Soft Set Parametric based Data Clustering for Building Data Set

By Iwan Tri Yanto

Soft Set Parametric based Data Clustering for Building Data Set

Iwan Tri Riyadi Yanto[1,6], Ahmad Azhari[2], Rohmat Saedudin[3], Sely Novita sari[4], Mustafa Mat Deris[3,5], Norhalina Senan[6]

[1] Department of Information Systems, Universitas Ahmad Dahlan,
[2]Department of Informatics, Universitas Ahmad Dahlan,
Yogyakarta, Indonesia
email: {yanto.itr@is.uad.ac.id, ahmad.azhari@tif.uad.ac.id}
[3]Department of Information Systems,
Telkom University,
email: rdrohmat@telkomuniversity.ac.id
[4]Faculty of Civil Engineering and Planning,
Institute Teknologi Nasional Yogyakarta
email: sely.novita @itny.ac.id
[5]Faculty of Applied Science and Technology,
[6]Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor
email: {mmustafa,halina}@uthm.edu.my

Abstract— **Identification of buildings for safety purposes is critical in order to anticipate unforeseen scenarios in the event of 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 in order 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, namely 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. This information can be utilized by the investigator or the government to make suggestions on how to treat the "less" variable in each cluster.**

Keywords— **Clustering; Soft Set; Multinomial Distribution; RVS.**

I. Introduction

Building development is increasing in breadth, not just in urban areas, but also in rural areas. Conversion of the environment is necessary in order to transform the area through the introduction of safe and energy-efficient structures [1]. The identification of 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 a large number of structures in a city (RVS). 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 [2]. 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 [3]. As a result, it is vital to adopt simpler processes that enable rapid assessment of the vulnerability profile of various types of buildings, allowing for more sophisticated evaluation procedures to be reserved for the most critical structures [4].

Numerous decision-making algorithms based on data mining have been applied to the RVS to classify the damage index of reinforced concrete (RC) buildings [2]. Another strategy is to classify buildings using a condition index scale [5]. Clustering is employed in [6] to monitor the thermal status of the building under a variety of external situations. Using the RVS data set, this article use 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. Numerous clustering approaches have been proposed. Xu, et al. et al. [7] proposed the 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. [8] increased the performance of fuzzy k-modes by replacing fuzzy centroids for 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. [9] 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 thought of as a fuzzy-based clustering algorithm. On the other hand, almost all fuzzy categorical data clustering techniques previously described represent data sets as binary values. Yanto et al.[9]propose changing FkP by employing rough set theory's indecipherability relation. Not all of the 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. Material 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 that was introduced in the United States. It makes use of 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 [10]. Rapid Visual Screening (RVS) is a visual examination technique used in Guwahati [11], Nepal [12], and a hospital [13]. Rapid Visual Screening (RVS) is one technique for lowering the vulnerability and condition of soil and structure to natural disasters, most notably earthquakes. RVS data is collected by the completion of the RVS form. FEMA's fundamental building assessment (standard wall) (Federal Emergency Management Agency). Following completion of the RVS form, each building's final score is determined in line with the provisions of FEMA 154-2002 [14].

B. *Data collection*

The data is primary data collected at Kulon Progo, Yogyakarta. The field survey is performed by directly looking at existing buildings then adoption into a simple building valuation method [1]. A basic building form involves the parts of a building that a building must own to make the building structurally sound [2]. The variables are conducted from 11 parts consists of 40 components of the standard basic building. Thus, the survey consists 40 observations where there are 3-4 observations of each variables. The list of variable is given in table 1.

TABLE 1.

THE LIST OF VARIABLE

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

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 [19]. The 144 structures were gathered from three Kulon Progo villages: Kalirejo, Sangon, and Kalikubo.

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. It use MMD to determine the highest probability and multi soft set decomposition to break the data down into numerous sets with comparable values. It is defineable as

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

Subject to

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

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

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

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

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

where $U = \{u_1, u_2, \dots, U_n\}$ is finite set of instance, $A = \{a_1, a_2, \dots a_m\}$ is finite set of attribute $(F, E) = \left((F, a_1), (F, a_2), \cdots, (F, a_{|A|})\right)$ can be defined as a multi soft set over universe U as in [3], 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 externally validate the performance of the strategies. 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 [21][5].

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.

The experiment is repeated twenty times 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.

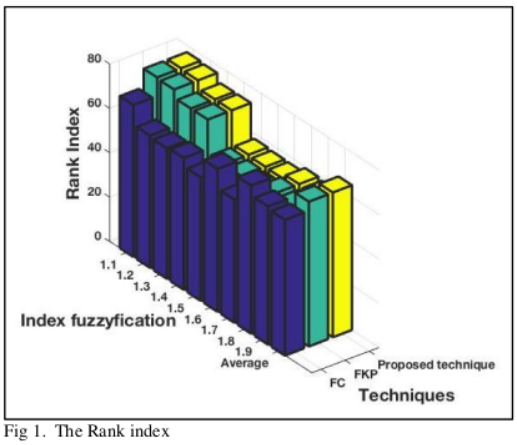

Fig 1. The Rank index

TABLE 2.
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, as well as to assess the safety.	Not Safe

TABLE 3 .
TIME RESPONSES

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

B. Internal validity based on number of cluster

This section describes the performance of the three techniques to know the stability in term of number of cluster created respect to 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 that the technique called divergent if it creates cluster follows the number of cluster given, convergent to 1 means that the number of member of cluster is only one. Since the data collection is obtained 144 building then the number of cluster is set up to 2-100 (< 144). Figure 2, shows that the FC technique creates number of cluster in accordance with the number of cluster given. Meanwhile, the FkP technique convergent to 1 after number of cluster given is more than 45. The proposed technique has good stability with convergent into 50-60 number of cluster with respect to increasing the number of cluster given. Then, the Dunn index is performed to determine the quality of cluster both of itself and with respect to increasing number of cluster. It can be seen that the technique has lowest Dunn index when the number of clusters is increasing up to 25. For more than 25 number of clusters setting, the technique able to keep the number of clusters created and to obtain Dunn index value. It can be seen in Figure 3.

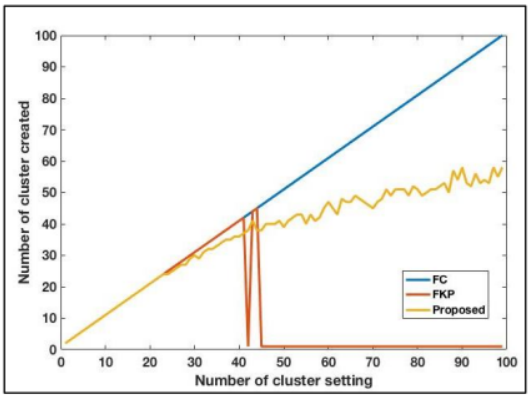

Fig 2. The cluster created

C. Implementation on dataset

Based on the rank index values, the technique has good performance in the 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 with respect to increasing number of clusters. Figure 5, is subfigure on the Dunn index in the range of 2- 10 number of clusters. It can be seen that the best number of clusters is 2 or 3 on the first level and 4 on second level, because it has higher Dunn index. Thus, the data is clustered into 3 clusters using the proposed technique and also it is explored for 4 number or clusters.

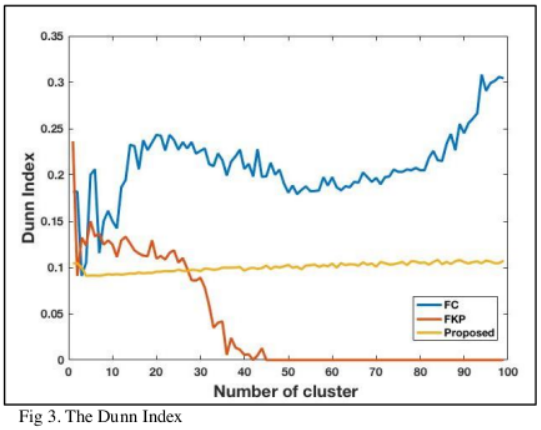

Fig 3. The Dunn Index

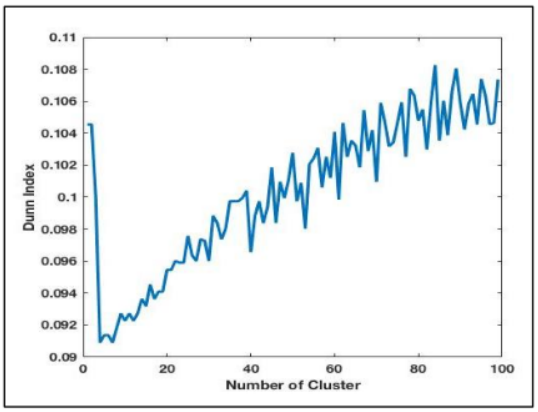

Fig 4. The Dunn index of the data using proposed approach

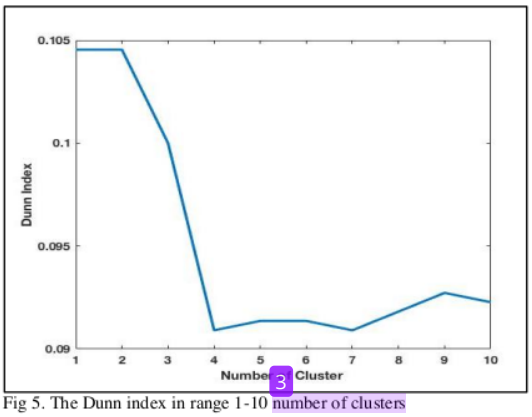

Fig 5. The Dunn index in range 1-10 number of clusters

Table 4. shows the clustering results with the distribution number of building of each areas. The building are clustered into 3 cluster with the condition based on the average index scale of all member. It also summarizes the number of member of each areas. Meanwhile, The data are clustered into 4 cluster as shown in Table 5. It is interesting if it compared with condition index scale as in table 1 where the cluster C3 is on zone unsaved but almost closed to secure zone. It may be suggested to investigator to determine different recommendation between cluster C2 and cluster C3. Then,

the clustered buildings are identified where which variable is significant that make C2 and C3 separated. The mean of data variable of each cluster is classify by threshold value 0.6, where the variable is less if the mean values < 0.6, otherwise the variable is ok. The result is summarized in Table 6. This shows the first four variables with less condition i.e Plan Drawing, Floor plan, Connection, Stance are obtained.

TABLE 4.

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 5.

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

IV. CONCLUSION

Several techniques namely Fuzzy centroid, Fuzzy K-partition and multi soft set based clustering has been explored and implemented to grouping building using RVS dataset. The experiment shows that the multi soft set based clustering achieves best performance in term of Rank index, Dunn index and response time, compared than baseline techniques. From the proposed technique, 4 clusters based on the first four variables with less condition i.e Plan Drawing, Floor plan, Connection, Stance are obtained. The four clusters are C1 (not safe condition) contains 18 building, C2 (unsafe) contains 46 buildings, C3 (unsafe / practically secure) contains 30 buildings and C4 (Secure) contains 50 buildings. This can be used by investigator or government 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, 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] T. Hadibarata and R. Rubiyatno, “Active learning strategies in environmental engineering course: A case study in curtin university Malaysia,” Jurnal Pendidikan IPA Indonesia, vol. 8, no. 4. pp. 456–463, 2019.
[18] A. Irsadi, S. Anggoro, T. R. Soeprobowati, M. Helmi, and A. S. E. Khair, “Shoreline and mangrove analysis along semarang-demak, Indonesia for sustainable environmental management,” Jurnal Pendidikan IPA Indonesia, vol. 8, no. 1. pp. 1–11, 2019.
[19] 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.
[20] T. Herawan, M. M. Deris, and J. H. Abawajy, “Matrices Representation of Multi Soft-Sets and Its Application,” in Computational Science and Its Applications -- ICCSA 2010: International Conference, Fukuoka, Japan, March 23-26, 2010, Proceedings, Part III, D. Taniar, O. Gervasi, B. Murgante, E. Pardede, and B. O. Apduhan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 201–214.
[21] 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.

Soft Set Parametric based Data Clustering for Building Data Set

ORIGINALITY REPORT

15%
SIMILARITY INDEX

PRIMARY SOURCES

1	ejournal.radenintan.ac.id Internet	102 words — 4%
2	eprints.uad.ac.id Internet	46 words — 2%
3	"Recent Advances on Soft Computing and Data Mining", Springer Science and Business Media LLC, 2017 Crossref	43 words — 2%
4	joiv.org Internet	26 words — 1%
5	Iwan Tri Riyadi Yanto, Younes Saadi, Dedy Hartama, Dewi Pramudi Ismi, Andri Pranolo. "A framework of fuzzy partition based on Artificial Bee Colony for categorical data clustering", 2016 2nd International Conference on Science in Information Technology (ICSITech), 2016 Crossref	15 words — 1%
6	Girish Chandra Joshi, Ratnesh Kumar. "Preliminary seismic vulnerability assessment of Mussoorie Town, Uttarakhand (India)", Journal of Building Appraisal, 2010 Crossref	14 words — 1%

7	Christopher Rojahn. "Christopher Rojahn", The Structural Design of Tall and Special Buildings, 09/2005 Crossref	13 words — < 1%
8	Rabiei Mamat, Tutut Herawan, Mustafa Mat Deris. "chapter 16 SAR", IGI Global, 2013 Crossref	12 words — < 1%
9	www.uajy.ac.id Internet	12 words — < 1%
10	www.fujipress.jp Internet	11 words — < 1%
11	Mohammadreza Yadollahi, Azlan Adnan, Rosli Mohamad Zin. "Seismic Vulnerability Functional Method for Rapid Visual Screening of Existing Buildings", Archives of Civil Engineering, 2012 Crossref	10 words — < 1%
12	Wahyu Riyanto, Djoko Irawan, Tri Joko Wahyu Adi, Data Iranata, Aniendhita Rizki Amalia. "Earthquake Vulnerability Assessment of High-Rise Buildings in Surabaya using RViSITS Android Application", IOP Conference Series: Materials Science and Engineering, 2020 Crossref	10 words — < 1%
13	www.mdpi.com Internet	10 words — < 1%
14	Thi Phuong Quyen Nguyen, R.J. Kuo. "Partition-and-merge based fuzzy genetic clustering algorithm for categorical data", Applied Soft Computing, 2019 Crossref	9 words — < 1%

15	Alessandra De Angelis, Marisa Pecce. "Seismic nonstructural vulnerability assessment in school buildings", Natural Hazards, 2015 Crossref	8 words — < 1%
16	Iwan Tri Riyadi Yanto, Edi Sutoyo, Ani Apriani, Okki Verdiansyah. "Fuzzy Soft Set for Rock Igneous Clasification", 2018 International Symposium on Advanced Intelligent Informatics (SAIN), 2018 Crossref	8 words — < 1%
17	Lecture Notes in Computer Science, 2010. Crossref	8 words — < 1%
18	Sami Naouali, Semeh Ben Salem, Zied Chtourou. "Clustering Categorical Data: A Survey", International Journal of Information Technology & Decision Making, 2020 Crossref	8 words — < 1%
19	eprints.utm.my Internet	8 words — < 1%
20	hdl.handle.net Internet	8 words — < 1%
21	Piyoosh Rautela, Girish Chandra Joshi, Bhupendra Bhaisora, Chanderkala Dhyani, Suman Ghildiyal, Ashish Rawat. "Seismic vulnerability of Nainital and Mussoorie, two major Lesser Himalayan tourist destinations of India", International Journal of Disaster Risk Reduction, 2015 Crossref	6 words — < 1%

EXCLUDE QUOTES OFF EXCLUDE MATCHES OFF

EXCLUDE BIBLIOGRAPHY ON