.000
				$\varepsilon_{total} = 0.3710$		
GK	Cluster 1	200	$\widehat{Y_1} = -5.1234 + 3.0984X_{1j}$	0.0819	0.915	0.000
	Cluster 2	199	$\widehat{Y_2} = 14.0092 - 3.0027X_{2j}$	0.0894	0.884	0.000
	Cluster 3	201	$\widehat{Y_3} = -14.3914 + 2.9227X_{3j}$	0.1184	0.833	0.000
				$\varepsilon_{total} = 0.2896$		

Example 1 The data set in the first example was produced as two-dimensional, with 300 observations, Gaussian distributed, and consisting of two clusters. This data set consists of two ellipsoid-scattered clusters, one with 150 and the other with 150 observations, where, mean values and variances are determined as, $\mu_1 = (2.3, 2)^T$, $\mu_2 = (4, 2)^T$, $\Sigma_1 = \begin{pmatrix} 0.1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 0.1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$. The calculated DBI values for the cases where the number of clusters for the generated data set was 2, 3, 4, and 5, and their representations according to the number of clusters are given in Fig. 1.

Using both Euclidean and Mahalanobis distance metrics, DBI values were obtained for cluster numbers between 2 and 5 and are presented in Table 1. Table 1 shows that the optimal number of clusters is 2 based on the DBI value obtained using the Mahalanobis distance, and 3 based on the DBI value using the Euclidean distance. However, from the scatter plot in Fig. 1, it can be seen that the number of clusters should be 2. This result was obtained using DBI based on Mahalanobis distance. This is because the data exhibit an ellipsoidal scatter.

With the knowledge that the optimum number of clusters for the data in the first example is 2, Fig. 2 shows the clustering results using FCM, Type-2 FCM and GK.

 https://doi.org/10.1007/s44196-025-01131-9

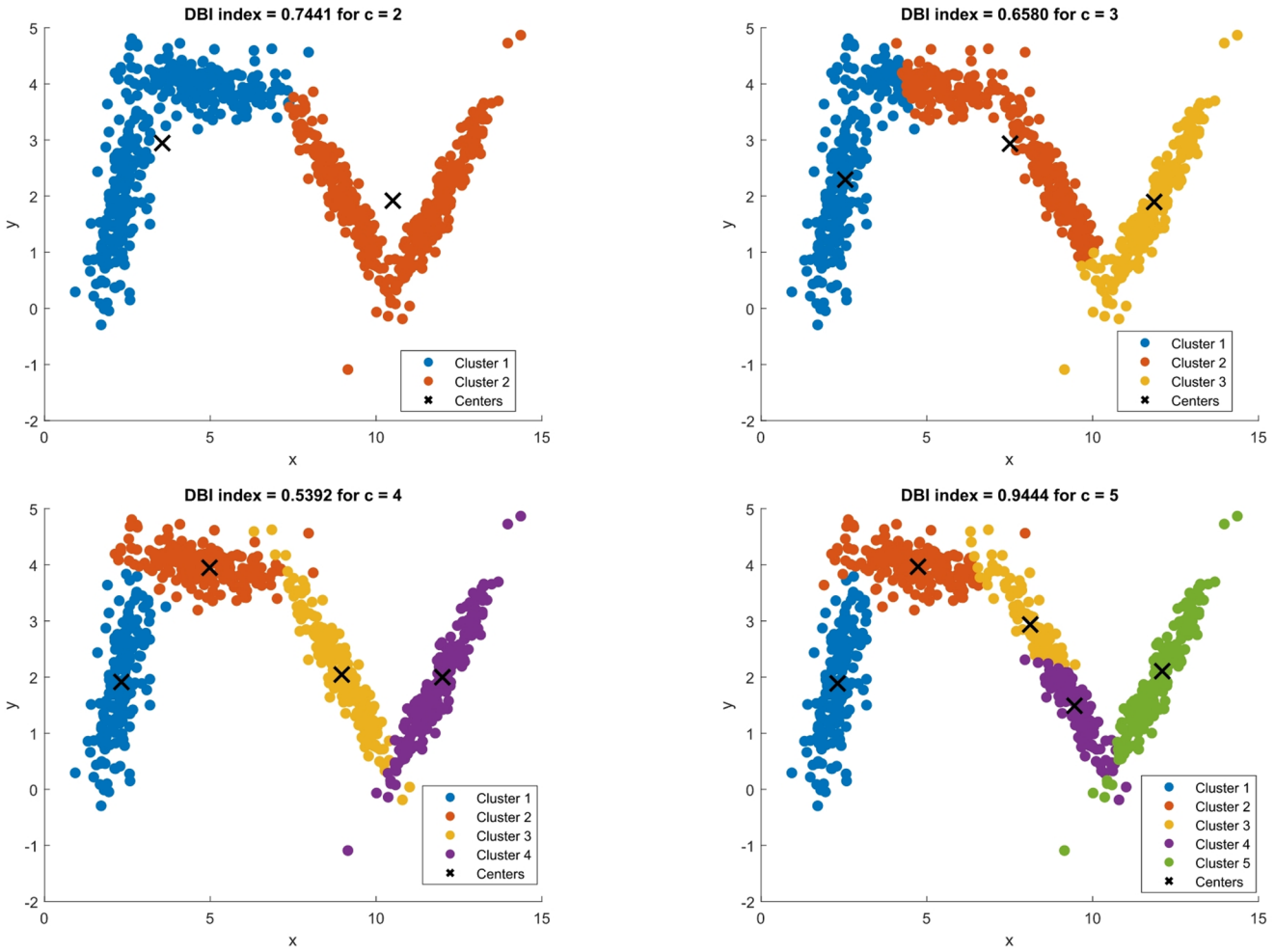


Fig. 5 DBI values for Dataset 3 corresponding to different numbers of clusters (2, 3, 4 and 5)

Table 5 DBI values based on Mahalanobis and Euclidean distances for different cluster numbers for Dataset 3

Methods	2 Cluster	3 Cluster	4 Cluster	5 Cluster
DBI (Mahalanobis)	0.7441	0.6580	0.5392	0.9444
DBI (Euclidean)	0.5137	0.6541	0.6061	0.6408

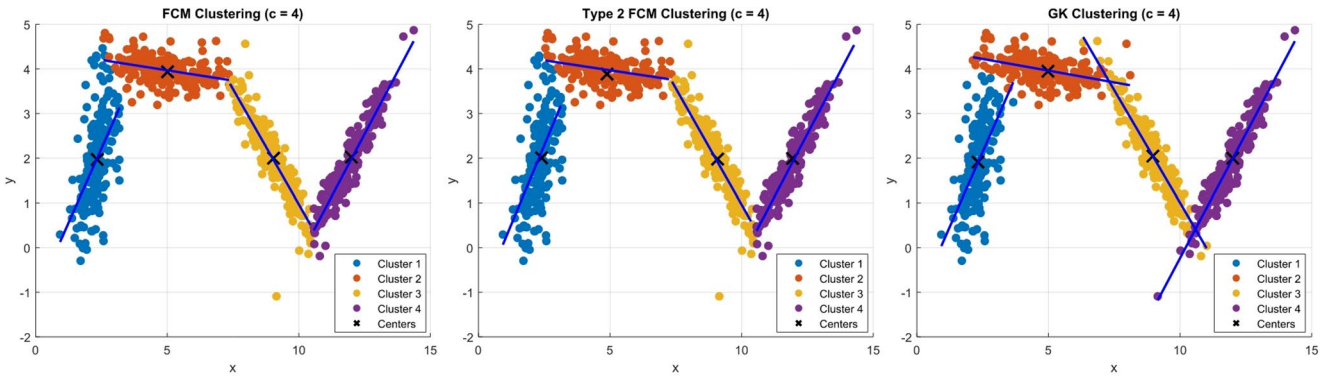


Fig. 6 Model estimates for clusters obtained with FCM, Type-2 FCM and GK for Dataset 3

Table 6 Linear regression models, ε, R^2 and p values for Dataset 3

Method	Cluster	n	Model	ε	R^2	p
FCM	Cluster 1	204	$\widehat{Y_1} = -1.1064 + 1.3297X_{1j}$	0.4948	0.322	0.000
	Cluster 2	190	$\widehat{Y_2} = 4.4326 - 0.0927X_{2j}$	0.0600	0.103	0.000
	Cluster 3	205	$\widehat{Y_3} = 11.2804 - 1.0313X_{3j}$	0.0968	0.801	0.000
	Cluster 4	201	$\widehat{Y_4} = -11.4218 + 1.1172X_{4j}$	0.0672	0.892	0.000
				$\varepsilon_{total} = 0.7188$		
Type-2 FCM ($m_1 = 5$ and $m_2 = 5$)	Cluster 1	201	$\widehat{Y_1} = -1.0261 + 1.2970X_{1j}$	0.6173	0.317	0.000
	Cluster 2	192	$\widehat{Y_2} = 4.4451 - 0.0953X_{2j}$	0.0763	0.115	0.000
	Cluster 3	206	$\widehat{Y_3} = 11.3750 - 1.0417X_{3j}$	0.1405	0.805	0.000
	Cluster 4	201	$\widehat{Y_4} = -11.3931 + 1.1148X_{4j}$	0.0806	0.884	0.000
				$\varepsilon_{total} = 0.9147$		
GK	Cluster 1	195	$\widehat{Y_1} = -1.2094 + 1.3417X_{1j}$	0.4281	0.395	0.000
	Cluster 2	195	$\widehat{Y_2} = 4.4877 - 0.1053X_{2j}$	0.0520	0.179	0.000
	Cluster 3	205	$\widehat{Y_3} = 11.0849 - 1.0083X_{3j}$	0.0665	0.883	0.000
	Cluster 4	205	$\widehat{Y_4} = -11.3717 + 1.1138X_{4j}$	0.0657	0.912	0.000
				$\varepsilon_{total} = 0.6123$		

The estimated models for the clusters obtained from all methods are given in Table 2, together with the observation numbers (n), error amounts (ε) calculated using Eq. (25), coefficients of determination (R^2) and p values for Dataset 1.

It is observed from the p value at the significance level of $\alpha = 0.05$ that the prediction models for all clusters obtained from all methods, except the model for the clusters obtained from FCM, are significant. When the coefficients of determination (R^2) for all models are examined, it is seen that the highest coefficient of determination is achieved by the models obtained from the proposed method based on GK. The total error amount for the estimates obtained from the weighted models, which were created using the degrees of observations belonging to the clusters determined by the GK algorithm, was calculated as $\varepsilon_{GK} = 0.1852$. The total error amount of the models created by using the weights determined by the FCM and Type-2 FCM are.

$\varepsilon_{FCM} = 1.0469$ and $\varepsilon_{T2FCM} = 0.6884$, respectively. As a result, the models based on the weights obtained with the GK algorithm for this data set provided more effective results in terms of mean square errors.

Example 2 In this example, the dataset was produced as two-dimensional, with 600 observations, Gaussian distributed, and consisting of three clusters. This dataset consists of 3 ellipsoid-scattered clusters, each containing 200 observations where, mean values and variances are as follows: $\mu_1 = (0.1, 0.3)^T$, $\mu_2 = (4, 2)^T$, $\mu_3 = (5.6, 2)^T$, $\Sigma_1 = \begin{pmatrix} 0.1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0.1 & -0.3 \\ -0.3 & 1 \end{pmatrix}$ and $\Sigma_3 = \begin{pmatrix} 0.1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$. The calculated DBI values for the cases where the number of clusters for the generated data set was 2, 3, 4, and 5, and their representations according to the number of clusters are given in Fig. 3, and the index values for different cluster numbers are given in Table 3.

As seen in Table 3, the number of clusters yielding the smallest DBI value was determined to be 3 when using the Mahalanobis distance. In the same table, DBI using the Euclidean distance indicates that the optimal number of clusters should be 4. However, the scatter plot in Fig. 3 indicates that the number of clusters should be 3. This

 https://doi.org/10.1007/s44196-025-01131-9

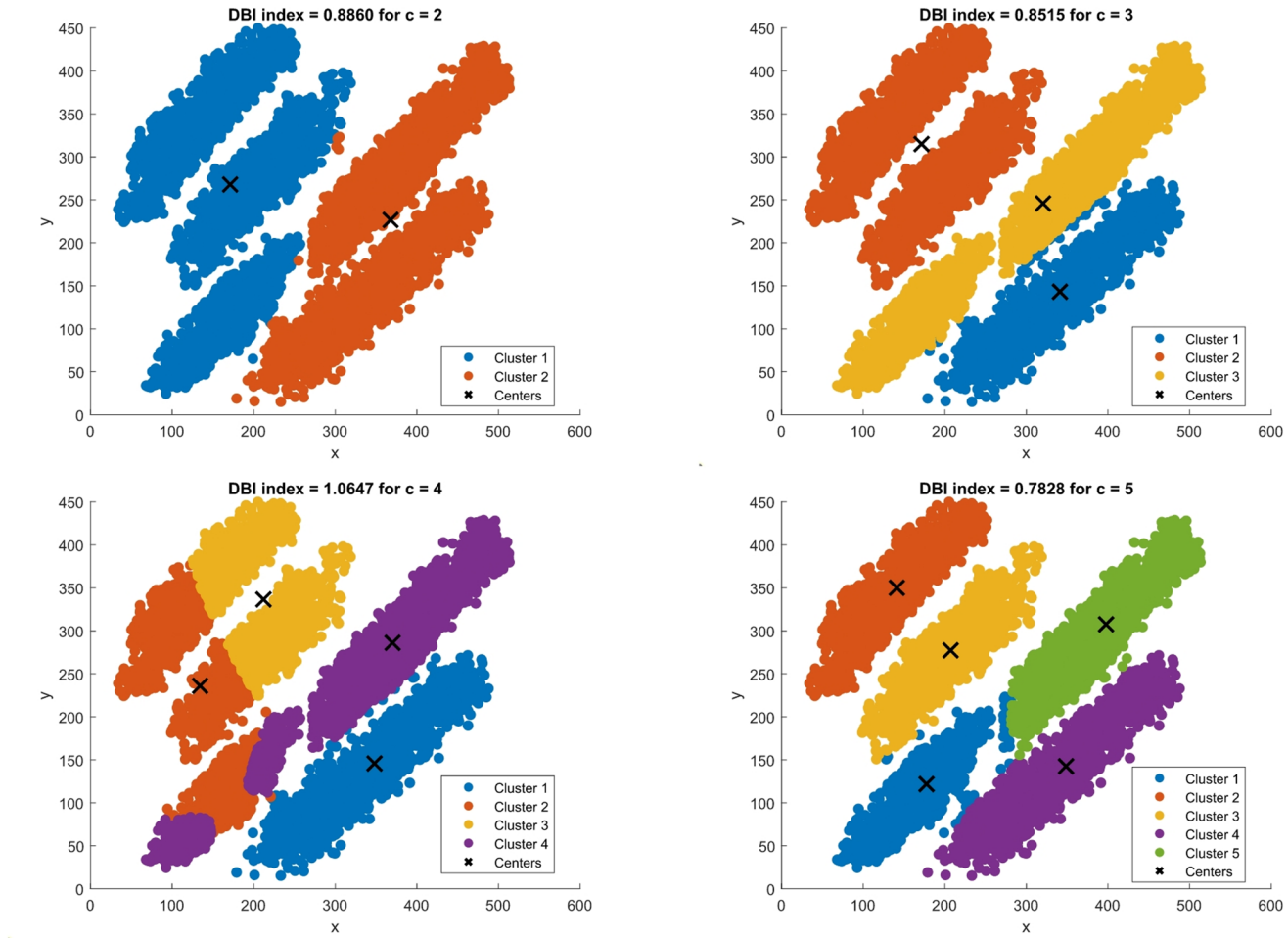


Fig. 7 DBI values for Dataset 4 corresponding to different numbers of clusters (2, 3, 4 and 5)

Table 7 DBI values based on Mahalanobis and Euclidean distances for different cluster numbers for Dataset 4

Methods	2 cluster	3 cluster	4 cluster	5 cluster
DBI (Mahalanobis)	0.8860	0.8515	1.0647	0.7828
DBI (Euclidean)	1.0542	1.8181	1.2502	0.9734

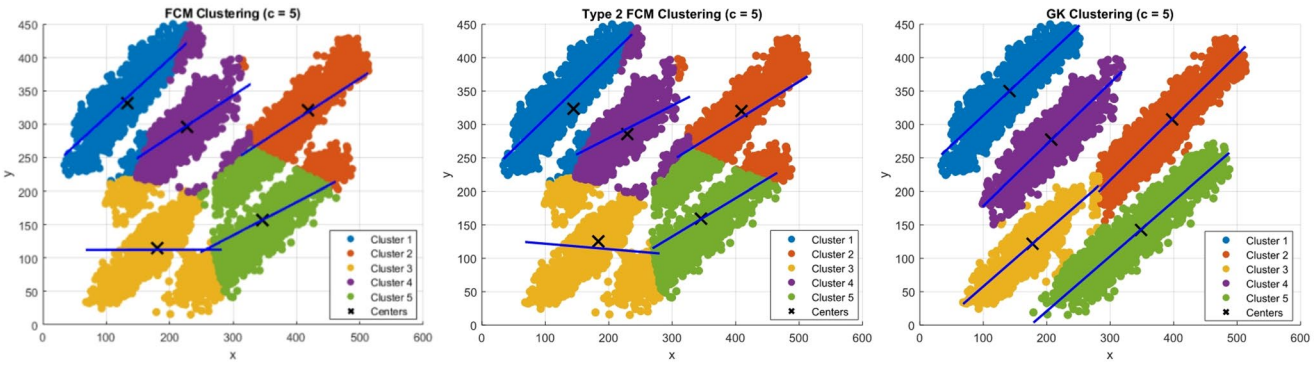


Fig. 8 Model estimates for clusters obtained with FCM, Type-2 FCM and GK for Dataset 4

Table 8 Linear regression models, ε, R^2 and p values for Dataset 4

Method	Cluster	n	Model	ε	R^2	p
FCM	Cluster 1	1117	$\widehat{Y_1} = 223.9491 + 0.8692X_{1j}$	623.6271	0.566	0.000
	Cluster 2	1300	$\widehat{Y_2} = 60.5745 + 0.6156X_{2j}$	1241.7254	0.274	0.000
	Cluster 3	1317	$\widehat{Y_3} = 111.5176 + 0.0043X_{3j}$	1327.1064	0.000	0.872
	Cluster 4	1062	$\widehat{Y_4} = 156.7268 + 0.6200X_{4j}$	874.6976	0.247	0.000
	Cluster 5	1399	$\widehat{Y_5} = -14.2994 + 0.4942X_{5j}$	1075.7332	0.193	0.000
				$\varepsilon_{total} = 5142.8897$		
Type-2 FCM ($m_1 = 14$ and $m_2 = 15$)	Cluster 1	1156	$\widehat{Y_1} = 217.5511 + 0.8979X_{1j}$	250.4230	0.598	0.000
	Cluster 2	1296	$\widehat{Y_2} = 77.0433 + 0.5695X_{2j}$	458.9357	0.249	0.000
	Cluster 3	1311	$\widehat{Y_3} = 130.8847 - 0.0811X_{3j}$	536.1951	0.006	0.004
	Cluster 4	1037	$\widehat{Y_4} = 172.0881 + 0.5227X_{4j}$	365.1980	0.207	0.000
	Cluster 5	1395	$\widehat{Y_5} = -34.2501 + 0.5602X_{5j}$	429.0923	0.240	0.000
				$\varepsilon_{total} = 2039.8440$		
GK	Cluster 1	1129	$\widehat{Y_1} = 224.4450 + 0.8901X_{1j}$	474.4853	0.771	0.000
	Cluster 2	1448	$\widehat{Y_2} = -65.1914 + 0.9390X_{2j}$	365.2005	0.846	0.000
	Cluster 3	922	$\widehat{Y_3} = -24.8573 + 0.8275X_{3j}$	342.4133	0.724	0.000
	Cluster 4	1166	$\widehat{Y_4} = 86.4920 + 0.09187X_{4j}$	444.9551	0.758	0.000
	Cluster 5	1530	$\widehat{Y_5} = -143.6122 + 0.8227X_{5j}$	337.2259	0.869	0.000
				$\varepsilon_{total} = 1964.2801$		

result was achieved with the DBI using the Mahalanobis distance. This is because the data in this example also exhibits an ellipsoidal scatter, as in Example 1.

With the knowledge that the optimum number of clusters for the data in the second example is 3, Fig. 4 shows the clustering results using FCM, Type-2 FCM and GK.

Three estimation models for each cluster in all methods are given in Table 4 together with the observation numbers (n), error amounts calculated using Eq. (25), coefficients of determination (R^2) and p values for Dataset 2.

It is observed from the p value at the significance level of α=0.05 that the prediction models for all clusters obtained from all methods are significant. When the coefficients of determination (R^2) for all models are examined, it is seen that the highest coefficient of determination is achieved by the models obtained from the proposed method based on GK.

When the total error amounts were obtained from the weighted models created using the observation degrees of the clusters determined by the GK algorithm, $\varepsilon_{GK} = 0.2896$ was calculated, and when obtained from the models created using the weights determined by the FCM and Type-2 FCM, $\varepsilon_{FCM} = 0.4220$, $\varepsilon_{T2FCM} = 0.3710$ was calculated, respectively. This result also shows that the models based on the weights obtained with the GK algorithm for this data set gave more effective results in terms of mean square error.

Example 3 In this example, we consider a two-dimensional data set of 800 observations, Gaussian distributed, and consisting of four clusters. This dataset consists of 4 ellipsoid-scattered clusters, each containing 200 observations where, mean values and variances are as follows: $\mu_1 = (2.3, 2)^T$, $\mu_2 = (5, 4)^T$, $\mu_3 = (9, 2)^T$, $\mu_4 = (12, 2)^T$, $\Sigma_1 = \begin{pmatrix} 0.2 & 0.3 \\ 0.3 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1.5 & -0.08 \\ -0.08 & 0.08 \end{pmatrix}$, $\Sigma_3 = \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 0.6 \end{pmatrix}$ and $\Sigma_4 = \begin{pmatrix} 0.7 & 0.8 \\ 0.8 & 1 \end{pmatrix}$. The calculated DBI values for the cases where the number of clusters for the generated data set was 2, 3, 4, and 5, and their

representations according to the number of clusters are given in Fig. 5, and the index values for different cluster numbers are given in Table 5.

Table 5 shows that the optimal number of clusters should be 4 with the DBI value using the Mahalanobis distance, and 2 with the DBI value using the Euclidean distance. However, the scatter plot in Fig. 5 indicates that the optimal number of clusters should be 4. This result was achieved with the DBI using the Mahalanobis distance. This is because the data, as in the previous examples, exhibits an ellipsoidal scatter.

With the knowledge that the optimum number of clusters for the data in the third example is 4, Fig. 6 shows the clustering results using FCM, Type-2 FCM and GK.

Models for all methods are given in Table 6 together with the observation numbers (n), error amounts calculated using Eq. (25), coefficients of determination (R^2) and p values for Dataset 3.

It is observed from the p value at the significance level of $\alpha = 0.05$ that the prediction models for all clusters obtained from all methods are significant. When the coefficients of determination (R^2) for all models are examined, it is seen that the highest coefficient of determination is achieved by the models obtained from the proposed method based on GK.

The total error amounts were obtained from both the weighted models created using the observation degrees of the clusters determined by the GK algorithm and the models created using the weights determined by the FCM and Type-2 FCM, and were calculated as $\varepsilon_{GK} = 0.6123$, $\varepsilon_{FCM} = 0.7187$, $\varepsilon_{T2FCM} = 0.9147$ respectively. From these values, it is seen that the models based on the weights obtained with the GK algorithm give more effective results in terms of mean square error.

Example 4 In the fourth example, a lines data set with predetermined clusters that can be accessed from the literature (https://www.kaggle.com/datasets/joonasyoon/clustering-exercises?select=lines2.csv) is considered. This 2-dimensional dataset is composed of five clusters, comprising a total of 6195 observations. The number of observations in each cluster is 1129, 1466, 882, 1169, and 1549, respectively. The calculated DBI values for the cases where the number of clusters for the data set was 2, 3, 4, and 5, and their representations according to the number of clusters are given in Fig. 7, and the index values for different cluster numbers are given in Table 7.

Table 7 shows that the smallest DBI value is achieved when the appropriate number of clusters for the data is determined as 5 according to both distance measures.

With the knowledge that the optimum number of clusters for the data in the fourth example is 5, Fig. 8 shows the clustering results using FCM, Type-2 FCM and GK.

The five models estimated for the clusters obtained from all methods are given in Table 8 together with the observation numbers (n) , error amounts calculated using Eq. (25), coefficients of determination (R^2) and p values for Dataset 4.

It is observed from the p value at the significance level of $\alpha = 0.05$ that the prediction models for all clusters obtained from all methods, except the model for the 3rd cluster obtained from FCM, are significant. When the coefficients of determination (R^2) for all models are examined, it is seen that the highest coefficient of determination is achieved by the models obtained from the proposed method based on GK.

Total error amounts were obtained from both the weighted models created using the observation degrees of the clusters determined by the GK and the models created using the weights determined by the FCM and Type-2 FCM. These values obtained as $\varepsilon_{GK} = 1964.2801$, $\varepsilon_{FCM} = 5142.8897$ and $\varepsilon_{T2FCM} = 2039.8440$ show that the models based on the weights obtained by the GK algorithm give more effective results in terms of mean square error.

Table 9 Models obtained from the methods, coefficients of determination, and total error amounts

Examples	Models	R^2	Total error amounts
Example 1 (2 cluster)	$\widehat{Y}_{1(FCM)} = 2.1288 - 0.1820X_{1j}$	0.010	$\varepsilon_{(FCM)} = 1.0469$
	$\widehat{Y}_{2(FCM)} = 0.9815 + 0.3533X_{2j}$	0.027	
	$\widehat{Y}_{1(T2FCM)} = -0.5485 + 1.0529X_{1j}$	0.155	$\varepsilon_{(T2FCM)} = 0.6884$
	$\widehat{Y}_{2(T2FCM)} = -3.7778 + 1.4889X_{2j}$	0.371	
	$\widehat{Y}_{1(GK)} = -4.5910 + 2.8603X_{1j}$	0.874	$\varepsilon_{(GK)} = 0.1852$
	$\widehat{Y}_{2(GK)} = -10.4880 + 3.1217X_{2j}$	0.915	
Example 2 (3 cluster)	$\widehat{Y}_{1(FCM)} = -4.9109 + 2.9976X_{1j}$	0.847	$\varepsilon_{(FCM)} = 0.4220$
	$\widehat{Y}_{2(FCM)} = 7.1182 - 1.2221X_{2j}$	0.581	
	$\widehat{Y}_{3(FCM)} = -6.3051 + 1.4763X_{3j}$	0.601	
	$\widehat{Y}_{1(T2FCM)} = -5.0921 + 3.0843X_{1j}$	0.912	$\varepsilon_{(T2FCM)} = 0.3710$
	$\widehat{Y}_{2(T2FCM)} = 13.9127 - 2.9766X_{2j}$	0.843	
	$\widehat{Y}_{3(T2FCM)} = -10.3769 + 2.2096X_{3j}$	0.782	
	$\widehat{Y}_{1(GK)} = -5.1234 + 3.0984X_{1j}$	0.915	$\varepsilon_{(GK)} = 0.2896$
	$\widehat{Y}_{2(GK)} = 14.0092 - 3.0027X_{2j}$	0.884	
	$\widehat{Y}_{3(GK)} = -14.3914 + 2.9227X_{3j}$	0.833	
Example 3 (4 cluster)	$\widehat{Y}_{1(FCM)} = -1.1064 + 1.3297X_{1j}$	0.322	$\varepsilon_{(FCM)} = 0.7188$
	$\widehat{Y}_{2(FCM)} = 4.4326 - 0.0927X_{2j}$	0.103	
	$\widehat{Y}_{3(FCM)} = 11.2804 - 1.0313X_{3j}$	0.801	
	$\widehat{Y}_{4(FCM)} = -11.4218 + 1.1172X_{4j}$	0.892	
	$\widehat{Y}_{1(T2FCM)} = -1.0261 + 1.2970X_{1j}$	0.317	$\varepsilon_{(T2FCM)} = 0.9147$
	$\widehat{Y}_{2(T2FCM)} = 4.4451 - 0.0953X_{2j}$	0.115	
	$\widehat{Y}_{3(T2FCM)} = 11.3750 - 1.0417X_{3j}$	0.805	
	$\widehat{Y}_{4(T2FCM)} = -11.3931 + 1.1148X_{4j}$	0.884	
	$\widehat{Y}_{1(GK)} = -1.2094 + 1.3417X_{1j}$	0.395	$\varepsilon_{(GK)} = 0.6123$
	$\widehat{Y}_{2(GK)} = 4.4877 - 0.1053X_{2j}$	0.179	
	$\widehat{Y}_{3(GK)} = 11.0849 - 1.0083X_{3j}$	0.883	
	$\widehat{Y}_{4(GK)} = -11.3717 + 1.1138X_{4j}$	0.912	
Example 4 (5 cluster)	$\widehat{Y}_{1(FCM)} = 223.9491 + 0.8692X_{1j}$	0.566	$\varepsilon_{(FCM)} = 5142.8897$
	$\widehat{Y}_{2(FCM)} = 60.5745 + 0.6156X_{2j}$	0.274	
	$\widehat{Y}_{3(FCM)} = 111.5176 + 0.0043X_{3j}$	0.000	
	$\widehat{Y}_{4(FCM)} = 156.7268 + 0.6200X_{4j}$	0.247	
	$\widehat{Y}_{5(FCM)} = -14.2994 + 0.4942X_{5j}$	0.193	
	$\widehat{Y}_{1(T2FCM)} = 217.5511 + 0.8979X_{1j}$	0.598	$\varepsilon_{(T2FCM)} = 2039.8440$
	$\widehat{Y}_{2(T2FCM)} = 77.0433 + 0.5695X_{2j}$	0.249	
	$\widehat{Y}_{3(T2FCM)} = 130.8847 - 0.0811X_{3j}$	0.006	
	$\widehat{Y}_{4(T2FCM)} = 172.0881 + 0.5227X_{4j}$	0.207	
	$\widehat{Y}_{5(T2FCM)} = -34.2501 + 0.5602X_{5j}$	0.240	
	$\widehat{Y}_{1(GK)} = 224.4450 + 0.8901X_{1j}$	0.771	$\varepsilon_{(GK)} = 1964.2801$
	$\widehat{Y}_{2(GK)} = -65.1914 + 0.9390X_{2j}$	0.846	
	$\widehat{Y}_{3(GK)} = 24.8573 + 0.8275X_{3j}$	0.724	
	$\widehat{Y}_{4(GK)} = 86.4920 + 0.09187X_{4j}$	0.758	
	$\widehat{Y}_{5(GK)} = -143.6122 + 0.8227X_{5j}$	0.869	

5 Conclusion and Results

Since the primary objective of this study is parameter estimation, model predictions on ellipsoidal-scattered data sets were obtained using methods based on Mahalanobis distance, which yields more effective results for such data distributions. In the case of ellipsoidal scattering of data, the proposed algorithm was run on various datasets to determine the regression model with higher coefficients of determination and lower error by estimating unknown parameters.

Considering cases where the data structure is ellipsoid, the DBI used to determine the optimal number of clusters has been adapted to use the Mahalanobis distance instead of the Euclidean distance. After the optimum number of clusters was determined, the clustering process was carried out using FCM, Type-2 FCM and GK clustering methods, and the degree of membership of each observation in the clusters was calculated. These membership degrees were determined as the weights to be used in the model and the unknown parameters related to the regression models were predicted by using the weights obtained with FCM, Type-2 FCM and GK on four different examples. The error amounts of the created models and the coefficients of determination of the models were calculated. Detailed results obtained by solving the examples considered are summarized in Table 9.

As seen in Table 9, for each example, the total error amounts of the models constructed with weights obtained from the GK clustering were lower than the total error amounts of the models constructed with weights obtained from the FCM and Type-2 FCM. Furthermore, when the coefficients of determination of the models are examined, it is observed that the validity of the models obtained from the proposed GK-based algorithm is higher than that of the other methods. This clearly shows that the use of weights obtained with clustering based on the proposed GK based algorithm provides more effective estimation models in estimating linear regression models for data sets with ellipsoidal scattering.

In the proposed method, the data contributions to the model during the parameter estimation phase are determined through weighting, unlike classical regression. The weights used are membership degrees obtained from the Mahalanobis distance-based GK clustering method. Because the proposed method relies on weighting in the parameter estimation process, it can be included in the class of existing robust regression methods. In this context, determining the data contributions to the model through membership degrees is anticipated to minimize the negative effects of possible outliers in the data set, enabling efficient parameter estimations even in data sets with outliers. In addition, as in the Type-2 FCM used for comparison in the application part of the study, a new parameter estimation method based on Type-2 fuzzy logic can be developed by defining the fuzzification index m in the GK algorithm proposed in this study as an interval fuzzy number format $[m_1, m_2]$. Another aspect of the problem that could be addressed in the future is the high number of dimensions.

Acknowledgements The authors wish to acknowledge the editor and anonymous reviewers for their insightful comments, which have improved the quality of this publication.

Author Contributions All authors in the article contributed equally to identifying the problem, designing the article and converting it into text, converting the algorithm created to solve the problem into software, and collecting and analyzing the data.

Funding The authors received no financial support for the research, authorship, and/or publication of this article.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

Ethical Approval and Consent to Participate All authors are contributing and accepting to submit the current work.

Consent for Publication All authors agree to submit and publish the submitted work.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Benatti, K.A., Pedroso, L.G., Ribeiro, A.A.: Fuzzy clustering with capacity constraints: algorithm, convergence analysis and numerical experiments. Expert Syst. Appl. (2024). https://doi.org/10.1016/j.eswa.2024.125191
2. Wu, H.-C.: Fuzzy estimates of regression parameters in linear regression models for imprecise input and output data. Comput. Stat. Data Anal. **42**, 203–217 (2003). https://doi.org/10.1016/S0167-9473(02)00116-0
3. Kula, K.S., Dalkiliç, T.E.: Parameter Estimation Based Type-II Fuzzy Logic. TWMS J. Pure Appl. Math. **4**, 187–193 (2013)
4. Tanaka, H., Uejima, S., Asai, K.: Linear regression analysis with fuzzy model. IEEE Trans. Syst. Man Cybern. **12**, 903–907 (1982). https://doi.org/10.1109/TSMC.1982.4308925
5. Ahmadini, A.A.H.: A novel technique for parameter estimation in intuitionistic fuzzy logistic regression model. Ain Shams Eng. J. **13**, 101518 (2022). https://doi.org/10.1016/j.asej.2021.06.004
6. Chou, T.-Y., Han, T.-C., Liang, G.-S., Hsu, C.-L.: Applying fuzzy regression forecasting model to predict Taiwan International Air Cargo volume. In: Proceedings of the Eastern Asia Society for Transportation Studies. p 123. https://doi.org/10.1109/ANNES.1995.499469 (2007)
7. Jung, H.-Y., Lee, W.-J., Yoon, J.H., Choi, S.H.: Likelihood inference based on fuzzy data in regression model. In: 2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS), IEEE. pp. 1175–1179. https://doi.org/10.1109/SCIS-ISIS.2014.7044744 (2014).
8. Dalkiliç, T.E., Sağirkaya, S.: Parameter prediction based on type-2 fuzzy clustering. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **26**, 877–892 (2018). https://doi.org/10.1142/S0218488518500393
9. Enke, D., Grauer, M., Mehdiyev, N.: Stock market prediction with multiple regression, fuzzy type-2 clustering and neural networks. Procedia Comput. Sci. **6**, 201–206 (2011). https://doi.org/10.1016/j.procs.2011.08.038
10. Fazel Zarandi, M.H., Gamasaee, R., Turksen, I.B.: A type-2 fuzzy c-regression clustering algorithm for Takagi–Sugeno system identification and its application in the steel industry. Inf. Sci. **187**, 179–203 (2012). https://doi.org/10.1016/j.ins.2011.10.015
11 Mokhtari, M., Allahviranloo, T., Behzadi, M.H., Lotfi, F.H.: Interval type-2 fuzzy least-squares estimation to formulate a regression model based on a new outlier detection method using a new distance. Comput. Appl. Math. **1**, 45 (2021)

12. Shi, J.: T-S fuzzy model identification based on an improved interval type-2 fuzzy c-regression model. J. Intell. Fuzzy Syst. **44**, 4495–4507 (2023). https://doi.org/10.3233/JIFS-221434
13. Jumadi Dehotman Sitompul, B., Salim Sitompul, O., Sihombing, P.: Enhancement clustering evaluation result of Davies-Bouldin index with determining initial centroid of K-Means algorithm J. . Phys. Conf. Ser. **1235**, 012015 (2019). https://doi.org/10.1088/1742-6596/1235/1/012015
14. Tarigan, D.A.: Optimization of the K-Means clustering algorithm using Davies Bouldin index in Iris data classification. Media Online) **4**, 545–552 (2023). https://doi.org/10.30865/klik.v4i1.964
15. Ros, F., Riad, R., Guillaume, S.: PDBI: A partitioning Davies-Bouldin index for clustering evaluation. Neurocomputing **528**, 178–199 (2023). https://doi.org/10.1016/j.neucom.2023.01.043
16. Rao, R.K., Mandhala, V.N.: Comparative analysis of clustering algorithms for financial fraud detection. Int. J. Saf. Secur. Eng. **15**, 777–785 (2025). https://doi.org/10.18280/ijsse.150414
17. Chao, G., Sun, S., Bi, J.: A survey on multiview clustering. IEEE Trans. Artif. Intell. **2**, 146–168 (2021). https://doi.org/10.1109/TAI.2021.3065894
18. Xu, R., WunschII, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**, 645–678 (2005). https://doi.org/10.1109/TNN.2005.845141
19. Feng, Q., Chen, L., Chen, C.L.P., Guo, L.: Deep fuzzy clustering - a representation learning approach. IEEE Trans. Fuzzy Syst. **28**, 1–1 (2020). https://doi.org/10.1109/TFUZZ.2020.2966173
20. Ruspini, E.H., Bezdek, J.C., Keller, J.M.: Fuzzy clustering: A historical perspective. IEEE Comput. Intell. Mag. **14**, 45–55 (2019). https://doi.org/10.1109/MCI.2018.2881643
21. Zadeh, L.A.: Fuzzy sets. Inf. Control. **8**, 338–353 (1965). https://doi.org/10.1016/S0019-9958(65)90241-X
22 Bellman, R.E., Zadeh, L.A.: Decision-making in a fuzzy environment. Manag. Sci. **17**(4), 141 (1970). https://doi.org/10.1287/Mnsc.17.4.B141
23. Zadeh, L.A., Fu, K.S., Tanaka, K., Shimura, M., Negoita, C.V.: Fuzzy sets and their applications to cognition and decision processes. IEEE Trans. Syst. Man Cybern. **7**, 122–123 (1977). https://doi.org/10.1109/TSMC.1977.4309670
24 Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. **3**(1974), 32–57 (1974)
25. Bezdek, J.C.: FCM: the fuzzy c-means clustering algorithm. Comput. Geosci. **10**, 191–203 (1984)
26 Höppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. Wiley, England (1999)
27. J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Springer US, Boston, MA, 1981. https://doi.org/10.1007/978-1-4757-0450-1.
28. Khan, H., Rasheed, M.T., Zhang, S., Wang, X., Liu, H.: Empirical study of outlier impact in classification context. Expert Syst. Appl. (2024). https://doi.org/10.1016/j.eswa.2024.124953
29. Gustafson, D., Kessel, W.: Fuzzy clustering with a fuzzy covariance matrix. In: 1978 IEEE Conference on Decision and Control Including the 17th Symposium on Adaptive Processes, IEEE. pp. 761–766. https://doi.org/10.1109/CDC.1978.268028 (1978)
30. Chen, Q., Yu, W., Zhao, X., Nie, F., Li, X.: Rooted Mahalanobis distance based Gustafson-Kessel fuzzy C-means. Inf. Sci. (2023). https://doi.org/10.1016/j.ins.2023.03.103
31. Abdullah, A., Banmongkol, C., Hoonchareon, N., Hidaka, K.: A study on the gustafson-kessel clustering algorithm in power system fault identification. J. Electr. Eng. Technol. **12**, 1798–1804 (2017). https://doi.org/10.5370/JEET.2017.12.5.1798
32. Miller, D.J., Nelson, C.A., Cannon, M.B., Cannon, K.P.: Comparison of fuzzy clustering methods and their applications to geophysics data. Appl. Comput. Intell. Soft Comput. **2009**, 1–16 (2009). https://doi.org/10.1155/2009/876361
33. Wang, H., Zheng, W.D., Huang, C.J., Li, S.G., Qian, Y., Jiang, X.C.: Application of Gustafson-Kessel fuzzy classification algorithm in the pattern recognition of partial discharge for GIS. Power Syst. Prot. Control **39**, 50–54 (2011)
34. Hadiloo, S., Mirzaei, S., Hashemi, H., Beiranvand, B.: Comparison between unsupervised and supervised fuzzy clustering method in interactive mode to obtain the best result for extract subtle patterns from seismic facies maps. Geopersia **8**, 27–34 (2018). https://doi.org/10.22059/GEOPE.2017.240099.648346
35. Ghosh, A., Mishra, N.S., Ghosh, S.: Fuzzy clustering algorithms for unsupervised change detection in remote sensing images. Inf. Sci. **181**, 699–715 (2011). https://doi.org/10.1016/j.ins.2010.10.016
36. Peng, Z., Zhang, X., Cao, J., Liu, B.: Recognition of biological tissue denaturation based on improved multiscale permutation entropy and GK fuzzy clustering. Information **13**, 140 (2022). https://doi.org/10.3390/info13030140
37. Zheng, J., Gong, Z., Yin, S., Wang, W., Wang, M., Lin, P., Zhou, H., Yang, Y.: Rapid determination of lambda-cyhalothrin residues on Chinese cabbage based on MIR spectroscopy and a Gustafson-Kessel noise clustering algorithm. RSC Adv. **12**, 18457–18465 (2022). https://doi.org/10.1039/d2ra01557a
38. ErbayDalkılıç, T., Akbaş, Y., Akbaş, S.: A novel approach for portfolio optimization using fuzzy AHP based on Gustafson Kessel clustering algorithm. Mehmet Akif Ersoy Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi (2024). https://doi.org/10.30798/makuiibf.1469103

39. Markiewicz, T., Dziekiewicz, M.: Quantification of the ilio-femoral arteries tortuosity and data cluster modelling for preoperative examination. Biocybern. Biomed. Eng. **42**, 1123–1136 (2022). https://doi.org/10.1016/j.bbe.2022.09.002
40. Selamat, A., Ismail, M.K.: Effectiveness of relevance feedback for Content Based Image Retrieval using Gustafson-Kessel algorithm. In: 2008 International Conference on Computer and Communication Engineering, IEEE. pp. 455–459. https://doi.org/10.1109/ICCCE.2008.4580646 (2008)
41. Podenok, L.P., Sadykhov, R.Kh.: Multispectral satellite image segmentation using fuzzy clustering and nonlinear filtering methods. In: 2008 International Machine Vision and Image Processing Conference, IEEE, 2008: pp. 43–48. https://doi.org/10.1109/IMVIP.2008.18 (2008)
42. Wang, W., Zhang, Y.: On fuzzy cluster validity indices. Fuzzy Sets Syst. **158**, 2095–2117 (2007). https://doi.org/10.1016/j.fss.2007.03.004
43. Kim, M., Ramakrishna, R.S.: New indices for cluster validity assessment. Pattern Recognit. Lett. **26**, 2353–2363 (2005). https://doi.org/10.1016/j.patrec.2005.04.007
44. Bezdek, J.C.: Cluster validity with fuzzy sets. J. Cybern. **3**, 58–73 (1973). https://doi.org/10.1080/01969727308546047
45. S. Sengupta, S. De, A. Konar, R. Janarthanan, An improved fuzzy clustering method using modified Fukuyama-Sugeno cluster validity index. In: 2011 International Conference on Recent Trends in Information Systems, IEEE. pp. 269–274. https://doi.org/10.1109/ReTIS.2011.6146880 (2011)
46. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. **13**, 841–847 (1991). https://doi.org/10.1109/34.85677
47. Pal, N.R., Bezdek, J.C.: On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Syst. **3**, 370–379 (1995). https://doi.org/10.1109/91.413225
48. Kim, D.W., Lee, K.H., Lee, D.: On cluster validity index for estimation of the optimal number of fuzzy clusters. Pattern Recogn. **37**, 2009–2025 (2004). https://doi.org/10.1016/j.patcog.2004.04.007
49. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7
50. Chaimontree, S., Atkinson, K., Coenen, F.: Best clustering configuration metrics: Towards multiagent based clustering. pp. 48–59. https://doi.org/10.1007/978-3-642-17316-5_5. (2010)
51. Bezdek, J.C., Pal, N.R.: Cluster validation with generalized Dunn's indices. In: Proceedings 1995 second New Zealand international two-stream conference on artificial neural networks and expert systems, IEEE Comput. Soc. Press, pp. 190–193. https://doi.org/10.1109/ANNES.1995.499469 (1995).
52. Ncir, C.-E., Hamza, A., Bouaguel, W.: Parallel and scalable dunn index for the validation of big data clusters. Parallel Comput. **102**, 102751 (2021). https://doi.org/10.1016/j.parco.2021.102751
53. Davies, D.L., Bouldin, D.W.: A cluster seperation measure. Trans. Pattern Anal. Mach. Intell, **1**, 224–227 (1977)
54. Manuel Andrade-Garda, J., Carlosena-Zubieta, A., María Soto-Ferreiro, R., Teran-Baamonde, J. , Thompson, M.: Classical linear regression by the least squares method. In: Basic Chemometric Techniques in Atomic Spectroscopy, The Royal Society of Chemistry, 2013: pp. 52–122. https://doi.org/10.1039/9781849739344-00052.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yeşim Akbaş[1] · Serkan Akbaş[1] · Türkan Erbay Dalkılıç[1]

✉ Yeşim Akbaş
yesimyeginoglu@ktu.edu.tr

Serkan Akbaş
serkanakbas@ktu.edu.tr

Türkan Erbay Dalkılıç
terbay@ktu.edu.tr

[1] Department of Computer Science, Faculty of Science, Karadeniz Technical University, Trabzon, Turkey