JOIV : Int. J. Inform. Visualization, 5(3) - September 2021 291-297

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Fast Clustering Environment Impact using Multi Soft Set Based on Multivariate Distribution

Iwan Tri Riyadi Yanto[a,e,*], Ani Apriani[b], Rahmat Hidayat[c] , Mustafa Mat Deris[d], Norhalina Senan[e]

[a] *Department of Information Systems, University of Ahmad Dahlan, Yogyakarta, Indonesia*
[b]*Faculty of Technology Mineral, Institut Teknologi Nasional Yogyakarta, Yogyakarta, Indonesia*
[c]*Department of Information Technology, Politeknik Negeri Padang, Padang, West Sumatera, Indonesia*
[d]*Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia*
[e]*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*— Every development activity is always related to human or community aspects. This can also lead to changes in the characteristics of the community. The community's increasing awareness and critical attitude need to be accommodated to avoid the emergence of social conflicts in the future. This research is to find out how the public perception about the impact of development on the environment. Two methods are used, i.e., MDA (Maximum Dependency Attribute) and MSMD (the Multi soft set multivariate distribution function). The MDA is to determine the most influential attribute and the Multi soft set multivariate distribution function (MSMD) is to group the selected data into classes with similar characteristics. This will help the police producer plan the right mediation and take quick activity to make strides in the quality of the social environment. The experiment conducted level of impact based on the clustering results with the greatest number of member clusters is cluster 1 (very low impact) with 32.25 % of total data following cluster 5 (Very High impact) with 24.25 % of total data. The experiment obtains the level of impact based on the clustering results. The greatest number of member clusters is cluster 1 (extremely low impact) with 32.25 % of total data following cluster 5 (Very High impact) with 24.25 % of total data. The scatter area impact is spread at districts 6, 7, 10, 11, the most of very high impact and districts 1,2,3,4,5,8 the lowest impact.**

***Keywords*—Environment impact MDA; clustering; multi soft set; multinomial distribution.**

Manuscript received 3 Jun. 2021; revised 9 Aug. 2021; accepted 17 Sep. 2021. Date of publication 30 Sep. 2021.

I. INTRODUCTION

Increasing the number of residents becomes a certainty in a country. This is a factor causing the growth of supporting facilities such as rapidly growing areas. Development is something that must be done to support people's lives. The goal of development is to improve living standards by utilizing accessible common and natural assets. One of the positive impacts of development is to provide convenience for the community. However, development also harms the surrounding community. The size of the impact is caused by the natural environment characteristics, participation in the development process, the accessibility of monetary and technological resources, and the type of government policy to be implemented [1]–[5]. Another influence that causes the impact of development that cannot be ignored is the spatial aspect. Urban areas, rural areas, suburban areas will certainly have differences in terms of the impact of development. This is due to different economic activities and the intensity of urban development [6], [7].

Yogyakarta is a city experiencing development activities. The city is well-known as Indonesia's primary destination for education and cultural exchange to achieve high levels of physical development. Investors in the trade and hospitality sectors are drawn to Yogyakarta because of its desirability. From an economic point of view, it has a positive impact on Yogyakarta. of Yogyakarta's Gross Domestic Product (GDP), the hospitality sector contributed 24.88%.

The positive impact of the development of the hotel sector in Yogyakarta cannot be separated from the negative impact. There are many complaints from the people of Yogyakarta against the rampant development of hotels and malls. The hallmark of Yogyakarta as a student city is that it is closed with many facilities that support fashion compared to learning facilities. Coupled with the negative impacts such as traffic

congestion, hot weather that causes temperature changes, decreased groundwater absorption, decreased turnover of small actors, and marginalization of local architecture [8]. Thus, it is necessary to research to determine the public's perception of the influence of development. This can create government or related agencies input to take policies on sustainable development [9], [10]. Detecting the biggest impact to the smallest impact using data mining is an effective way [11]. The mining of information is used to find and analyze modern data that will exist in information and summarize what comes about as valuable information. There are numerous exceptional ponders on information mining within the regions of clustering, affiliation rules, classification, strife examination, etc. [12]–[15]. The field of information mining that concerned with applications. To attain the inquire about objective, two strategies based on harsh set hypothesis are utilized: MDA and MSMD. The MDA strategy is utilized to positioning the reliance degrees of quality and select the foremost critical property or features [16]. Besides, the MSMD is a clustering method utilized to gather the information chosen for the lesson with comparative attributes [17]. Attributes selection, distinguishing the foremost powerful property, and gathering the information may the police maker to plan the right mediation and take quick activity to make strides the quality of the social environment.

II. Material and Methods

A. MDA

MDA is a rough set-based technique for ranking each attribute's dependency concerning the other attributes by Herawan *et al* [18]. It uses an attribute's reliance by selecting the attribute with the highest dependency among all others. The MDA technique has three steps: convert the data into the equivalence classes of an attribute, compute the degree of dependency attribute and select the highest one [18]–[20]. An equivalence relation can induce a unique partition. The indistinguishable relationship indicated IND is the starting point for structuring partitions $IND(B)$. Let $S(U,A,V,f)$ is an information system, $x,y \in U$ and attribute $B \subseteq A$ is $IND(B) \Leftrightarrow f(x,a) = f(y,a), \forall_a \in A$. The equivalence class U/B is the partition of U by $IND(B)$ and the U/B contains $x \in U$ is denoted $[x]_B$, Let $X \subseteq U$, The positive region$\underline{B}(X) = \{x \in X | [x]_B \subseteq X\}$ is the lower approximation. Then, the degree dependency D on C where $C, D \subseteq A, C \Longrightarrow_k D$ is defined as

$$k = \frac{Pos_C}{|U|}, 0 \le k \le 1,$$

$$Pos_C = \sum_{X \in \frac{U}{D}} \left|\underline{C}(X)\right|$$

B. The multi soft set multivariate distribution function

The data is analyzed using the clustering technique based on a multi soft set (MSMD). The technique uses a multinomial distribution function to find the highest probability and multi soft set based on decomposing the data into several sets with similar values.

Let $S = (U,A,V,f)$ be a categorical-valued information system, where $U = \{u_1, u_2, \dots, U_n\}$ is a finite set of instance, $A = \{a_1, a_2, \dots a_m\}$ is a finite set of the attribute, V is values set of each attribute A, f is mapping function $f: (U,A) \to V$ and $S = \left(U, a_i, V_{a_i}, f\right), i = 1,2,\cdots,|A|$ Boolean-valued information system, it can be decomposed to be multi-Boolean information system as

$$S = (U,A,V,f) = \begin{cases} S^1 = \left(U, a_1, V_{\{0,1\}}, f\right) \Leftrightarrow (F, a_1) \\ S^2 = \left(U, a_2, V_{\{0,1\}}, f\right) \Leftrightarrow (F, a_2) \\ \vdots \\ S^{|A|} = \left(U, a_{|A|}, V_{\{0,1\}}, f\right) \Leftrightarrow (F, a_{|A|}) \end{cases} = \left((F,a_1),(F,a_2),\cdots,\left(F,a_{|A|}\right)\right)$$

Then, $(F,E) = \left((F,a_1),(F,a_2),\cdots,\left(F,a_{|A|}\right)\right)$ is a multi soft set over universe U that represents a categorical-valued information system. $S = (U,A,V,f)$.

The multivariate multinomial distribution of multi soft set can be defined as is defined as:

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

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:

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

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

where $U = \{u_1, u_2, \dots, U_n\}$ is finite set of instances, $A = \{a_1, a_2, \dots a_m\}$ is finite set of attributes. $(F,E) = \left((F,a_1),(F,a_2),\cdots,\left(F,a_{|A|}\right)\right)$ can be defined as a multi soft set over universe U as in [21], where $(F,a_1),\cdots,\left(F,a_{|A|}\right) \subseteq (F,A)$ and $\left(F,a_{j_1}\right),\cdots,\left(F,a_{j_{|a_j|}}\right) \subseteq \left(F,a_j\right)$.

The MSMD algorithm is shown in algorithm 1.

Step 1. Fix $2 \le K \le |U|$ and fix $\varepsilon > 0$ and Max iter.

Give initials $z_{ik}^{(0)}$ and let $t = 1$.

Step 2. Compute $\lambda_{kjl}^{(it)}$ with $z_{ik}^{(it-1)}$

Step 3. Update $z_{ik}^{(it)}$ with $\lambda_{kjl}^{(it)}$

Step 4. Compare $z_{ik}^{(it)}$ to $z_{ik}^{(it-1)}$ in a convenient norm

IF $||z_{ik}^{(it)} - z_{ik}^{(it-1)}|| < \varepsilon$

or it = Maximum iteration set, THEN Stop

ELSE $it = it + 1$ and return to Step 2.

Algorithm 1. The MSMD algorithm

III. Results and Discussion

A. Analysis and Data Description

The poll collected data from 400 people, including 224 girls and 176 males from 14 different districts. The respondents react to the question on their perceptions of environmental impact in terms of physical and chemical aspects and economic aspects. In the Cronbach alpha test, the data has a reliability value of 0.953. Table 1 shows the data set's description.

TABLE I
Data Description

Data and Description	Attribute
Physic and chemical aspects Attributes: 10 Mean: 16 SD: 2.67	PC1: Quantity of water PC2: Quality of water PC3: Level of water absorption PC4: Temperature PC5: Level of air pollution PC6: Climate PC7: Level of noise PC 8: Land use PC 9: Public Facilities availability
Economic aspect Attributes: 9 Mean: 25 SD: 2.95	Eco 1: Immigration Eco 2: Employment Rate Eco 3: ESD Eco 4: Income Eco5: Expenditure Eco 6: Occupation shift Eco 7: open wellbeing Eco 8: Instructive office Eco 9: Devout office Eco 10: Wellbeing care facility

The goal of this study is to classify cities with similar environmental impacts. The procedure is divided into two stages. To begin, the most important attribute in the dataset was chosen by ranking the degree dependency using the MDA technique. Excluded attributes with a lower degree of reliance minimize the dataset's size. The smaller dataset is then clustered using the clustering technique. This can assist the police maker in designing the appropriate solution and take immediate action to improve the quality of the social environment.

B. Attribute Selection Using MDA

Nine qualities make up the physical and chemical aspects, with a mean of 25 and a standard deviation of 2.95. Table 2 shows the degree of dependency. Table 2 shows that the qualities PC3, PC4, PC6, PC7, PC8, and PC9 have the least amount of dependency. Thus, they are omitted from the future phases, leaving just PC1, PC2, and PC5. Furthermore, there is a heavy reliance on PC2, which is water quality as an important criterion.

The economic element has 10 features with a mean of 12 and a standard deviation of 1.59. Table 3 shows the degree of dependency. The highest features are Eco1 (immigration), which refers to the movement of people from outside the city of Yogyakarta, and Eco2 (Employee absorption). Because the qualities Eco3, Eco7, Eco8, and Eco10 have less reliance, they are eliminated from the next step, leaving only Eco1, Eco2, Eco4, Eco5, and Eco6. As a result, the selected qualities, namely PC1, PC2, PC5 in the physical and chemical aspects, and Eco1, Eco2, Eco4, Eco5, Ec6 in the economic aspect, have been grouped as a subset of the data. It is utilized as primary data to cluster the influence of the area using clustering techniques.

TABLE II
Degree Dependency

Physic and chemical aspects	Economic aspect	
Degree of dependency	**1stDegree of dependency**	**2ndDegree of dependency**
PC1: 0.0075	Eco1: 1	Eco1:0.01
PC2: 0.0475	Eco2: 1	Eco2:0.01
PC3: 0.0025	Eco3: 0	Eco3:0
PC4:0.0025	Eco4: 1	Eco4:0
PC5: 0.0075	Eco5:1	Eco5:0
PC6: 0.0025	Eco6:0.015	Eco6:0
PC7: 0.0025	Eco7: 0	Eco7:0
PC8: 0.0025	Eco8: 0	Eco8:0
PC9: 0.0025	Eco9:0	Eco9:0
	Eco10:0	Eco10:0

TABLE III
Response Time

Technique	Fuzzy Centroid	Fuzzy partition	K	MSMD
Average of time response (second)	8.9582	24.0682		0.0896

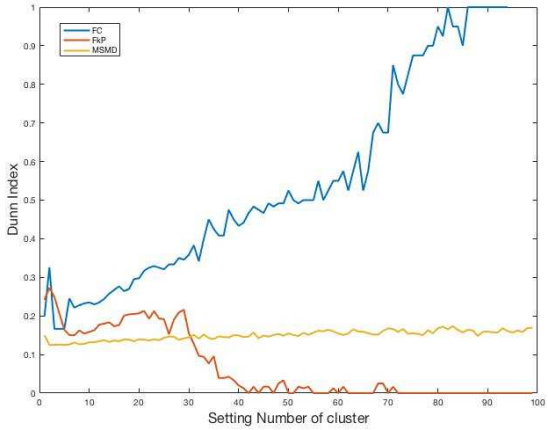


Fig. 1 The Dunn index respect to the set number of clusters (2-100)

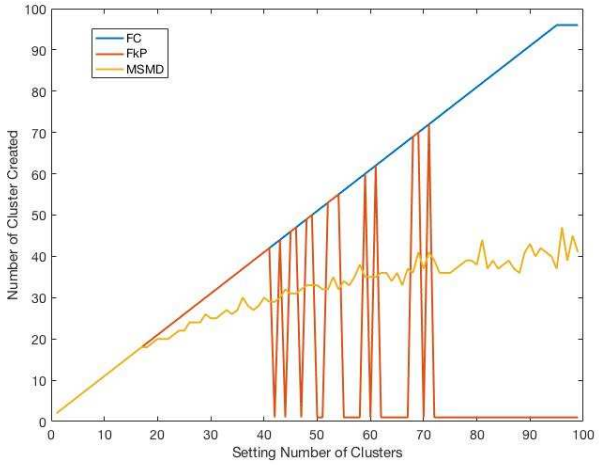
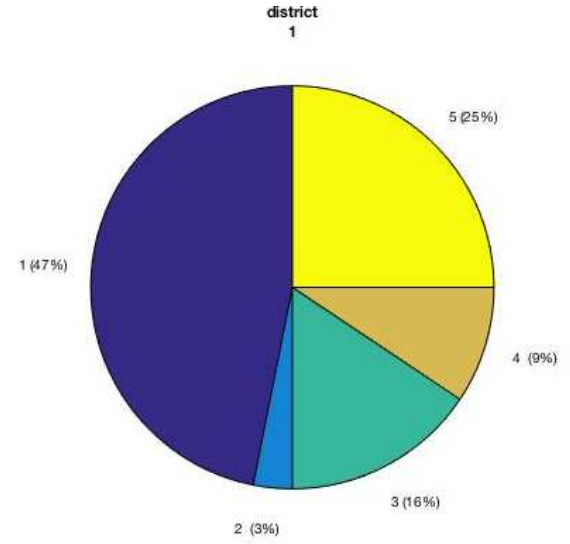
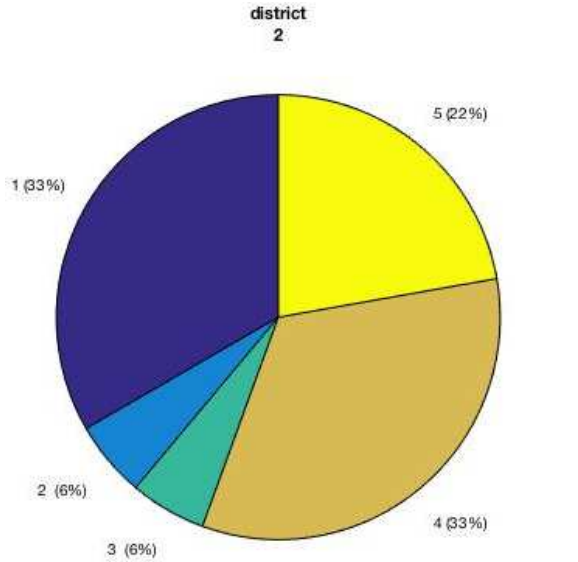
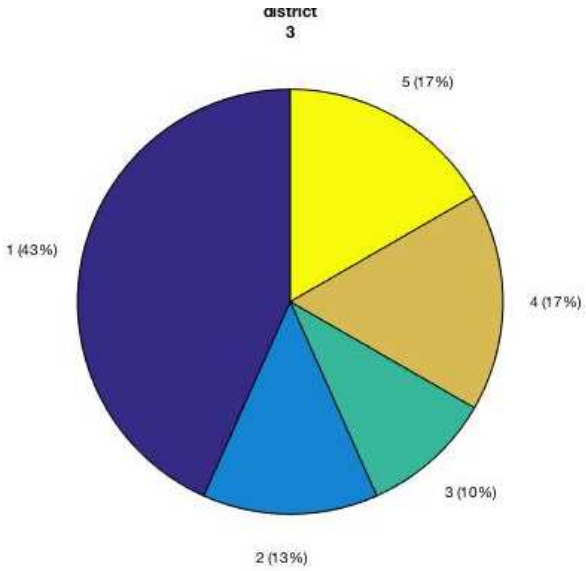
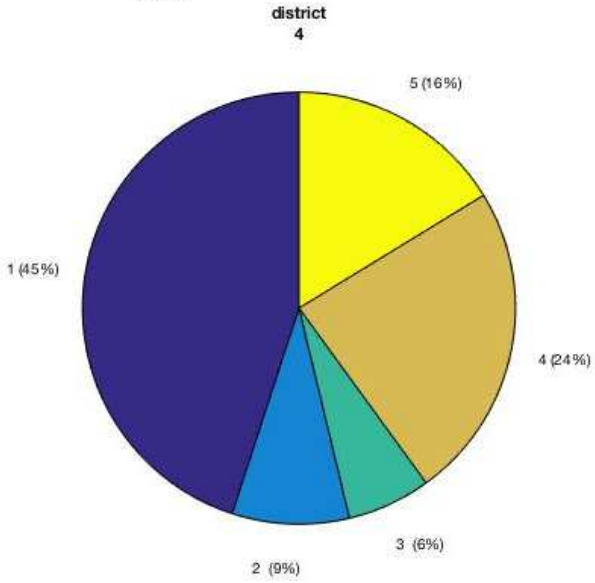


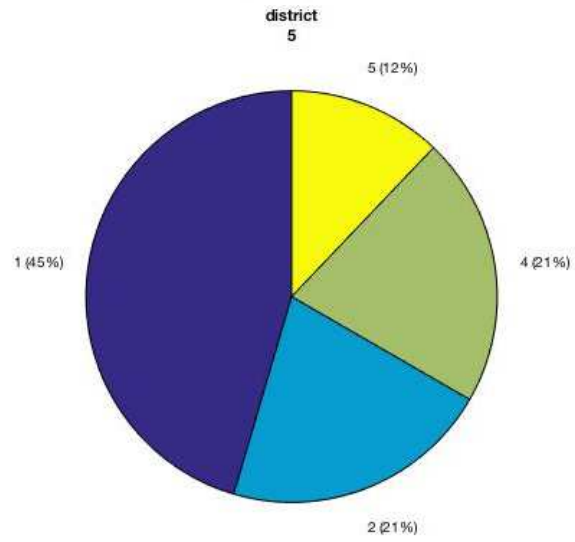
Fig. 2 The number of clusters created by techniques versus the setting number of clusters (2-100)

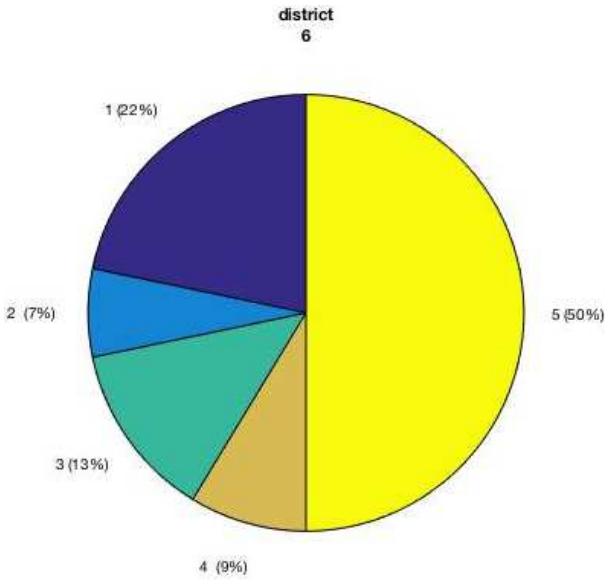

district
1
5 (25%)
1 (47%)
4 (9%)
3 (16%)
2 (3%)

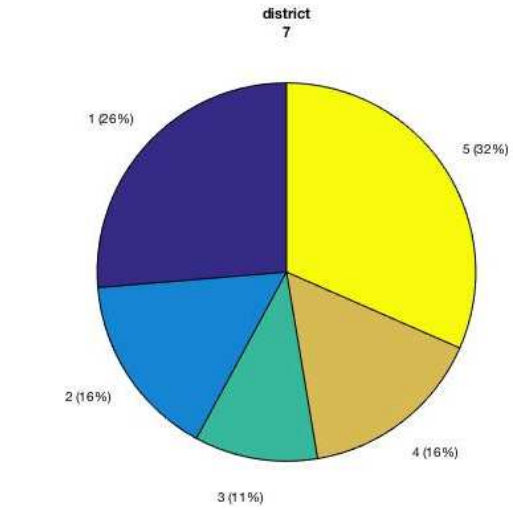

district
2
5 (22%)
1 (33%)
2 (6%)
3 (6%)
4 (33%)

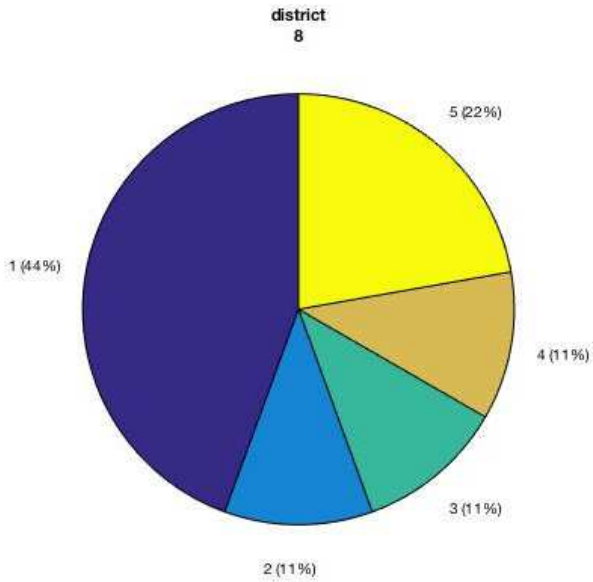

district
3
5 (17%)
1 (43%)
4 (17%)
3 (10%)
2 (13%)

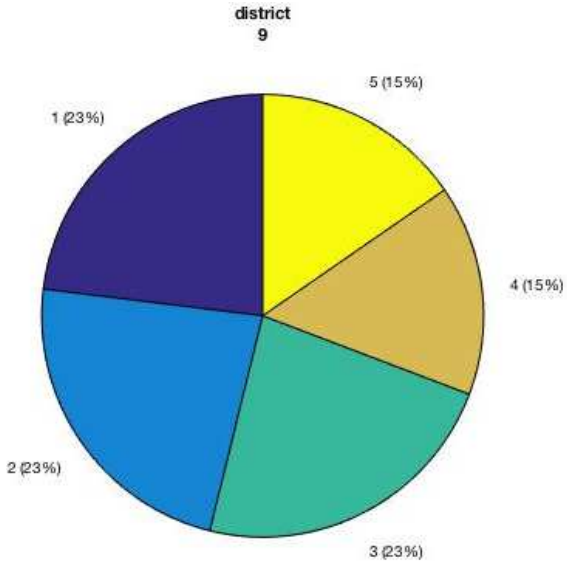

district
4
5 (16%)
1 (45%)
4 (24%)
3 (6%)
2 (9%)

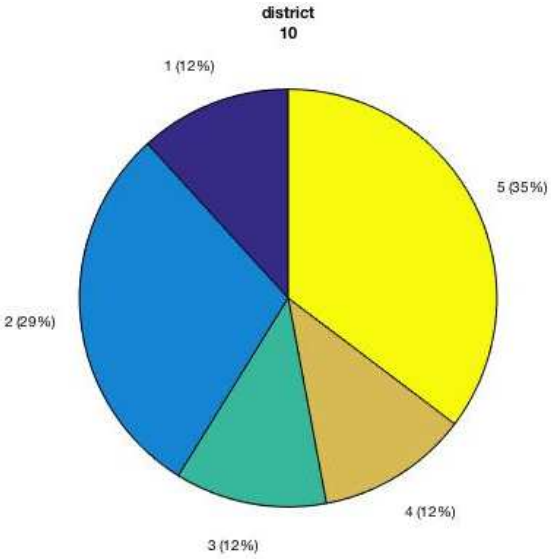

district
5
5 (12%)
1 (45%)
4 (21%)
2 (21%)

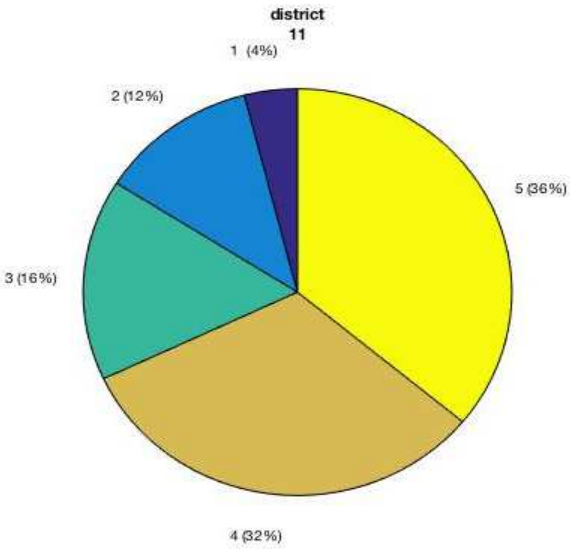

district
6
1 (22%)
2 (7%)
5 (50%)
3 (13%)
4 (9%)

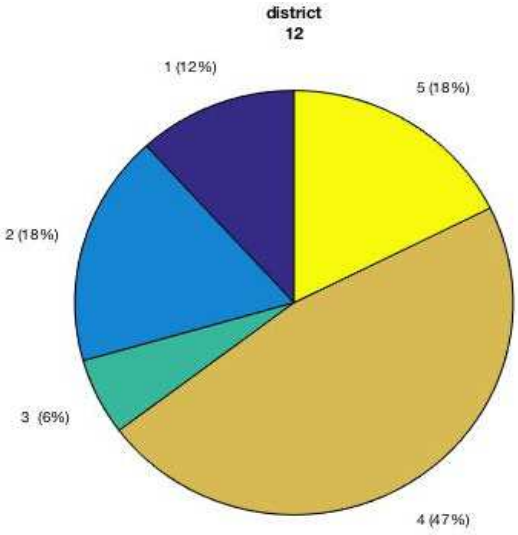

district
7
1 (26%)
5 (32%)
2 (16%)
4 (16%)
3 (11%)


district
8
5 (22%)
1 (44%)
4 (11%)
3 (11%)
2 (11%)


district
9
5 (15%)
1 (23%)
4 (15%)
2 (23%)
3 (23%)


district
10
1 (12%)
5 (35%)
2 (29%)
4 (12%)
3 (12%)


district
11
1 (4%)
2 (12%)
5 (36%)
3 (16%)
4 (32%)


district
12
1 (12%)
5 (18%)
2 (18%)
3 (6%)
4 (47%)

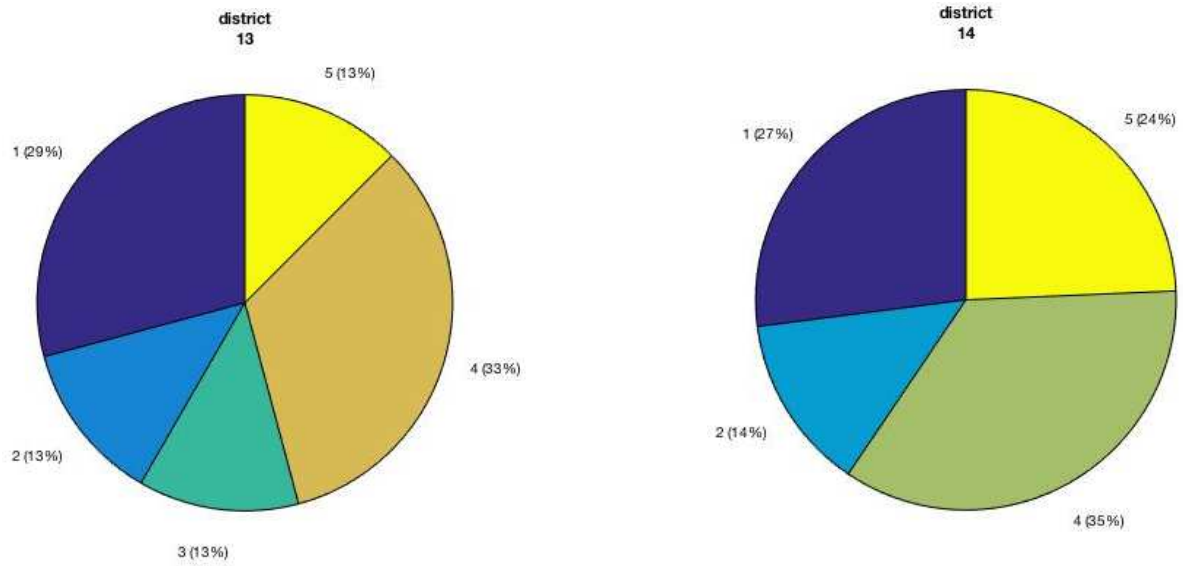


Fig. 3 The pie chart of cluster scatter of each district

TABLE IV
CLUSTER RESULTS

Clusters	Average of sum impact	Category of impact	Number of members
C1	15.93	Very Low	129
C2	16.87	Low	49
C3	17.05	Medium	36
C4	18.30	High	89
C5	19.41	Very High	97
	Total		400

C. Cluster Analysis Using Multi soft Multivariate Distribution Based

The technique called multi soft multivariate distribution (MSMD) based is using to cluster the data. The technique is compared first with the baseline technique, i.e., fuzzy centroid and fuzzy k-partition, to show the performance. The clustering result is analyzed using Dunn Index and stability testing to determine the desired number of clusters. The time response is also recorded to show the time processing the data. In the clustering phase, the variation number of clusters is two until 100. The experiment is run using MATLAB software on an I5 core CPU with 8Gb of RAM. Table IV shows the average response time. MSMD has the fastest processing than the other technique, i.e., 0.0896 seconds. Figure 1 shows the Dunn index versus the increasing number of clusters. It illustrates the stability of the MSMD technique to performance the cluster in terms of Dunn index by an increasing number of clusters. It is also supported by generating a number of clusters of each technique when the number of clusters is determined to increase. It can be shown in figure 2. The MSMD can maintain by creating a number of the cluster itself. Meanwhile, the baseline technique follows by creating the number of the cluster under the determined number of clusters or falling to a single number of clusters. Thus, the MSMD is the desirable technique than the other to analyze the environmental impact data set.

As in Figure 1, the Dunn index of MSMD is consistent on a number of clusters 2-100. to divide the data into several levels of impact, therefore the cluster is determined to be 5 clusters (category of impact). The cluster results are summarized in Table V. The greatest number of member clusters is cluster 1 (very low impact) with 32.25 % of total data, following cluster 5 (Very High impact) with 24.25 % of total data. The district with the most extremely high impact is districts 6, 7, 10, and 11. The district with the highest impact in districts 2, 12, 13, and 14. The distribution of each district can be seen in figure 3. 1,2,3,4,5,8

IV. CONCLUSIONS

This paper studies the use of the methods to cluster the environment affect dataset. The procedure includes two-stage, i.e., highlight determination based on the most extreme reliance quality and clustering utilizing MSMD procedure based on multi delicate set and multinomial dispersion work. The convenience of the procedure has been illustrated as a try comes about by comparing with the standard method. The result appears that the MSDM is more alluring in terms of the solidness based on Dunn record, Rank record, and timely reaction. Two perspectives of environment effect taken from fourteen area datasets are utilized to utilize both methods. The information is separated into 5 clusters as the level of effect. The conveyance of each locale is shown that can be utilized to create a suggestion to progress the quality of the social environment.

REFERENCES

[1] P. Clavel and R. Young, "'Civics': Patrick Geddes's theory of city development," *Landsc. Urban Plan.*, vol. 166, no. June, pp. 37–42, 2017, doi: 10.1016/j.landurbplan.2017.06.017.

[2] A. Kumari and A. K. Sharma, "Physical & social infrastructure in India & its relationship with economic development," *World Dev. Perspect.*, vol. 5, pp. 30–33, 2017, doi: 10.1016/j.wdp.2017.02.005.

[3] P. Ashcroft and L. Murphy Smith, "Impact of environmental regulation on financial reporting of pollution activity: A comparative study of U.S. and Canadian firms," *Res. Account. Regul.*, vol. 20, pp. 127–153, Jan. 2008, doi: 10.1016/S1052-0457(07)00207-X.

[4] J. K. Woo, D. S. H. Moon, and J. S. L. Lam, "The impact of environmental policy on ports and the associated economic opportunities," *Transp. Res. Part A Policy Pract.*, no. xxxx, pp. 0–1, 2017, doi: 10.1016/j.tra.2017.09.001.

[5] P. Ashcroft and L. Murphy Smith, "Impact of environmental regulation on financial reporting of pollution activity: A comparative study of U.S. and Canadian firms," *Res. Account. Regul.*, vol. 20, pp. 127–153, 2008, doi: https://doi.org/10.1016/S1052-0457(07)00207-X.

[6] K. Howard and R. Gerber, “Impacts of urban areas and urban growth on groundwater in the Great Lakes Basin of North America,” *J. Great Lakes Res.*, vol. 44, no. 1, pp. 1–13, 2018, doi: https://doi.org/10.1016/j.jglr.2017.11.012.

[7] A. R. Shahtahmassebi *et al.*, “How do modern transportation projects impact on development of impervious surfaces via new urban area and urban intensification? Evidence from Hangzhou Bay Bridge, China,” *Land use policy*, vol. 77, pp. 479–497, 2018, doi: https://doi.org/10.1016/j.landusepol.2018.05.059.

[8] P. K. Yogyakarta, “Dampak Pertumbuhan Hotel Terhadap Perubahan Karakteristik Perwilayahan Kota Yogyakarta Tika Ainunnisa Fitria,” vol. 10, pp. 52–57, 2016.

[9] M. T. Dugan, E. H. Turner, M. A. Thompson, and S. M. Murray, “Measuring the financial impact of environmental regulations on the trucking industry,” *Res. Account. Regul.*, vol. 29, no. 2, pp. 152–158, 2017, doi: 10.1016/j.racreg.2017.09.007.

[10] C. Mary Schooling, E. W. L. Lau, K. Y. K. Tin, and G. M. Leung, “Social disparities and cause-specific mortality during economic development,” *Soc. Sci. Med.*, vol. 70, no. 10, pp. 1550–1557, 2010, doi: 10.1016/j.socscimed.2010.01.015.

[11] T. Woldai and A. G. Fabbri, “The Impact of Mining on The Environment,” in *Deposit and Geoenvironmental Models for Resource Exploitation and Environmental Security*, A. G. Fabbri, G. Gaál, and R. B. McCammon, Eds. Dordrecht: Springer Netherlands, 2002, pp. 345–364.

[12] I. T. R. Yanto, “Minimum error classification clustering,” *Int. J. Softw. Eng. its Appl.*, vol. 7, no. 5, pp. 221–232, 2013, doi: 10.14257/ijseia.2013.7.5.20.

[13] I. T. R. Yanto, A. Rahman, and Y. Saaadi, “Soft Maximal Association Rule for web user mining,” 2017, doi: 10.1109/ICSITech.2016.7852659.

[14] I. T. R. Yanto, E. Sutoyo, A. Apriani, and O. Verdiansyah, “Fuzzy Soft Set for Rock Igneous Clasification,” 2019, doi: 10.1109/SAIN.2018.8673383.

[15] M. Muhajir and B. Rian, “Association Rule Algorithm Sequential Pattern Discovery using Equivalent Classes (SPADE) to Analyze the Genesis Pattern of Landslides in Indonesia,” vol. 1, no. 3, pp. 158–164, 2015.

[16] N. Senan, R. Ibrahim, N. M. Nawi, I. T. R. Yanto, and T. Herawan, “Soft Set Theory for Feature Selection of Traditional Malay Musical Instrument Sounds,” 2010, pp. 253–260.

[17] I. T. R. Yanto, M. A. Ismail, and T. Herawan, “A modified Fuzzy k-Partition based on indiscernibility relation for categorical data clustering,” *Eng. Appl. Artif. Intell.*, vol. 53, pp. 41–52, Aug. 2016, doi: 10.1016/j.engappai.2016.01.026.

[18] T. Herawan, M. M. Deris, and J. H. Abawajy, “A rough set approach for selecting clustering attribute,” *Knowledge-Based Syst.*, vol. 23, no. 3, pp. 220–231, Apr. 2010, doi: 10.1016/j.knosys.2009.12.003.

[19] N. Senan, R. Ibrahim, N. Mohd Nawi, I. T. R. Yanto, and T. Herawan, *Rough set approach for attributes selection of traditional Malay musical instruments sounds classification*, vol. 151 CCIS, no. PART 2. 2011.

[20] D. W. Jacob, M. F. M. Fudzee, M. A. Salamat, R. R. Saedudin, I. T. R. Yanto, and T. Herawan, *An application of rough set theory for clustering performance expectancy of indonesian e-government dataset*, vol. 549 AISC. 2017.

[21] 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.