International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 5
ISSN: 2088-5334

Soft Set Parametric-based Data Clustering for Building Data Set

Iwan Tri Riyadi Yanto [a,f,*], Ahmad Azhari [b], Rohmat Saedudin [c], Sely Novita Sari [d], Mustafa Mat Deris [c,e], Norhalina Senan [f]

[a] Department of Information Systems, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[b] Department of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[c] Department of Information Systems, Telkom University, Bandung, West Java, Indonesia
[d] Faculty of Civil Engineering and Planning, Institute Teknologi Nasional Yogyakarta, Indonesia
[e] Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia
[f] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia
Corresponding author: *yanto.itr@is.uad.ac.id

***Abstract*—Identifying buildings for safety purposes is critical to anticipate unforeseen scenarios during a disaster. Rapid Visual Screening (RVS) is one of the procedures that can be used to determine a building's hazardous structure. The growing number of buildings necessitates grouping to provide recommendations for improving the analysis or conducting a more extensive review of the same building group. This article investigates the application of fuzzy clustering to the RVS dataset. Numerous strategies are compared, including fuzzy centroid clustering, fuzzy K-partition clustering, and multi-soft set clustering. The technique is applied to the RVS data set from Kulon Progo, Yogyakarta, which has 144 cases for grouping construction. Four clusters are formed from four distinct variables with fewer conditions: Plan Drawing, Floor Plan, Connection, and Stance. The experiment is based on the rank index, the Dunn index, and response time. The results indicate that multi-soft set-based clustering outperforms other baseline approaches. The investigator or government can utilize this information to suggest treating each cluster's "less" variable.**

***Keywords*— Clustering; soft set; multinomial distribution; RVS.**

Manuscript received 11 Aug. 2021; revised 15 Oct. 2021; accepted 10 Jan. 2022. Date of publication 31 Oct. 2022.

I. INTRODUCTION

Building development is increasing in breadth, not just in urban areas but also in rural areas. Conversion of the environment is necessary to transform the area by introducing safe and energy-efficient structures [1]. Identifying a structure is critical in determining whether it is safe or requires repair or reconstruction. Rapid Visual Screening is one method for estimating the seismic vulnerability of many structures in a city (RVS) [2], [3]. It is based on correlations between the predicted seismic performance of the buildings and their structural typology (frame, shear wall, monolith, in-fill), material composition (steel, reinforced concrete, reinforced/unreinforced masonry, wood, composite), design methods, and other details. The RVS approach was created as a screening tool for identifying constructions that may be dangerous [4]. RVS enables users to classify surveyed structures into two categories: those that pose no concern to life safety and those that may be seismically hazardous and should be further analyzed by a design specialist. Comprehensive seismic vulnerability assessment is a technically demanding method that can be done on a limited number of structures [5]. As a result, it is vital to adopt simpler processes that enable rapid assessment of the vulnerability profile of various buildings, allowing for more sophisticated evaluation procedures to be reserved for the most critical structures [6].

Numerous decision-making algorithms based on data mining have been applied to the RVS to classify the damage index of reinforced concrete (RC) buildings [7]–[9]. Another strategy is classifying buildings using a condition index scale [10]. Clustering is employed in [11] to monitor the thermal status of the building under a variety of external situations. Using the RVS data set, this article uses the clustering technique to divide the building into multiple categories based on shared traits or situations. It is crucial to distinguish the process of homogeneity formation. Clustering is a data mining technique that allows vast amounts of data to be divided into smaller groupings [12],[13]. Numerous clustering approaches have been proposed. Xu et al. et al. [14]

proposed fuzzy k-modes. It is based on the matching dissimilarity metric. Due to the potential for artifacts associated with the usage of hard centroids, Kim et al. [15] increased the performance of fuzzy k-modes by replacing fuzzy centroids (FC) with hard centroids. It is a non-parametric technique based on the principle of minimizing the sum of squared errors within clusters. Miin-Shen et al. [16] introduced the Fuzzy k-partitioning (FkP) algorithm, a parametric approach based on the likelihood function of multivariate multinomial distributions.

Additionally, the FkP technique for categorical data can be considered a fuzzy-based clustering algorithm. On the other hand, almost all fuzzy categorical data clustering techniques previously described represent data sets as binary values. On the other hand, categorical data have multi-valued attribute that can be represented as a multi-soft [17], [18]. The multi-soft set used for multi-valued attribute has advantages in representing the categorical data without the need to be converted into binary values. Based on this advantages, Yanto *et al.* proposed propose a clustering technique based on soft set theory for categorical data via multinomial distribution called MDD [19].

Not all strategies described above have been studied to determine RVS clustering's performance. Thus, we undertake an experiment to determine the feasibility of grouping the RVS dataset using a fuzzy parametric model.

II. Materials and Method

A. Rapid Visual Screening (RVS)

Rapid Visual Screening (RVS) is a technique created by FEMA for quickly identifying inventories that may be seismically hazardous. Rapid visual screening (RVS) is a technique for assessing a building's sensitivity to earthquake risks based on visual inspections from the outside and, if necessary, from within the structure. It is relatively straightforward to implement. Rapid Visual Screening (RVS) is a new method of visually inspecting buildings introduced in the United States. It uses a set of fields that provide primary data about the structures analyzed, such as the number of floors, construction years, building addresses, building pictures, and building sketches representing the building's floor plan and elevation [20]. Rapid Visual Screening (RVS) is a visual examination technique used in Guwahati [20], Nepal [21], and a hospital [22]. Rapid Visual Screening (RVS) is one technique for lowering the vulnerability and condition of soil and structure to natural disasters, most notably earthquakes. The completion of the RVS form collects RVS data. FEMA's fundamental building assessment (standard wall) (Federal Emergency Management Agency). After completion of the RVS form, each building's final score is determined in line with FEMA 154-2002 [23][24].

B. Data Collection

The data is primary data collected at Kulon Progo, Yogyakarta. The field survey is performed by directly looking at existing buildings and then adapting them into a simple building valuation method [25]. A basic building form involves the parts of a building that a building must own to make the building structurally sound [26]. The variables are conducted from 11 parts comprising 40 components of the standard basic building. Thus, the survey consists of 40 observations with 3- 4 observations of each variable. The list of variables is given in Table 1.

Simply check the "Yes" column to see whether or not the building part fits and the "No" shape or column to determine whether or not the building part does not exist. If a part of the building shape fits but the size does not, the bias can be filled in the Less. The 144 structures were gathered from three Kulon Progo villages: Kalirejo, Sangon, and Kalikubo.

TABLE I
THE LIST OF VARIABLES

No	Variable
1	Plan Drawing
2	Floor plan
3	House Foundation
4	Sloof
5	Column
6	Wall
7	Ring Back
8	Reinforcement Details
9	Connection
10	Mountains
11	Stance

C. Analysis Technique

The data is analyzed using the clustering technique to determine which buildings are in a comparable state of repair. Several baseline techniques, including FC and FkP, are compared to the proposed multivariate multinomial distribution (MMD) technique based on several soft sets [19]. It uses MMD to determine the highest probability and multi-soft set decomposition to break the data into numerous sets with comparable values [27] [28]. It is defined as:

$$Max\ L_{CML}(z,\lambda)=\sum_{i=1}^{|U|}\sum_{k=1}^{K}z_{ik}\sum_{j=1}^{|A|}\sum_{l=1}^{|a_j|}\ln\left(\lambda_{kjl}^{u_i}\right)^{|F,a_{jl}|} \quad (1)$$

Subject to

$$\sum_{k=1}^{K}z_{ik}=1, for\ i=1,2,\ldots,|U|.$$

$$\sum_{l=1}^{|a_j|}\lambda_{kjl}=1.$$

The maximization of the objective function $L_{CML}(z,\lambda)$ can be obtained by updating the equation as follows:

$$\lambda_{kjl}=\frac{\sum_{u_i\in(F,a_{jl})}z_{ik}(u_i)}{|U|} \quad (3)$$

$$z_{ik}=\begin{cases}1 & if\quad \sum_{j=1}^{|A|}\ln\lambda_{kjl}^{u_i}=\max_{1\le k'\le K}\sum_{j=1}^{|A|}\ln\lambda_{k'jl}^{u_i}\\ 0 & otherwise\end{cases} \quad (4)$$

where $U=\{u_1,u_2,\ldots,U_n\}$ is a finite set of instances, $A=\{a_1,a_2,\ldots a_m\}$ is a finite set of attributes. $(F,E)=\left((F,a_1),(F,a_2),\cdots,(F,a_{|A|})\right)$ can be defined as a multi-soft set over universe U as in [29], where $(F,a_1),\cdots,(F,a_{|A|})\subseteq(F,A)$ and $\left(F,a_{j_1}\right),\cdots,\left(F,a_{j_{|a_j|}}\right)\subseteq(F,a_j)$.

III. Results and Discussion

A. External Validity

The rank index is used to validate the performance of the strategies externally. External validity demands the computation of the rank index using external classes and comparing it to the cluster formed by the procedures. The data will be divided into three categories for this purpose based on a simple percentage of building damage determined through an examination of existing forms, namely secure percentage > 70%, less secure percentage 40-69 percent, unsafe percentage 40%, and unsafe percentage 40%, as shown in Table 2 [30]. Calculate the percentage value by multiplying the response 'Yes' by 1.0, the response 'Less' by 0.5, and the response 'No' by 0. The sum of all data points is divided by forty (the simple number of building components) and multiplied by one hundred percent to obtain the proportion of basic buildings using the simple building evaluation technique.

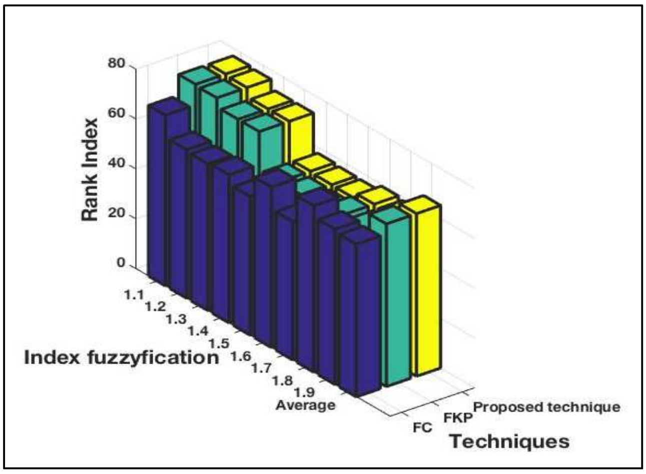

Fig. 1 The Rank indexes

TABLE II
CONDITION INDEX SCALE.

Zone	Condition Index	Condition Description	Handling Measure	Building Categorization
1	70-100	Well	No immediate action is required.	Secure
2	40-69	Intermediate	To determine the appropriate course of action, it is necessary to conduct an alternative economic analysis of improvements	Unsafe
3	0-39	Bad	A thorough evaluation is required to determine the necessary repair, rehabilitation, and reconstruction actions and assess the safety.	Not Safe

The experiment is repeated for each technique on a PC equipped with an Intel i5-8400 six-core processor running at 2.8 GHz and 8 GB RAM and the MATLAB programming language. Averages are used to calculate the rank index and time response. The first graph illustrates performance in terms of the total average. Increase the index fuzziness of each approach by 1.1–1.9. The FkP and proposed technique outperform the FC with an almost identical overall average of rank index. Additionally, Table 3 demonstrates that the suggested strategy outperforms baseline techniques in terms of time response, with an improvement of up to 98 percent.

TABLE III
TIME RESPONSES

	FC	FKP	Proposed	Improvement
Time Response (second)	8.3592	5.8948	0.1105	98.13 %

B. Internal validity based on the Number of Clusters

This section describes the performance of the three techniques to know the stability in terms of the number of clusters created concerning the increasing number of clusters. Whether the techniques will follow the number of clusters setting or can limit the number of clusters themselves. We define the technique called divergent if it creates a cluster following the number of clusters given; convergent to 1 means that the cluster members are only one. Since the data collection is obtained from 144 buildings, the number of clusters is set up to 2-100 (< 144). Figure 2 shows that the FC technique creates a number of clusters under the number of clusters given.

Meanwhile, the FkP technique convergent to 1 after the number of clusters given is more than 45. The proposed technique has good stability with convergent into 50-60 number of the cluster concerning increasing the number of clusters given. Then, the Dunn index is performed to determine the quality of the cluster in itself and concerning the increasing number of clusters. It can be seen that the technique has the lowest Dunn index when the number of clusters increases up to 25. For more than 25 clusters setting, the technique can keep the number of clusters created and obtain the Dunn index value. It can be seen in Figure 3.

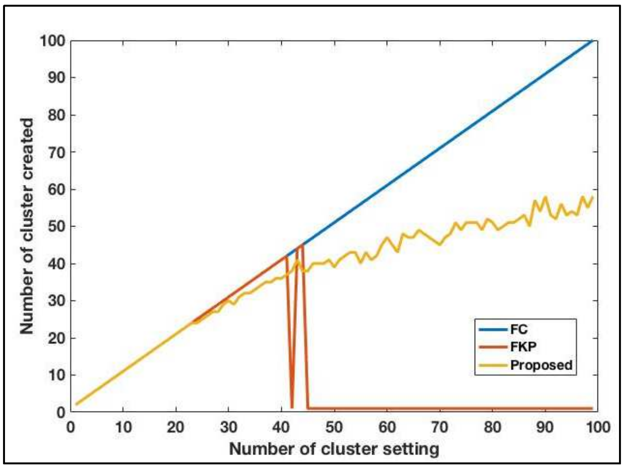

Fig. 2 The cluster created

C. Implementation on Dataset

Based on the rank index values, the technique performs well in index fuzziness 1.1 and 1.2. We select 1.2 as index fuzziness to implement on the dataset. Then Figure 4 illustrates the Dunn index of the proposed technique concerning the increasing number of clusters. Figure 5 is a subfigure on the Dunn index in the range of 2-10 clusters. The best number of clusters is 2 or 3 on the first level and four on the second level because it has a higher Dunn index. Thus, the data is clustered into 3 clusters using the proposed technique and is explored for four numbers or clusters.

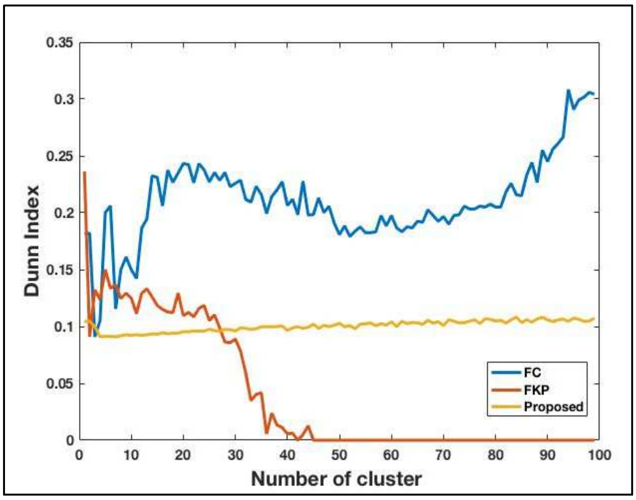

Fig. 3 The Dunn Index

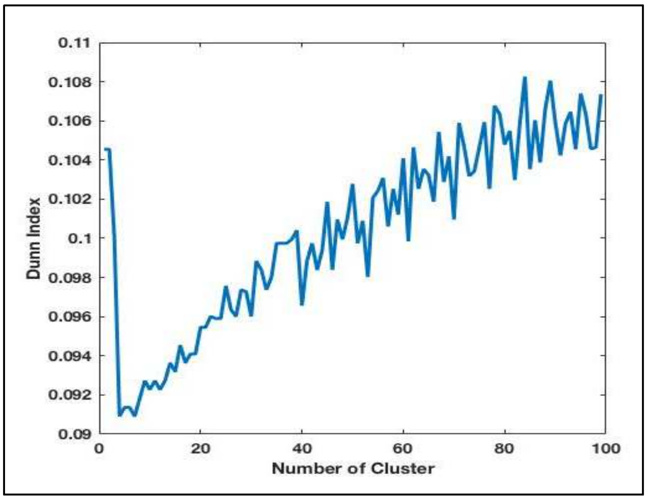

Fig. 4 The Dunn index of the data using proposed approach

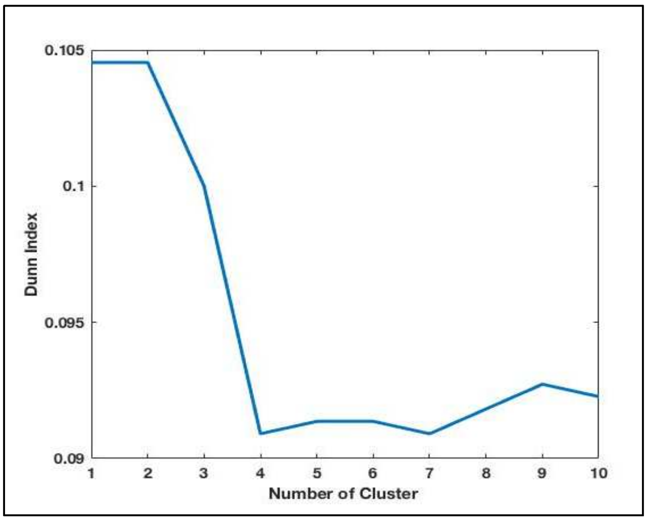

Fig. 5 The Dunn index in range 1-10 number of clusters

Table 4. shows the clustering results with the distribution number of the building in each area. The building is clustered into 3 clusters with the condition based on the average index scale of all members. It also summarizes the number of members of each area. Meanwhile, the data are clustered into 4 clusters, as shown in Table 5. It is interesting if it is compared with the condition index scale as in Table 1, where cluster C3 is on the zone unsaved but almost close to the secure zone. It may be suggested to the investigator to determine different recommendations between clusters C2 and C3. Then, the clustered buildings are identified, which variable is significant that makes C2 and C3 separated. The mean data variable of each cluster is classified by a threshold value of 0.6, where the variable is less if the mean value < 0.6. Otherwise, the variable is ok. The result is summarized in Table 6. This shows that the first four variables with fewer conditions, i.e., Plan Drawing, Floor plan, Connection, and Stance, are obtained.

TABLE IV
THE CLUSTERING RESULTS OF RVS DATASET WITH 3 CLUSTERS

Clusters	Index Scale	Category	Number of Members			
			Total	Kalikubo	Kalirejo	Sangon
C1	21.1697	Not Safe	18	12	3	3
C2	64.2456	unsafe	76	4	34	38
C3	72.7715	Secure	50	35	5	10

TABLE V
THE CLUSTERING RESULTS OF RVS DATASET WITH 4 CLUSTERS

Clusters	Index Scale	Category	Number of members			
			Total	Kalikubo	Kalirejo	Sangon
C1	21.1697	Not Safe	18	12	3	3
C2	59.3395	unsafe	46	0	11	35
C3	69.1518	unsafe (practically secure)	30	4	23	3
C4	72.7715	Secure	50	35	5	10

TABLE VI
THE CONDITION OF EACH CLUSTER

No	Variable	Condition			
		C1	C2	C3	C4
1	Plan Drawing	less	less	less	ok
2	Floor plan	less	ok	less	ok
3	House Foundation	less	ok	ok	ok
4	Sloof	less	ok	ok	ok
5	Column	less	ok	ok	ok
6	Wall	less	ok	ok	ok
7	Ring Back	less	ok	ok	ok
8	Reinforcement Details	less	ok	ok	ok
9	Connection	less	less	ok	less
10	Mountains	less	ok	ok	ok
11	Stance	ok	less	less	less

IV. CONCLUSION

Several techniques, namely Fuzzy centroid, Fuzzy K-partition, and multi-soft set-based clustering have been explored and implemented in group building using the RVS dataset. The experiment shows that the multi-soft set-based clustering achieves the best performance in terms of Rank index, Dunn index, and response time compared to baseline techniques. From the proposed technique, 4 clusters based on the first four variables with fewer conditions, i.e., Plan Drawing, Floor plan, Connection, Stance, are obtained. The four clusters are C1 (not safe condition), which contains 18 buildings, and C2 (unsafe), which contains 46 buildings. C3 (unsafe / practically secure) contains 30 buildings, and C4 (Secure) contains 50 buildings. The investigator or government can use this to provide recommendations to determine different treatments for the "less" variable in each cluster.

REFERENCES

[1] A. Barbaresi, M. Bovo, and D. Torreggiani, "The dual influence of the envelope on the thermal performance of conditioned and unconditioned buildings," *Sustain. Cities Soc.*, vol. 61, p. 102298, 2020.

[2] E. Harirchian *et al.*, "A review on application of soft computing techniques for the rapid visual safety evaluation and damage classification of existing buildings," *J. Build. Eng.*, vol. 43, p. 102536, 2021.

[3] S. Ishack, S. P. Bhattacharya, and D. Maity, "Rapid Visual Screening method for vertically irregular buildings based on Seismic Vulnerability Indicator," *Int. J. Disaster Risk Reduct.*, vol. 54, p. 102037, 2021.

[4] E. Harirchian, K. Jadhav, K. Mohammad, S. E. A. Hosseini, and T. Lahmer, "A comparative study of MCDM methods integrated with rapid visual seismic vulnerability assessment of existing RC structures," *Appl. Sci.*, vol. 10, no. 18, 2020.

[5] M. M. Kassem, F. Mohamed Nazri, and E. Noroozinejad Farsangi, "The seismic vulnerability assessment methodologies: A state-of-the-art review," *Ain Shams Eng. J.*, vol. 11, no. 4, pp. 849–864, 2020.

[6] A. Darko, A. P. C. Chan, Y. Yang, and M. O. Tetteh, "Building information modeling (BIM)-based modular integrated construction risk management – Critical survey and future needs," *Comput. Ind.*, vol. 123, p. 103327, 2020.

[7] P. Zhou and Y. Chang, "Automated classification of building structures for urban built environment identification using machine learning," *J. Build. Eng.*, vol. 43, p. 103008, 2021.

[8] H. Du *et al.*, "A classification method of building structures based on multi-feature fusion of UAV remote sensing images," *Earthq. Res. Adv.*, vol. 1, no. 4, p. 100069, 2021.

[9] Z. Ye, K. Cheng, S.-C. Hsu, H.-H. Wei, and C. M. Cheung, "Identifying critical building-oriented features in city-block-level building energy consumption: A data-driven machine learning approach," *Appl. Energy*, vol. 301, p. 117453, 2021.

[10] W. Smith, "The role of environment clubs in promoting ecocentrism in secondary schools: student identity and relationship to the earth," *J. Environ. Educ.*, vol. 50, no. 1, pp. 52–71, Jan. 2019.

[11] N. Khan, M. Ahmed, and N. Roy, "Temporal Clustering Based Thermal Condition Monitoring in Building," *Sustain. Comput. Informatics Syst.*, p. 100441, Sep. 2020.

[12] G. M. Gonçalves and L. L. Lourenço, "Mathematical formulations for the K clusters with fixed cardinality problem," *Comput. Ind. Eng.*, vol. 135, pp. 593–600, 2019.

[13] G. J. McLachlan, S. I. Rathnayake, and S. X. Lee, "2.24 - Model-Based Clustering☆," S. Brown, R. Tauler, and B. B. T.-C. C. (Second E. Walczak, Eds. Oxford: Elsevier, 2020, pp. 509–529.

[14] M. K. N. Huang, "A fuzzy k-modes algorithm for clustering categorical data - Fuzzy Systems, IEEE Transactions on," *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 4, pp. 446–452, 1999.

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

[16] M.-S. Yang, Y.-H. Chiang, C.-C. Chen, and C.-Y. Lai, "A fuzzy k-partitions model for categorical data and its comparison to the GoM model," *Fuzzy Sets Syst.*, vol. 159, no. 4, pp. 390–405, 2008.

[17] M. S. Khan, G. Mujtaba, M. A. Al-garadi, N. H. Friday, A. Waqas, and F. R. Qasmi, "Multi-soft sets-based decision making using rank and fix valued attributes," in *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2018, pp. 1–11.

[18] B. Pardasani, "Multi Softset for Decision Making," *Int. J. Sci. Res.*, vol. 7, no. 11, pp. 55–56, 2018.

[19] I. Tri, R. Yanto, R. Saedudin, S. Novita, M. Mat, and N. Senan, "Soft Set Multivariate Distribution for Categorical Data Clustering," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 11, no. 5, pp. 1841–1846, 2021.

[20] T. Sarmah and S. Das, "Earthquake Vulnerability Assessment for RCC Buildings of Guwahati City using Rapid Visual Screening," in *Procedia Engineering*, 2018, vol. 212, pp. 214–221.

[21] M. M. Akhlaghi, S. Bose, M. E. Mohammadi, B. Moaveni, A. Stavridis, and R. L. Wood, "Post-earthquake damage identification of an RC school building in Nepal using ambient vibration and point cloud data," *Eng. Struct.*, vol. 227, p. 111413, Jan. 2021.

[22] D. Perrone, M. A. Aiello, M. Pecce, and F. Rossi, "Rapid visual screening for seismic evaluation of RC hospital buildings," *Structures*, vol. 3, pp. 57–70, Aug. 2015.

[23] B. Lizundia *et al.*, "Rapid visual screening of buildings for potential seismic hazards: FEMA 154 and FEMA 155 updates," *NCEE 2014 - 10th U.S. Natl. Conf. Earthq. Eng. Front. Earthq. Eng.*, no. January, 2014.

[24] S. Sari, S. N. Sari, R. Prastowo, R. Junaidi, and A. Machmud, "Rapid Visual Screening of Building for Potential Ground Movement in Kalirejo, Kulonprogo, Yogyakarta," *J. Ilm. Pendidik. Fis. Al-Biruni*, vol. 9, no. 1, pp. 51–59, Apr. 2020.

[25] A. K. Mallick and A. Mukhopadhyay, "Different Schemes for Improving Fuzzy Clustering Through Supervised Learning BT - Computational Intelligence, Communications, and Business Analytics," 2019, pp. 155–164.

[26] J. Arora, K. Khatter, and M. Tushir, "Fuzzy c-Means Clustering Strategies: A Review of Distance Measures BT - Software Engineering," 2019, pp. 153–162.

[27] I. T. R. Yanto, R. Setiyowati, M. M. Deris, and N. Senan, "Fast Hard Clustering Based on Soft Set Multinomial Distribution Function BT - Recent Advances in Soft Computing and Data Mining," 2022, pp. 3–13.

[28] I. T. R. Yanto, M. M. Deris, and N. Senan, "PSS: New Parametric Based Clustering for Data Category," in *Recent Advances in Soft Computing and Data Mining*, 2022, pp. 14–24.

[29] T. Herawan, M. M. Deris, and J. H. Abawajy, "Matrices Representation of Multi Soft-Sets and Its Application BT - Computational Science and Its Applications – ICCSA 2010," 2010, pp. 201–214.

[30] J. Nakajima and A. Hasegawa, "Subduction of the Philippine Sea plate beneath southwestern Japan: Slab geometry and its relationship to arc magmatism," *J. Geophys. Res.*, vol. 112, no. B8, p. B08306, Aug. 2007.