Indian Journal of Engineering & Materials Sciences
Vol. 32, June 2025, pp. 364-378
DOI: 10.56042/ijems.v32i03.15474

Developing an integrated clustering algorithm based on fuzzy *C*-means for characterizing sound signals of vibration test rig

Teuku Edisah Putra[a*], Hizir Sofyan[b], Arif Saputra[b,c], Dien Lessy[d], Husaini[a] & Teuku Ariessa Sukhairi[a]

[a]Department of Mechanical Engineering, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[b]Department of Statistics, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
[c]Department of Epidemiology, Prince of Songkla University, Hat Yai 90110, Thailand
[d]Institute of Digital Signal Processing, Universität Duisburg-Essen, Duisburg 47057, Germany

Received: 03 December 2024; accepted: 19 February 2025

Fuzzy-based integrated clustering application, or FICA, has been developed in this research. It effectively performs data clustering by considering seven clustering quality indexes, namely the partition coefficient, classification entropy, modified partition coefficient, Fuzzy silhouette width, Xie-Beni, partition index, and separation index, leading to more optimal cluster results. In addition, FICA is also equipped with correlation coefficient and coefficient of determination functionalities that describe the relationship in the clustered data. For validation purposes, a data set of 30 sound signals with sampling frequencies ranging from 11-20 kHz has been measured with a voltage of 6, 9, and 12 Volts for 30 seconds. Similarly, key parameters such as sampling frequency, sound pressure level, and power spectral density have been obtained from sound signals and have been clustered and compared with open-source software. Validation results have shown that only Fuzzy silhouette width reports different results, although the difference does not affect the selection of the optimal number of clusters. Both FICA and the software has recommended the same number of clusters, which was 2, for the two and three dimensions. In conclusion, an integrated, user-friendly, and accurate clustering application for engineering data has been successfully developed.

Keywords: Clustering, Clustering quality index, Coefficient of determination, Correlation coefficient, Fuzzy *C*-means, Optimal number of clusters

1 Introduction

Clustering is anunsupervisedata mining method that is very important for complex decision-making in various domains. It is divided into hierarchical and non-hierarchical techniques. Subsequently, the hierarchical technique clusters data have similar characteristics in a structured manner where the number of clusters is unknown. The number of clusters is obtained by agglomerative (merging) and divisive (separating) methods. While the non-hierarchical technique clusters a set of data based on centroids with a predetermined number of clusters. It is preferable for large data sets due to the higher speed than the hierarchical technique.

Several clustering software tools are available, but most are developed using a programming language. Therefore, additional skills are needed to understand the algorithm and mathematical models as well as translate commands into the programming language for easy usage. Selecting relevant variables is crucial for meaningful clustering, avoiding empirical variable clusters.

Determining the optimal number of clusters is a crucial task, often requiring the experimentation of different values. This is because many clustering algorithms optimize objective functions that may not be in line perfectly with the desired quality. Consequently, various algorithms may yield locally optimal solutions. To obtain an optimal objective function, it is important to evaluate the resulting partition to improve the quality through a process known as cluster validity. This method assesses the quality of clustering results, particularly useful for comparing different algorithms.

Considering the importance of data clustering and the limited programming skills among engineering professionals, this research aims to develop an algorithm based on Fuzzy *C*-Means (FCM). This algorithm is implemented using an application known as Fuzzy-based Integrated Clustering Application (FICA). In this application, the optimal number of clusters is selected based on the clustering quality

*Corresponding author (E-mail: edi@usk.ac.id)

indexes. Additionally, it presents correlation coefficient and coefficient of determination to measure the strength of the relationship between data.

2 Materials and Methods

One of the most popular adopted non-hierarchical techniques is *K*-Means, developed by MacQueen[1]. Subsequently, FCM, proposed by Dunn[2] and later popularized by Bezdek[3], was developed from *K*-Means using Fuzzy logic[4]to reallocate data into each cluster. FCM has some advantages[5,6], such as including more effective and smoother results, simpler, easy to implement, ability to clyahoo assify larger datasets, robustness, and increased tolerance to outliers. In addition, it is believed that FCM outperforms *K*-Means in clustering data and has been applied in engineering data analysis for over a decade[7].

FCM was used to construct FICA withthe use of commercial programming language software. The clustering process in FICA is summarized in Fig. 1. This process started with data input. If a data set *U*is available, both input and output from Fuzzy system can be represented as follows:

$$U = (u_1, u_2, u_3, \dots, u_n) \quad \dots(1)$$

The data is represented in the form of an $n \times m$ matrix which consists of an*x* and *y*-axis for two-dimensional (2-D) clustering or *x*, *y*, and *z*-axis for three-dimensional (3-D) clustering. The form of the matrix X_{ij} with data on the i^{th} observation (i = 1, 2, 3, …, n) and the j^{th} variable (j = 1, 2, 3, …, m) can be written:

$$X_{ij} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{bmatrix} \quad \dots(2)$$

Subsequently, the expected smallest error/ threshold ξ, maximum iteration $a_{\max}$, fuzziness w, and the maximum number of clusters C are also determined.

Once all data inputs are completed, FICA proceeds to calculate the membership degrees, which show the probability of data points belonging to a particular cluster. The membership degree μ_{ik} of the i^{th} data in the k^{th}cluster can be expressed as:

$$\mu_{ik}(U_i) \epsilon [0,1] \text{with} (1 \le i \le n; 1 \le k \le C) \quad \dots(3)$$

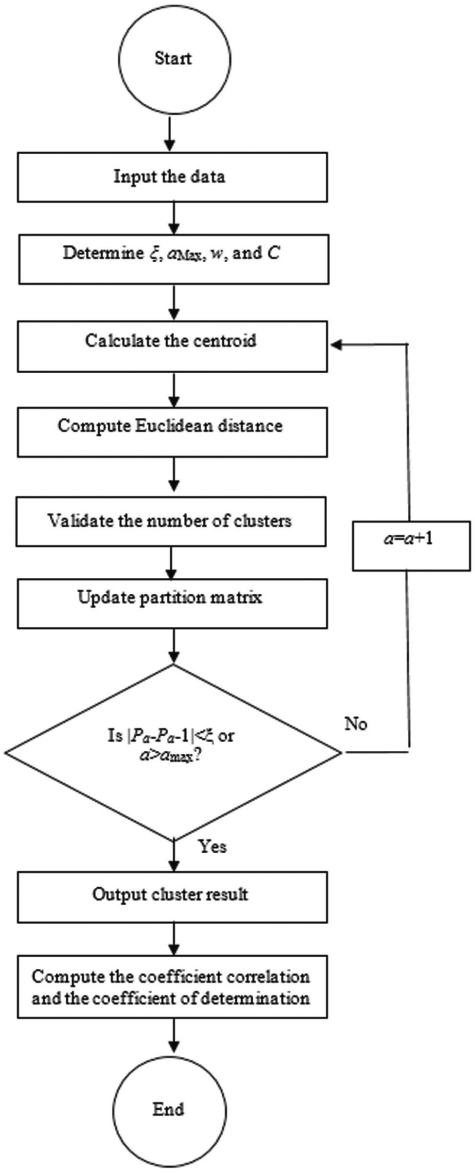

Fig. 1 — The flow diagram of FICA.

Random numbers of membership degree μ_{ik}(i= 1, 2, 3, …, n; k= 1, 2, 3, …, C),serving as elements in the initial partition matrix U can be expressed below:

$$\mu_{ik} = \begin{bmatrix} \mu_{11}[\mu_1] & \mu_{12}[\mu_1] & \cdots & \mu_{1C}[\mu_1] \\ \mu_{21}[\mu_2] & \mu_{22}[\mu_2] & \cdots & \mu_{2C}[\mu_2] \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n1}[\mu_n] & \mu_{n2}[\mu_n] & \cdots & \mu_{nC}[\mu_n] \end{bmatrix} with \sum_{k=1}^{C} \mu_{ik} = 1 \quad \dots(4)$$

The total sum of each column in this matrix can be calculated by:

$$Q_k = \sum_{i=1}^{n} \mu_{ik} \quad \ldots(5)$$

This process includes calculating the membership degree for each data and assigning each data to the nearest centroid, ensuring that data points within a cluster are as close to each other as possible and as far apart from members of other clusters as possible. A data point can belong to multiple clusters, reflecting the probability of its position in each cluster. The membership degree can be presented by arbitrary values in the interval from 0 to 1[8].This implies that the total membership degree of data across all clusters should be equal to 1. Data with a higher probability of belonging to one cluster will have membership degree close to 1,while for another cluster, it will be close to 0. For example, a data point can have membership degree of 0.5 for cluster I, 0.3 for cluster II, and 0.2 for cluster III.

Once all clusters are formed, the centroid of the k^{th}cluster on the j^{th} variable for each cluster is determined using the formula:

$$V_{kj} = \frac{\sum_{i=1}^{n}((\mu_{ik})^{w} X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{w}} \quad \ldots(6)$$

The position of each data point is determined based on membership degree,using Euclidean distance[9] as Fuzzy index, considering the predetermined number of clusters. The advantage of this method is that adding a new data point will not affect the distance between two existing data, allowing data points to be members of multiple clusters due to its ability to measure geometric distances in multi-dimensional space[10,11,12].

Euclidean distance between the matrix X_{ij} and the centroid V_{kj} can be written as[13]:

$$d_{ik} = d\left(X_{ij} - V_{kj}\right) \quad \ldots(7)$$

$$d_{ik} = \sqrt{(X_{i1} - V_{k1})^2 + (X_{i2} - V_{k2})^2 + \cdots + (X_{im} - V_{km})^2} \quad \ldots(8)$$

Changes in the membershipdegree of partition matrix elements can be calculated by:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right]^{\frac{-1}{w-1}}} \quad \ldots(9)$$

The number of clusters is selected based on the clustering quality indexes. Subsequently, several clustering quality indexes have been proposed in the literature, but none is perfect. Therefore, in this research, seven clustering quality indexes were used.

Partition coefficient (PC) index introduced by Bezdek[14]is to measure the quality of clustering using membership degree of data. The optimal number of clusters is determined by the largest index value, ranging from 1/*C* to 1. Clusters with index value close to 1/*C* suggest that the clustering algorithm is struggling to group the data effectively. Therefore, clusters are considered more optimal when PC index value approaches 1. PC index value can be calculated by:

$$PC = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{C}(\mu_{ik})^2 \quad \ldots(10)$$

Bezdek[14] also proposed classification entropy (CE) index to measure the degree of cluster blurring using the degree of membership and data. CE index value is in the range of 0 to ln(*C*). Clusters are considered more optimal when CE value approaches 0, showing a higher degree of data separation and less cluster blurring. CE index value can be calculated by:

$$CE = -\frac{1}{n}\left(\sum_{i=1}^{n}\sum_{k=1}^{C}(\mu_{ik}\, log(\mu_{ik}))\right) \quad \ldots(11)$$

To address the limitations of PC and CE indexes, a modified partition coefficient (MPC) index[15] can be calculated by using the following formula[10]:

$$MPC = 1 - \frac{C}{C-1}(1 - PC) \quad \ldots(12)$$

Fuzzy silhouette width (FSW) proposed by Rousseeuw[16]is anextension of silhouette width index used to determine the optimal number of clusters using Fuzzy logic[17]. FSW index uses the membership degree on the matrix and silhouette width index. FSW index can be obtained using the following equation[18]:

$$FSW = \frac{\sum_{i=1}^{n}(\mu_{ir} - \mu_{iq})^{w} S_i}{\sum_{i=1}^{n}(\mu_{ir} - \mu_{iq})^{w}} \quad \ldots(13)$$

with:

$$S_i = \frac{b_i - a_i}{max(b_i, a_i)} \quad \ldots(14)$$

$$a_i = \frac{1}{n(C_{(i)})}\sum_{j \in C_{(i)}} d_{ij} \quad \ldots(15)$$

$$b_i = min_{C_k \in C \setminus C_{(i)}} \sum_{j \in C_{(k)}} \frac{d_{ij}}{n(C_k)} \quad \ldots(16)$$

where $\mu_{ir}$is the first largest element in the partition matrix, μ_{iq} is the second largest element in the partition matrix, $a_i$is the mean distance between i^{th}

data and other objects in the same cluster, b_i is the mean distance between i^{th} data and other objects in different clusters, d_{ij} is the distance between the i^{th} and j^{th} data, C_i is the distance between data in the same cluster, and C_k is the distance between data in the different clusters.

Xie-Beni (XB) index proposed by Xie & Beni[19] calculates the compactness of the mean separation data between clusters. This includes the distance between the data points and their respective centroid as well as the distance between the different centroids[20]. A more optimal clusteris achieved when XB index value decreases[12] but tends to change monotonically. To address this challenge, it is necessary to determine the maximum number of clusters that will be examined first. In addition, XB value is also influenced by the weighting rank, with a higher rank leading to a near-infinite XB value. It can be calculated by:

$$XB = \frac{\sum_{i=1}^{n}\sum_{k=1}^{C}(\mu_{ik})^{w}\|X_i - V_k\|^2}{n * min_{ik}\|V_i - V_k\|^2} \quad \ldots(17)$$

where $\|X_i\text{-}V_k\|$ is the Euclidean distance from the data point X_i to the centroidV_k, and $\|V_i\text{-}V_k\|$ is the Euclidean distance between the centroids.

The partition index (SC) measures the ratio of the sum of compactness and separation of clusters. It is a sum of individual cluster validity measures normalized by Fuzzy cardinality of each cluster. A lower value of SC shows a better partition, and it can be defined as[21]:

$$SC = \sum_{k=1}^{C}\frac{\sum_{i=1}^{n}(\mu_{ik})^{w}\|X_i - V_k\|^2}{n\sum_{i=1}^{n}\|V_i - V_k\|^2} \quad \ldots(18)$$

In contrast to SC,the separation index (S) uses a minimum-distance separation for partition validity. It is defined by[21]:

$$S = \frac{\sum_{i=1}^{n}\sum_{k=1}^{C}(\mu_{ik})^{2}\|X_i - V_k\|^2}{n min_{ik}\|V_i - V_k\|^2} \quad \ldots(19)$$

To minimize variation within one cluster and maximize variation between clusters, the initial objective function is set at P_0= 0. The objective function in thea^{th} iteration P_a can be calculated by:

$$P_a = \sum_{i=1}^{n}\sum_{k=1}^{C}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right](\mu_{ik})^{w}\right) \quad \ldots(20)$$

This iteration process calculates the centroid of the k^{th}cluster in the a^{th} iteration, where $a = a + 1$. This process continues by repeatedly moving the centroid to various places as long as (i) the change in the membership degree is still above the threshold, (ii) the change in the centroid is still above the threshold, or (iii) the change in the objective function is still above the threshold. This process stops when the closest distance is obtained, there is no further transfer of objects between clusters or the maximum number of iterations has been reached, where$|P_a\text{-}P_a\text{-}1|<\xi$ or $a>a_{\max}$.

The relationship between clustered data can be determined. Subsequently, the correlation coefficient developed by Pearson[22] is a measure of the linear relationship between two random variables and is used to measure the extent to which the points in the data lie around a straight line. For determining the correlation coefficient, the relationship between the variables should be linear. This factor estimates the interdependence between two random variables and represents the percentage value of a point that is the nearest to a line of best fit. The value varies between -1 to +1, while 0 represents the absence of any correlation.

A negative correlation indicates that when one variable increases, the other decreasesin accordance with a linear relationship. Meanwhile, a positive correlation value shows that as one variable value increases, the value of the second variable also increases, following a linear relationship. The correlation coefficient serves as dependable indicator of the strength of this linear relationship when therelationship between the variables is genuinely linear. However, when the relationship or pattern between these variables isnon-linear, the reliability and validity of the correlation coefficient diminish. Additionally, when a regression line intersects every point on a scatter plot, it can account for all variations, but as the line moves farther from these points, its ability to explain variations decreases.

The Pearson correlation coefficient r between random variables X and Y can be written as:

$$r = \frac{n\sum_{i=1}^{n}X_iY_i - \left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n}X_i^2 - \left(\sum_{i=1}^{n}X_i\right)^2\right]\left[n\sum_{i=1}^{n}Y_i^2 - \left(\sum_{i=1}^{n}Y_i\right)^2\right]}} \quad \ldots(21)$$

where X_i and Y_i are the values of the i^{th} random variables X and Y, respectively, with $i = 1, 2, 3, \ldots, n$.

The coefficient of determination R^2 developed by Wright[23] explains the proportion of the total variation in the dependent variable values that can be accounted for by the independent variable through a linear relationship. The coefficient of determination value

ranges from 0 to 1 and can be obtained using the equation:

$$R^2 = 1 - \frac{\sum(X_i - Y_i)^2}{\sum(X_i - \bar{x})^2} \quad \ldots(22)$$

where $\bar{\lambda}$ is the mean value.

The development of FICA was started by translating the equations above into algorithms. These algorithms are displayed in a Graphical User Interface, making it easier for users. FICA is provided in the form of an executable file (*.exe) and a setup installation package which can be run directly on a computer with a minimum operating system of Windows 7, processor i3, memory 2 GB, Internet connection 10 Mbps and hard disk 64 GB.

In Fig. 2, Part *A* provides a menu for uploading data for the *x*, *y*, and *z* axes. Part *B* contains the initial parameters, namely threshold, maximum iteration, fuzziness, and maximum number of clusters. By pressing the RUN button, FICA tabulates the values for each clustering quality index in Part *C*. Part *D* displays the optimum value and number of clusters recommended by each clustering quality index. Visualization of the distribution of data selected based on the most recommendations in Part *D* is displayed in Part *E*. Finally, the correlation coefficient and coefficient of determination are displayed in Part *F*.

3 Results and Discussion

To test the capability, FICA was used to cluster sound signals acquired at vibration test rig (Fig. 3). This tool has a steel rod supported at both ends and exciter with unbalanced disc, which is installed in the middle of the rod. The disc was rotated by applying a voltage to exciter of 6, 9, and 12 Volts, with

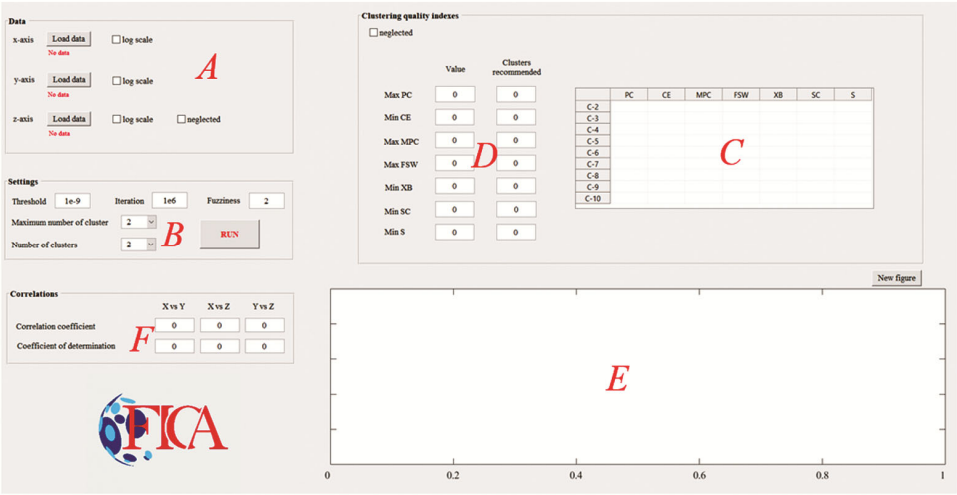

Fig. 2 — View of FICA.

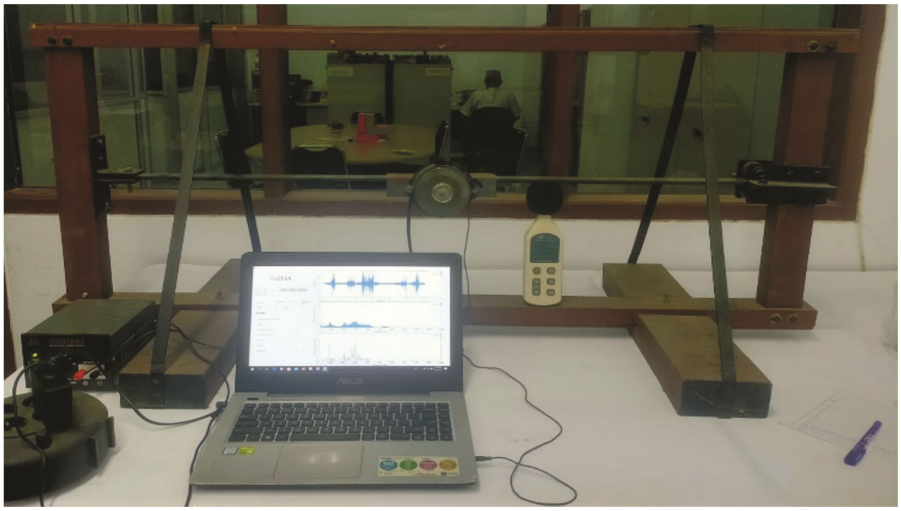

Fig. 3 — Process of measuring sound signals on vibration test rig.

1,468,7,545 and 10,906 radiant per minute, respectively. Sound signals at each voltage were measured with a sampling frequency of 11-20 kHz for 30 seconds, and in total, 30 sound signals were obtained. During the measurement process, a sound level meter was also used to measure the sound pressure level (SPL)for each sound signal.

Sound signals produced at the sampling frequency of 11-20 kHz are shown in Figs (4-6), and the SPL values are tabulated in Table 1. The sampling

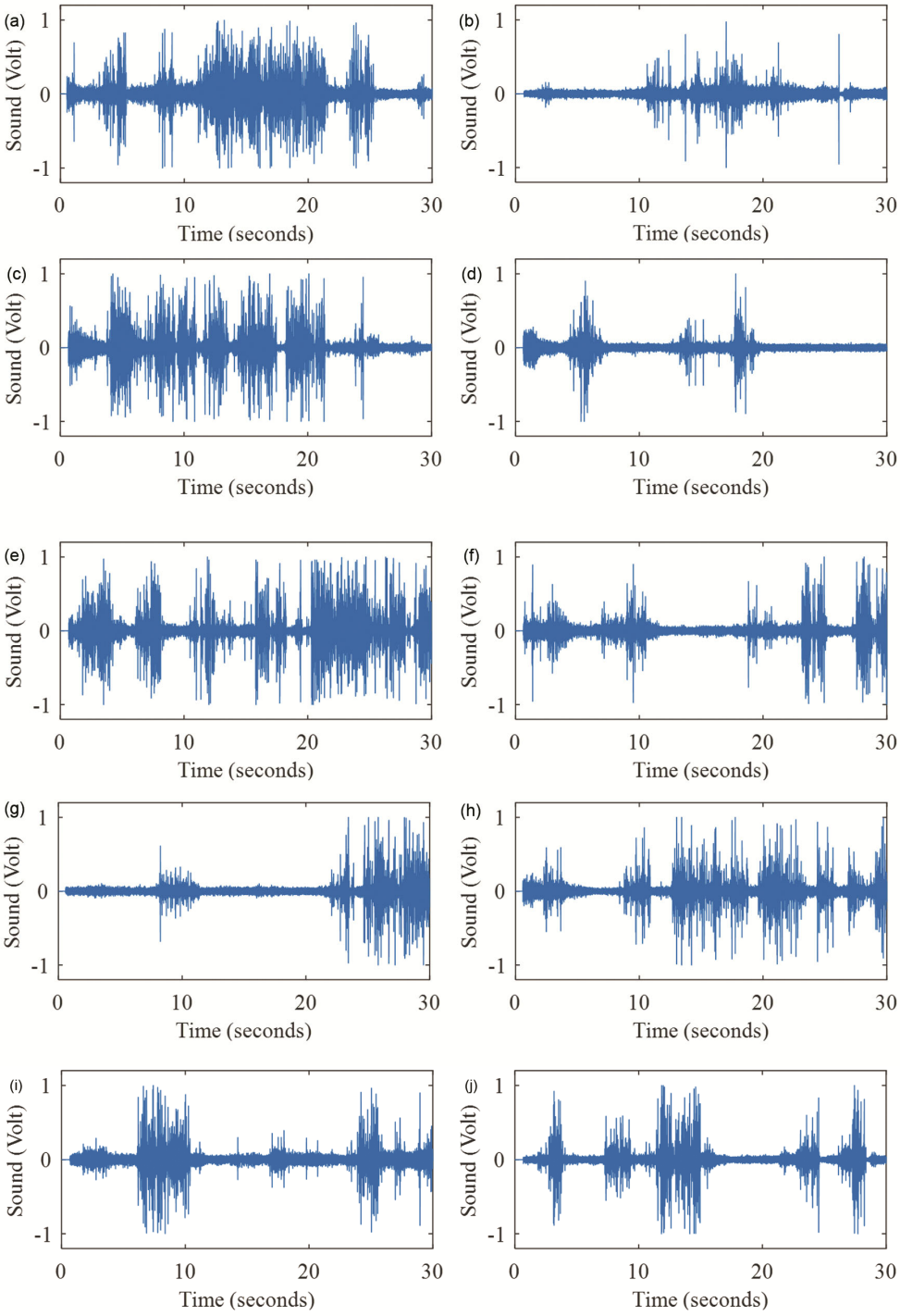

Fig. 4 — Sound signals measured at 6 Volts with various sampling frequencies: (a) 11 kHz, (b) 12 kHz, (c) 13 kHz, (d) 14 kHz, (e) 15 kHz, (f) 16 kHz, (g) 17 kHz, (h) 18 kHz, (i) 19 kHz, and (j) 20 kHz.

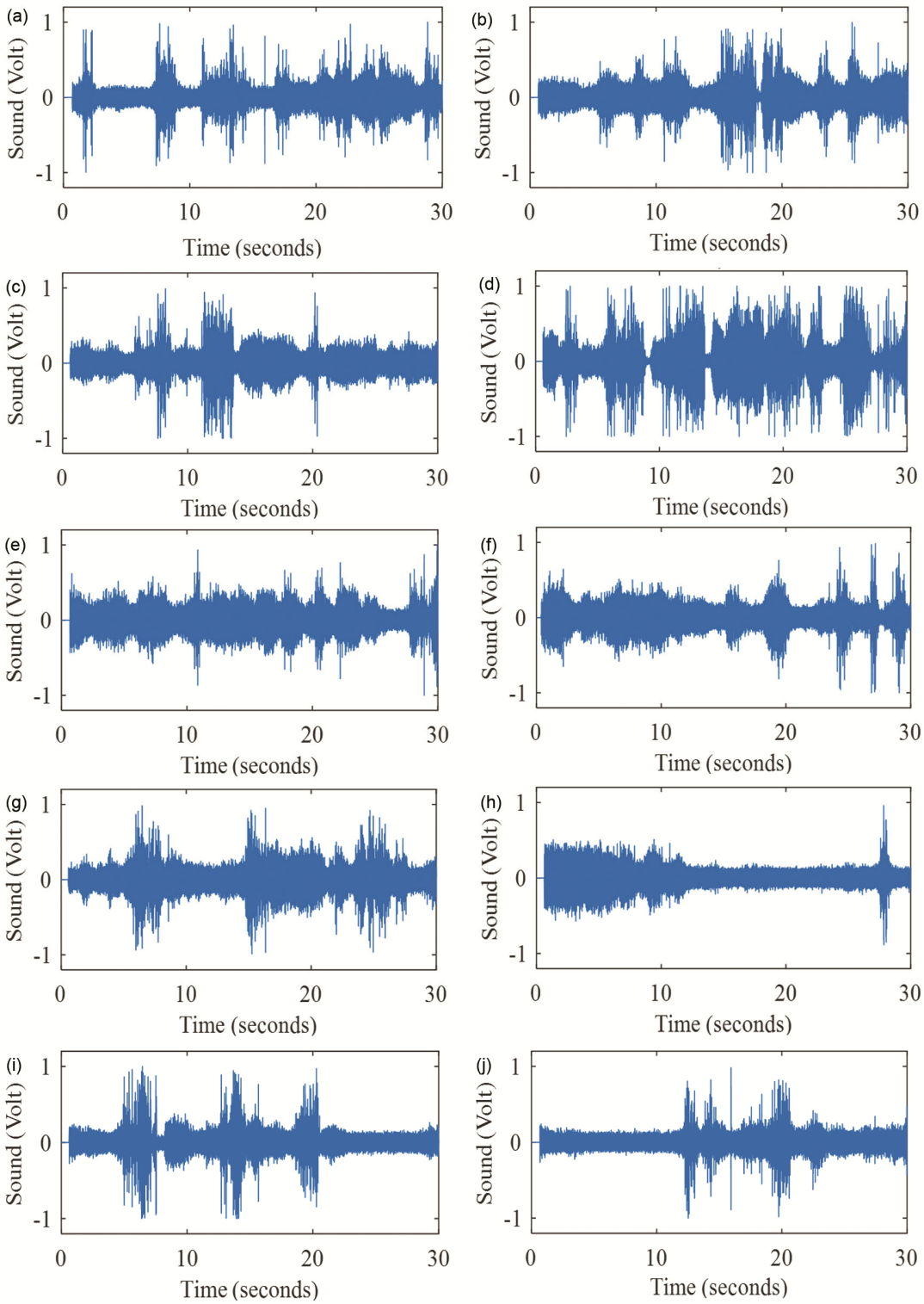

Fig. 5 — Sound signals measured at 9 Volts with various sampling frequencies: (a) 11 kHz, (b) 12 kHz, (c) 13 kHz, (d) 14 kHz, (e) 15 kHz, (f) 16 kHz, (g) 17 kHz, (h) 18 kHz, (i) 19 kHz, and (j) 20 kHz.

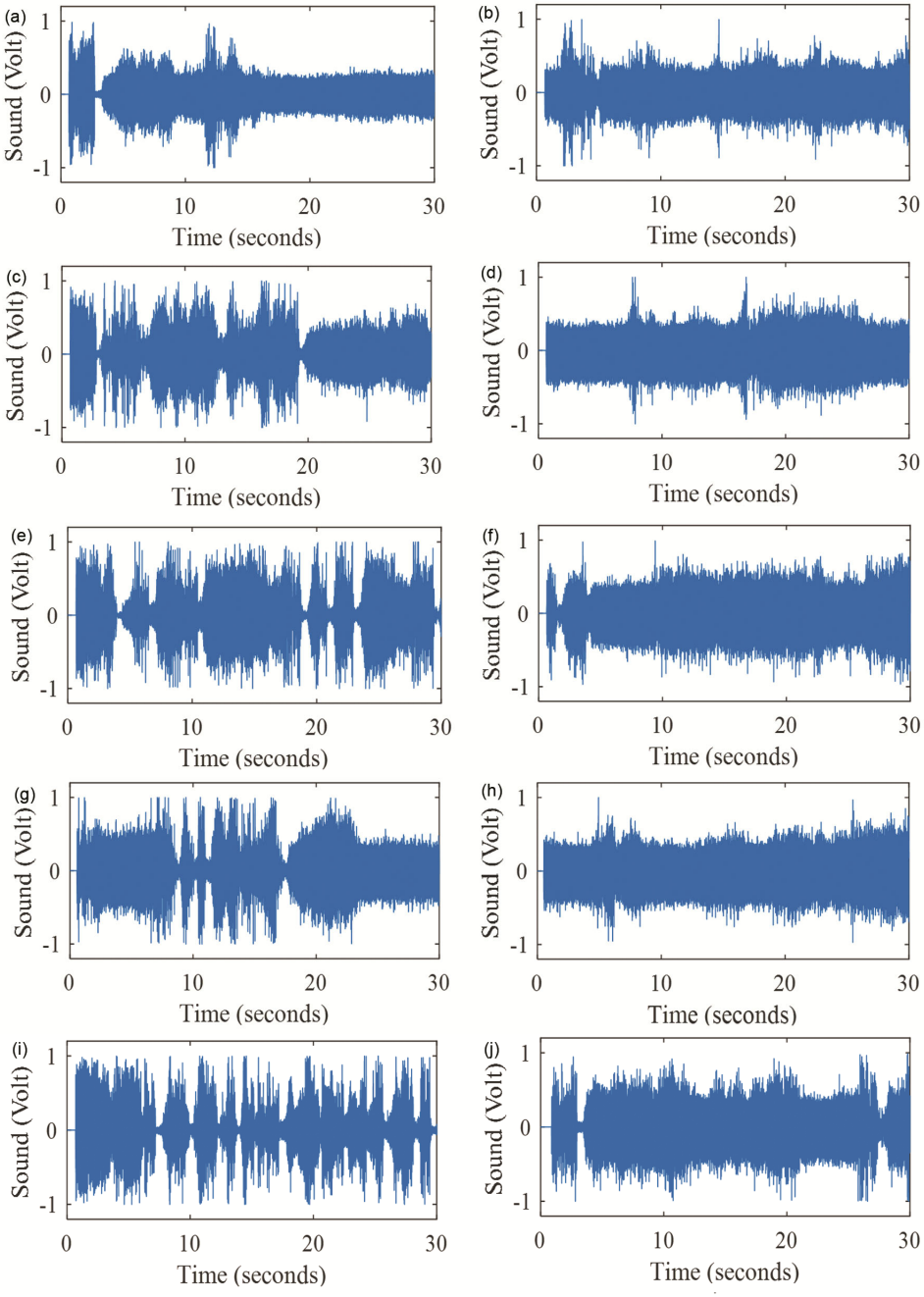

Fig. 6 — Sound signals measured at 12 Volts with various sampling frequencies: (a) 11 kHz, (b) 12 kHz, (c) 13 kHz, (d) 14 kHz, (e) 15 kHz, (f) 16 kHz, (g) 17 kHz, (h) 18 kHz, (i) 19 kHz, and (j) 20 kHz.

frequency and SPL were then clustered with the threshold and the maximum iteration were 1E-9 and 1E6, respectively. Meanwhile, the fuzziness was $2^{24}$and the maximum number of clusters was7.

The number of clusters was selected based on the largest PC, MPC and FSW values and the smallest CE, XB, SC and S values. Therefore, FICA process started by calculating values generated from each index, and identified the maximum index values of PC, MPC, and FSW as well as the minimum index values of CE, XB, SC, and S, as tabulated in Table 2. Based on the index values, PC, CE, and S recommended 2 clusters, FSW and XB recommended 5 clusters, and MPC and SC recommended 7 clusters.

The optimal number of clusters was 2 because it was recommended by most indexes, namely PC, CE, and S. As shown in Fig. 7, it only divided the data based on the sampling frequency in the *x*-axis, having a higher range compared to SPL in *y*-axis. Low sampling frequencies are in cluster 1 and high sampling frequencies are in cluster 2. Different conditions are shown in Fig. 8 with 5 clusters recommended by FSW and XB as well as in Fig. 9 with 7 clusters recommended by MPC and SC. This method divided the data into smaller clusters.

The correlation coefficient and the coefficient of determination between the sampling frequency and SPL were 0.187 and 0.035, respectively. This shows that there was a positive relationship between these two parameters, if the sampling frequency is high, then the noise is also high. Conversely, if the frequency is low, then the noise is also low.

The constituent frequency components of sound signals were identified. One of the most commonly used methods in frequency domain analysis is power spectral density (PSD). It is an analysis spectrum taking into account the signal energy, which can be expressed as:

$$S_{(f)} = \frac{1}{n}\left|X_{(f)}\right|^2 \quad \ldots(23)$$

where $X_{(f)}$is the complex frequency domain representation at frequency index to the input sequence in the time domain at time index $x_{(i)}$, stated by:

$$X_{(f)} = \sum_{t=0}^{n-1} x_{(i)} e^{-i\frac{2\pi}{n}ft} \quad \ldots(24)$$

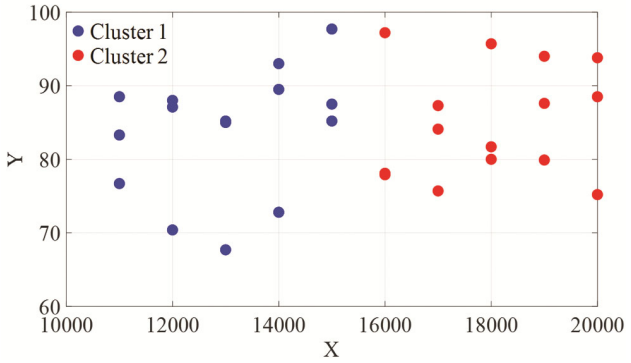

Fig. 7 — The 2-D FICA clustering results for 2 clusters.

Table 1 — SPL values for each sound signal.

Sampling frequency	Voltages		
	6 Volts	9 Volts	12 Volts
11 kHz	76,7	83,3	88,5
12 kHz	70,4	88	87,1
13 kHz	67,7	85,2	85
14 kHz	72,8	93	89,5
15 kHz	85,2	87,5	97,7
16 kHz	78,1	77,9	97,2
17 kHz	75,7	84,1	87,3
18 kHz	80	81,7	95,7
19 kHz	79,9	94	87,6
20 kHz	75,2	88,5	93,8

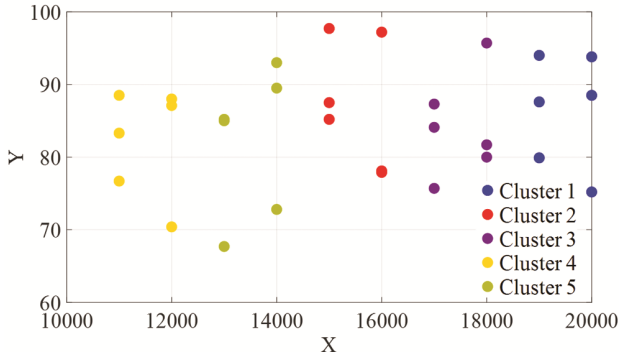

Fig. 8 — The 2-D FICA clustering results for 5 clusters.

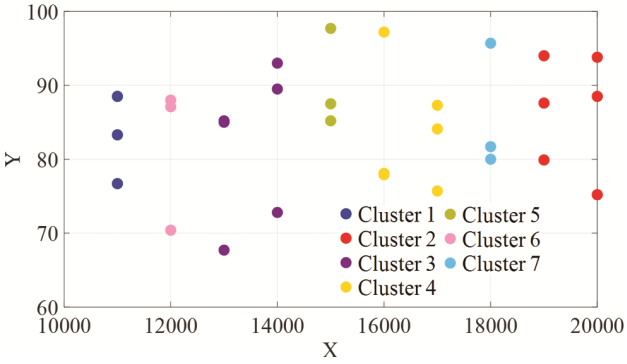

Fig. 9 — The 2-D FICA clustering results for 7 clusters.

Table 2 — Validation clusters for the 2-D resulted from FICA.

Recommended number of clusters	PC	CE	MPC	FSW	XB	SC	S
2	0.855	0.244	0.711	0.789	0.060	3.605	0.270
3	0.803	0.360	0.705	0.724	0.057	0.850	0.296
4	0.775	0.433	0.700	0.708	0.055	0.327	0.326
5	0.770	0.496	0.713	0.790	0.053	0.160	0.359
6	0.740	0.512	0.688	0.682	0.065	0.093	0.380
7	0.778	0.455	0.741	0.685	0.079	0.056	0.406

where f is the frequency, t is the time, and i is the imaginary unit.

The resulting PSDsfor each sound signal are shown in Figs (10-12). The area of the PSD graphs was then calculated and became a parameter for the 3-D clustering. Other parameters used were the same as in the 2-D clustering, namely threshold of 1E-9, maximum iteration of 1E6, fuzziness of 2, and maximum number of clusters of 7.

The values generated from each index are tabulated in Table 3. Based on the index values, the number of clusters selected was also 2 because it was recommended

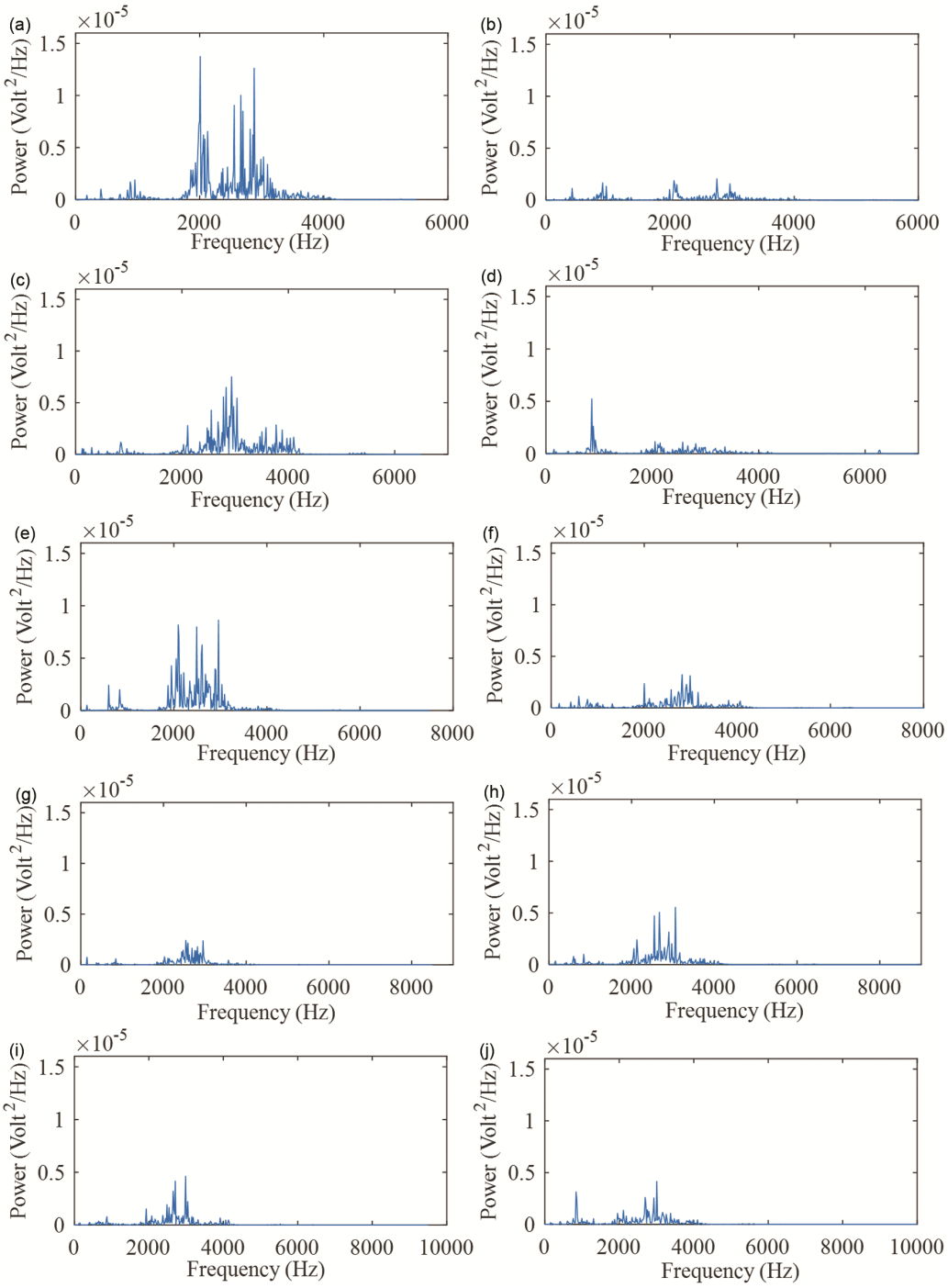

Fig. 10 — PSD of sound signals measured at 6 Volts with various sampling frequencies: (a) 11 kHz, (b) 12 kHz, (c) 13 kHz, (d) 14 kHz, (e) 15 kHz, (f) 16 kHz, (g) 17 kHz, (h) 18 kHz, (i) 19 kHz, and (j) 20 kHz.

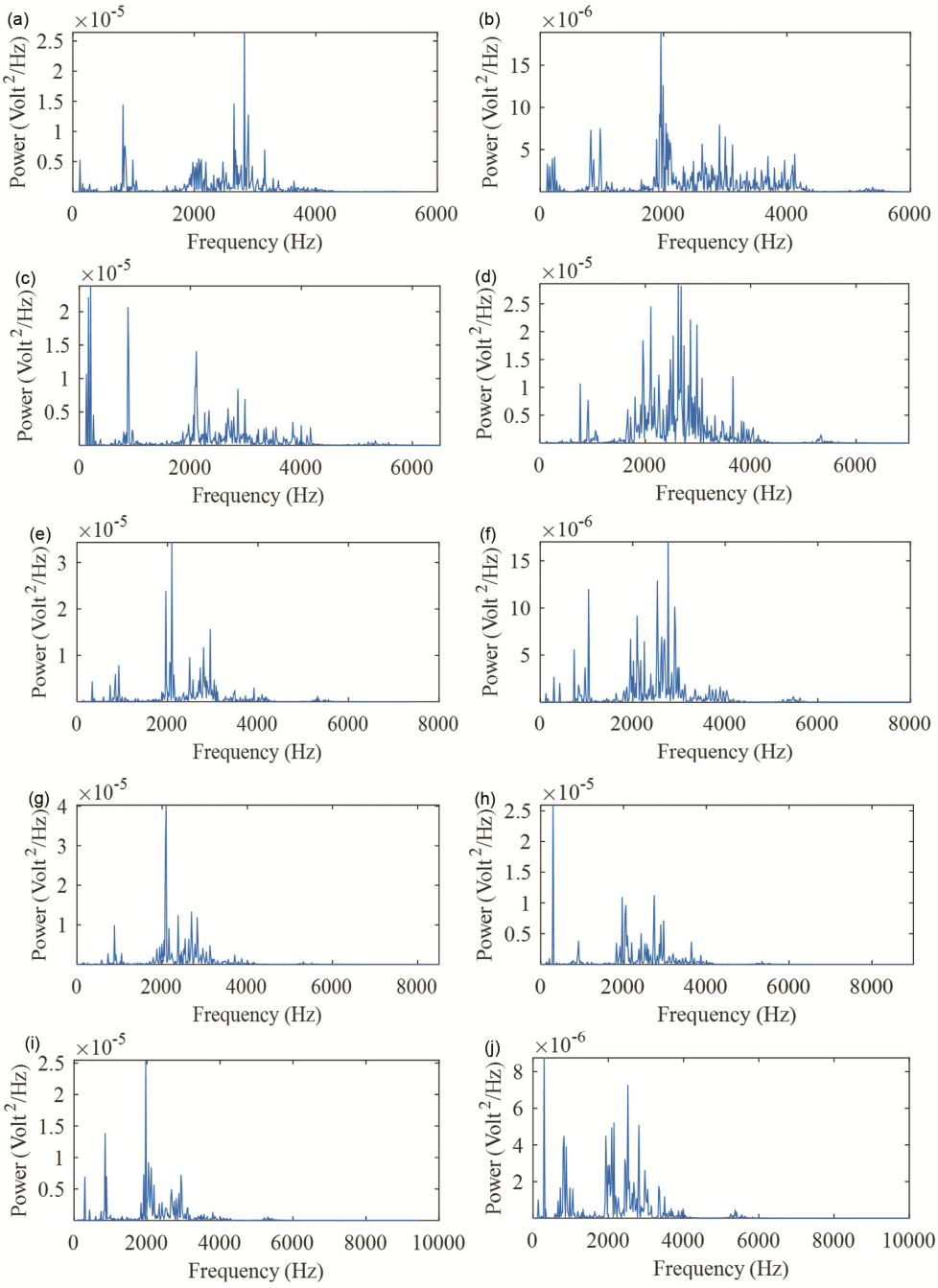

Fig. 11 — PSD of sound signals measured at 9 Volts with various sampling frequencies: (a) 11 kHz, (b) 12 kHz, (c) 13 kHz, (d) 14 kHz, (e) 15 kHz, (f) 16 kHz, (g) 17 kHz, (h) 18 kHz, (i) 19 kHz, and (j) 20 kHz.

by the majority of the indexes, including PC, CE, FSW, and S. Meanwhile, XB suggested 5 clusters, while MPC and SC recommended 7 clusters. The clustering results for 2, 5, and 7 clusters are shown in Figs (13-15), respectively. These divisions were primarily based on the sampling frequency along the *x*-axis, as it had the widest compared to SPL along the *y*-axis and PSD along the *z*- axis. The smaller number of clusters resulted in a finer subdivision of the data.

The correlation coefficient and the coefficient of determination between the sampling frequency and PSD were -0.21 and 0.044, respectively. Anegative correlation indicated that as the sampling frequency increased, the PSDdecreased, and vice versa.

Table 3 — Validation clusters for the 3-D resulted from FICA.

Number of clusters	PC	CE	MPC	FSW	XB	SC	S
2	0.855	0.244	0.711	0.789	0.060	3.605	0.270
3	0.803	0.360	0.705	0.724	0.057	0.850	0.296
4	0.775	0.433	0.700	0.708	0.055	0.327	0.326
5	0.770	0.496	0.713	0.789	0.053	0.160	0.359
6	0.740	0.512	0.688	0.682	0.065	0.093	0.380
7	0.777	0.461	0.739	0.685	0.078	0.051	0.406

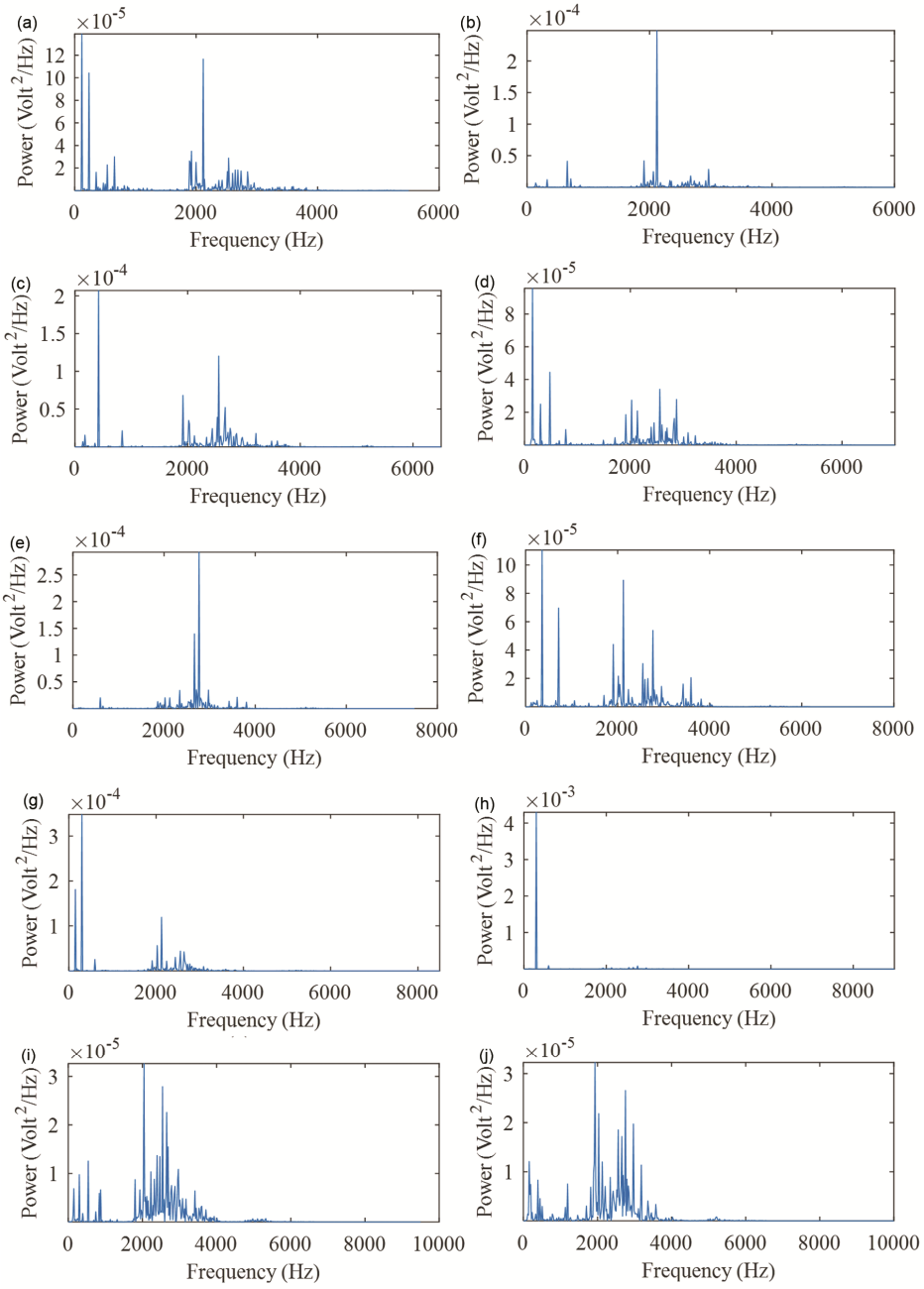

Fig. 12 — PSD of sound signals measured at 12 Volts with various sampling frequencies: (a) 11 kHz, (b) 12 kHz, (c) 13 kHz, (d) 14 kHz, (e) 15 kHz, (f) 16 kHz, (g) 17 kHz, (h) 18 kHz, (i) 19 kHz, and (j) 20 kHz.

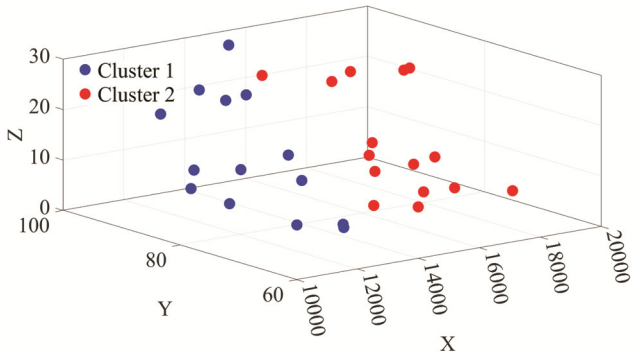

Fig. 13 — The 3-D FICA clustering results for 2 clusters.

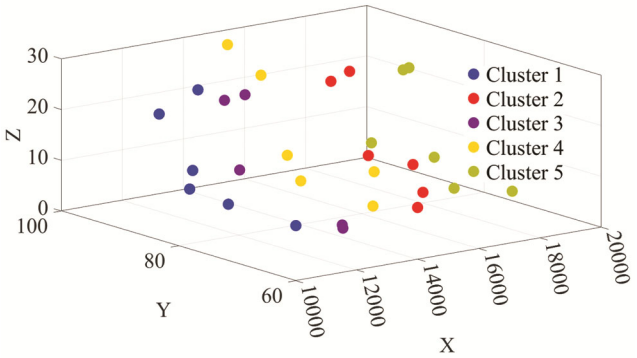

Fig. 14 — The 3-D FICA clustering results for 5 clusters.

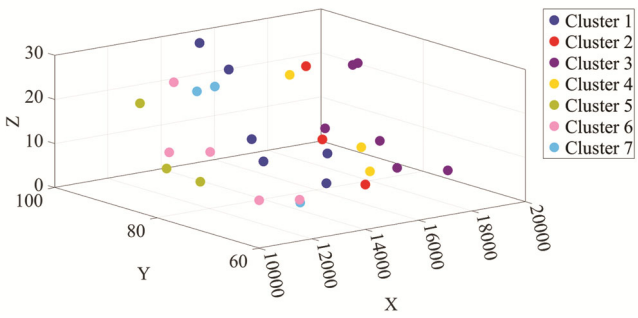

Fig. 15 — The 3-D FICA clustering results for 7 clusters.

Meanwhile, the correlation coefficient and the coefficient of determination between SPL and PSD were 0.658 and 0.433, respectively. This showed a stronger relationship between SPL and PSD compared to the relationship between sampling frequency and SPL or sampling frequency and PSD.

To validate the results, the open-source software *R* was used, which was a statistical software developed by Ross Ihahka and Robert Gentlemen. The software used the *S* programming language developed by John Chambers and colleagues at Bell Laboratories[25] and yielded satisfactory results in clustering strain signals[26].

The analysis performed with *R* software provided clustering quality indexes, as shown in Table 4. When compared with the values obtained from FICA, only FSW values showed differences for all clusters. Meanwhile, the other indices differed only for the 7th cluster. The value of *R* software was higher than the value obtained by FICA. The most significant difference occurred in the 4thand 6th clusters, namely 14%.

The differences did not affect the determination of the optimal number of clusters. The optimal number of clusters obtained from *R* software matched the recommendation of FICA, with 2 clusters being selected based on the guidance of PC, CE, FSW, and S. Meanwhile, XB suggested 5 clusters, while MPC and SC recommended 7 clusters. The 2-D clustering result generated by *R* software is depicted in Fig. 16. Visually, there was no discernible from the clustering results produced by FICA, as shown in Fig. 7. However, FICA graph is more interactive and flexible for editing. The correlation coefficient obtained was 0.187 while the coefficient of determination was 0.035. This value was the same as that obtained by FICA.

In the 3-D clustering analysis, Table 5presents the clustering quality indexes obtained from *R* software. When compared with FICA, differences also occurred

Table 4 — Validation clusters for the 2-D resulted from *R* software.

Number of clusters	PC	CE	MPC	FSW	XB	SC	S
2	0.855	0.244	0.711	0.843	0.060	3.605	0.270
3	0.803	0.360	0.705	0.822	0.057	0.850	0.296
4	0.775	0.433	0.700	0.825	0.055	0.327	0.326
5	0.770	0.496	0.713	0.794	0.053	0.160	0.359
6	0.740	0.512	0.688	0.795	0.065	0.093	0.380
7	0.783	0.445	0.747	0.785	0.107	0.066	0.400

Table 5 — Validation clusters for the 3-D resulted from *R* software.

Number of clusters	PC	CE	MPC	FSW	XB	SC	S
2	0.855	0.244	0.711	0.843	0.060	3.605	0.270
3	0.803	0.360	0.705	0.822	0.057	0.850	0.296
4	0.775	0.433	0.700	0.825	0.055	0.327	0.326
5	0.770	0.496	0.713	0.794	0.053	0.160	0.359
6	0.740	0.512	0.688	0.795	0.065	0.093	0.380
7	0.778	0.456	0.741	0.809	0.079	0.056	0.406

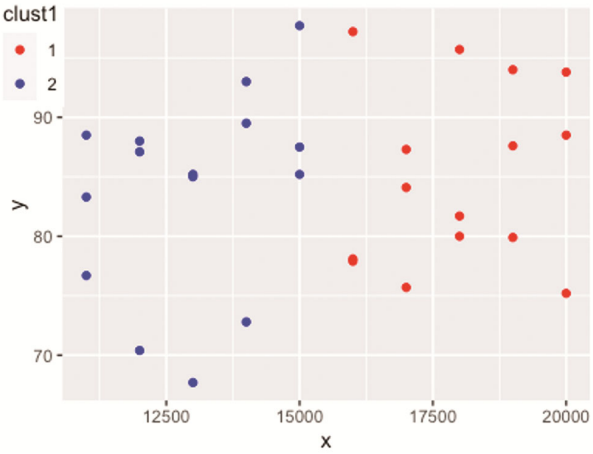

Fig. 16 — *R* software clustering results for the 2-D.

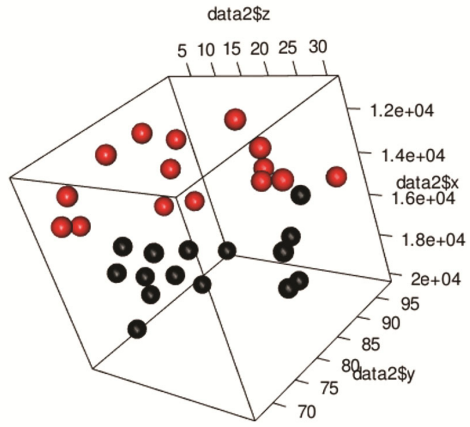

Fig. 17 — *R* software clustering results for the 3-D.

in FSW index, while the other indexes were the same. The value of *R* software was higher than the value given by FICA. The biggest difference occurred in the 7th cluster, namely 15%. However, the optimal number of clusters obtained from FICA was also the same as that recommended by *R* software. The number of clusters of 2 was selected based on the recommendations of PC, CE, FSW, and S. Meanwhile, XB recommended 5 clusters as well and MPC and SC recommended 7 clusters.

The 3-D clustering using *R* software is shown in Fig. 17. In contrast to FICA graph in Fig. 13, additional capabilities are required to edit the *R*-generated graph. For the relationship between sampling frequency and PSD, the correlation coefficient was -0.21 and the coefficient of determination was 0.044. Meanwhile, regarding the relationship between SPL and PSD, the correlation coefficient was 0.658 and the coefficient of determination was 0.433. These values were also the same as the values obtained by FICA.

4 Conclusion

This research aimed to develop a clustering application called FICA. This application's advantage lies in its ability to determine the optimal number of clusters by considering recommendations from various clustering quality indexes, including PC, CE, MPC, FSW, XB, SC, and S. The correlation coefficient and coefficient of determination are also shown. For simulation purposes, sampling frequency, sound pressure level, and power spectral density of sound signals were clustered and the results were compared to an open-source software. Based on the comparison, the same number of clusters was recommended. This means that, with just a few easy steps, users can obtain data clustering in the 2-D or the 3-D views as well as more accurate relationships between data.

Acknowledgements

The authors are grateful to the Ministry of Education, Culture, Research, and Technology, Indonesia, for the financial support provided through grant no. 634/UN11.2.1/PT.01.03/DPRM/2023.

References

1 Mac Queen J B, Proc 5th Berkeley Symp Math Stat & Prob, 1 (1967) 281.
2 Dunn J C, *J Cybern*, 3 (1974) 32.
3 Bezdek J C, Pattern Recognition with Fuzzy Objective Function Algorithms (Plenum Press, New York), 1981.
4 Zadeh L A, *Inform Control*, 8 (1965) 338.
5 Zhang W, Huang T & Chen J, *Math Probl Eng*, 2019, Article ID.
6 Irvanizam, Zi N N, Zuhra R, Amrusi & Sofyan H, *Axioms*, 9 (2020) 104.

7 Abdullah S, Putra T E, Nuawi M Z, Nopiah Z M, Arifin A & Abdullah L, *WSEAS Trans Math*, 5 (2010) 345.
8 Irvanizam, Syahrini I, Afidh R P F, Andika M R & Sofyan H, Proc 6th Int Conf Cyber & IT Serv Manage (CITSM), (2019) 1.
9 Ali S M & Silvey S D, *J Roy Stat Soc Ser B*, 28 (1966) 131.
10 Suleman A, *Pattern Recogn Lett*, 56 (2015) 1.
11 Irani J, Pise N & Phatak M, *Int J Comput Appl*, 134 (2016) 9.
12 Deeksha & Sahu S, *Int J Res Appl Sci Eng Technol*, 5 (2017) 1711.
13 Iglesias F & Kastner W, *Energies*, 6 (2013) 579.
14 Bezdek J C, *J Cybern*, 3 (1974) 58.
15 Bezdek J C, Proc 8th Annu Int Conf Numer Taxon, 143 (1975) 166.
16 Rousseeuw P J, *J Comput Appl Math*, 20 (1987) 53.
17 Douzal-Chouakria A, Vilar J A & Marteau P F, Proc 1st ECML PKDD Workshop, (2016).
18 Horta D, de Andrade I C & Campello R J G B, *Theor Comput Sci*, 412 (2011) 5854.
19 Xie X & Beni G, *IEEE Trans Pattern Anal Mach Intell*, 13 (1991) 841.
20 Singh M, Bhattacharjee R, Sharma N & Verma A, Proc 4th Int Conf Image Inform Process (ICIIP), (2017) 95.
21 Bensaid A M, Hall L O, Bezdek J C, Clarke L P, Silbiger M L, Arrington J A & Murtagh R F, *IEEE Trans Fuzzy Syst*, 4 (1996) 185.
22 Pearson K, Proc Roy Soc Lond, 58 (1895) 240.
23 Wright S, *J Agric Res*, 20 (1921) 557.
24 Gueorguieva N, Valova I & Georgiev G, *Procedia Comput Sci*, 114 (2017) 224.
25 Hui E G M, Learn R for Applied Statistics: with Data Visualizations, Regressions, and Statistics (Apress, Singapore), 2019.
26 Saputra A, Sofyan H & Putra T E, *Lect Notes Mech Eng*, (2021) 401.