JOIV : Int. J. Inform. Visualization, 7(1) - March 2023 1-6

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

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

K-Means Algorithm Analysis for Election Cluster Prediction

Sri Ngudi Wahyuni [a,*], Nazmun Nahar Khanom [b], Yuli Astuti [a]

[a] Department of Informatics Management, Universitas Amikom Yogyakarta, Sleman, Yogyakarta 55283, Indonesia
[b] Department of Information and Communication Technology, University of Professionals, Mirpur Cantonment, Dhaka, 1216, Bangladesh
Corresponding author: *yuni@amikom.ac.id

***Abstract*—The general election is a democratic process that is carried out in every country whose system of government is presidential, including Indonesia, which conducts it every five years. In fact, some people abstain, leading to budget wasting and missing target. Thus, it is very important to identify clusters of general election districts and map the number of voters to map the budget for the upcoming election. This process needs prediction to help reduce budgeting risk as an early warning. Based on the latest election data taken from Margokaton, Yogyakarta, Indonesia, many people voted in 2021, but the number of abstainers is high. In this case, cluster prediction is important to identify the election participants in each area. The K-Means algorithm could also predict abstainer areas in election activities to facilitate early mitigation in drafting election budgeting. Therefore, this study aimed to identify the pattern of voters in the election using the K-means algorithm. The data parameters comprised the list of voters, Unused ballot papers, and the sum of abstainers. This study is important because it contributes to reducing the election budget of each area. The data obtained from the Indonesia Ministry of Internal Affairs official website in 2021 were processed using the RapidMiner tool. The results showed more than 11% of the non-voters in cluster 1, 16% in Cluster 2, and 8% in cluster 3. The evaluation of clusters value is 2.04, indicating that the clustering using K-means is suitable, as shown by the DBI value close to 0. The results indicate that testing the cluster optimization of the K-Means algorithm using DBI is highly recommended. Based on this prediction result, the government needs special attention to clusters with many abstainers to decrease the number of abstainers and prevent overbudgeting. These results indicate the need to review the election participant data in 2024. Furthermore, there is a need for continuous socialization and education about election activities to reduce the number of abstainers and prevent overbudgeting.**

***Keywords*— K-Means algorithm; cluster; prediction; election; Davies Bouldin index.**

Manuscript received 14 Aug. 2022; revised 16 Nov. 2022; accepted 1 Dec. 2022. Date of publication 31 Mar. 2023.

I. Introduction

A village is a legal community unit with territorial boundaries authorized to regulate and manage government affairs and community interests based on initiatives recognized in Indonesia. A village head, the highest leader of the village government, leads it. The head is elected in a democratic party that involves the entire community. Therefore, the village head election is important to determine a candidate suitable as a leader that meets the criteria of a village head. To avoid a budget deficit, it is important to map out the participants and the budget in the village head election. In this regard, K-means is the simplest algorithm that could be used to conduct group mapping in determining the budget for the democratic party activity. The many voter participants that do not exercise their rights in village head election cause losses in the budget, making democratic activities run improperly. The clustering technique aims to find patterns from a set of data [1]. K-means clustering is an unsupervised learning algorithm included in the non-hierarchical cluster analysis used to group data based on variables or features. It is widely used in business, health, education, social, and many other fields. Therefore, the K-means algorithm could be used to cluster data based on their similarities. The clustering is based on the distance between the data and the centroid point of the cluster obtained through iteration. This analysis process requires determining the number of *k* as the input algorithm. K-means could be used for market and image segmentation, image compression, and remote sensing image classification [2]. Election pattern prediction in machine learning is efficient and fast budgeting, making it very important.

Previous studies employed the K-Means algorithm, such as in dengue prediction clustering [3], [4]. Tada et al. [5] used the K-means algorithm for clustering semiconductor detectors. Ahmad and Dey [6] focused on clustering data, Khawaja et al. [7] clustered networking, and Vadyala [8] examined COVID-19 by combining the K-means with LSTM algorithms. Grant et al. [9] examined a clinical profile of patients at Kaiser

Permanente, Northern California. The study grouped patients based on comorbidity scores and previous emergency room admissions to map health services through clinical interpretation. The mapping of health services was conducted correctly, and prioritized the cluster of the patient groups. Furthermore, Cheves et al. detected clusters of residential electricity consumption areas in Argentina. The study recommended determining the typology of house construction and its quality to achieve energy efficiency. The results indicated an average electricity consumption of each census radius of 1010, which is Great La Plata [10]. Debao, Yinxia, and Min [11] also mapped the job applicants' talents through the Chinese recruitment website zhaopin.com.

Another study used the K-means to cluster energy consumption for agriculture to control water due to minimal rain and boost food security in South Africa [12]. Moreover, using spatial parameters, Janrao, Mishra, and Bharadi [13] clustered sugarcane zones to plant management and fertilization. The study found that the K-means algorithm is recommended for clustering and contributes to food security efforts in India. Trivedi et al. performed disease detection on leaves based on the images of the leaves using K-means. The results were used to deal with plant diseases [14], [15]. Furthermore, Boz [16] conducted an agricultural assessment by clustering tea plantations with livestock to test economic sustainability in Turkey. Ghassemi and Hashemi [17] clustered transportation allocation by grouping texts from the web [11], [18]. Other studies also used the algorithm to cluster food markets and security [18] and earth vulnerability [19]. Similarly, the algorithm was used to map image processing [20]–[24] and healthcare tomography images [25], [26]. The algorithm was widely used to map areas with a higher spread of COVID-19, including Indonesia, where the clustering was used to determine regional lockdowns. The aim was to reduce the spread of the pandemic from high-risk areas to other regions. Therefore, the algorithm could help decrease the spread in several countries [27], [28].

This study focused on clustering election participants. Previous studies examined the selection of wireless sensor networking [29]–[31] but did not discuss election participant cluster predictions using the K-means algorithm. Therefore, it is important to investigate the election participant data and the election budgeting for 2024. The next sections present the study methodology, results and discussion, conclusion, and references.

II. Material and Method

A. Dataset processing

Data were taken from the Indonesia Ministry of Internal Affairs official website in 2021. The parameters used were the voters' list, unused ballot papers, and the number of Abstainers. The data collected were processed using RapidMiner. Figure 1 shows the stages of the study. The first step was data preprocessing to remove the information of eligible participants that died from the voter list. The aim was to remain with only participants alive, aged at least 17, with ID cards and allowed to participate in general elections. The next step was determining the number of clusters or *k* in the population studied. The data centroid was also determined to define Euclidian distance by conducting iteration. The final step was cluster performance evaluation using the Davies–Bouldin index.

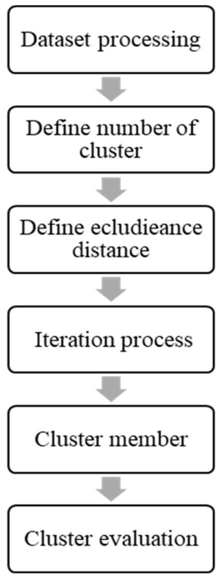

Fig. 1 Methodology

B. K-means Algorithm

K-means algorithm could overcome clustering problems. It is a fast-iterative algorithm widely used in many point-based grouping applications that begin with a central cluster initially placed randomly and moved from the cluster's center. This algorithm uses several iterative processes to approach the optimal value [32]. To solve this heuristic region grouping problem, the algorithm starts with the initial candidate solution *{c1, ..., ck}* $\subset$ *Rd* chosen arbitrarily as P until every ci. The algorithm calculates all Pi's and sets the points in *P* closest to ci. For every *1* $\leq i \leq k$, it replaces ci with the average of *Pi* because this calculation is derived from the average set of Pi [33]. K-Means Algorithm is a typical grouping algorithm in data mining. It is the simplest and most widely used for grouping large amounts of data. MacQueen first proposed the algorithmo solve data clustering in 1967 [34]. The phases of the K-means algorithm are: (1) k-values are selected randomly; (2) the closest distance from each point to the center point is measured. This iteration is performed to obtain the smallest criterion function. When the target object x, x_i indicates the average of the c_i cluster, the criteria function is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{x=c_i} |x - x_i|^2 \quad (1)$$

Where E is the sum squared error of all objects in the database. This phase is repeated until there is no change in the median value in the cluster. In this study, the initial step of K-means was to calculate the distance between the data and the centre of the cluster using Euclidian distance. The Euclidean distance between one vector $x = (x_1, x_2, \ldots x_n)$ and another vector $y = (y_1, y_2, \ldots y_n)$. Euclidean distance $d(x_i$, *d (xi, yi)* is obtained as follows [35]:

$$d(x_t, y_t) = \left[\sum_{i=1}^{n} (x_i - y_i)^2\right]^{1/2} \quad (2)$$

where,
D (i, j): Distance from data-*i* to the cluster centre *j*.
Xki: Data-*i* to data-k attribute

Xkj: Central-*j* to data-k attribute.
Expected number of clusters *k*, and database *D = {d1, d2, ... dn}* comprising data objects.

C. Cluster Evaluation

Davies-Bouldin Index is a method used to test cluster performance. It is an internal validation process for clusters based on data density with centroids and separations and ratios between clusters. The calculation results of the three values were used to obtain Bouldin-Davies Index [36], [37]. The steps for defining the DB index value are included in the formula below:

1) *First,* the result of computing the Sum of the Square Within (SSW) the Cluster was used to define the distribution in one cluster class.

$$SSW_i = \frac{1}{m_i}\sum_{j=1}^{m_i} d(x_j, c_i) \quad (3)$$

The value of *d* in equation 3 could use the dissimilar formula employed in the grouping process. This ensures that the validation has the same purpose as the grouping process.

2) *Second,* the Sum of squares between clusters was defined using equation 4.

$$SSB_{I,J} = D(c_i, c_j) \quad (4)$$

Where SSB = Sum of Square Between cluster and $d\ (ci..cj)$ is Centroid distance ci with centroid *cj*.

3) *Third,* the R and DBI values were computed using equations 5 and 6.

$$R_i = maxR_{ij} \;\; j = 1, \dots k, i \neq j \quad (5)$$

$$DBI = \frac{1}{k}\sum_{i=1}^{k} R_i \quad (6)$$

Where R is the distance between clusters, *Var* is the variance from data, x_i is data, and $\overline{x}$ is the average of each cluster. Equation 7 shows the variance of the data formula.

$$var(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \quad (7)$$

III. Result and Discussion

A. Iteration process

The first experiment step was randomly determining the *k* values in Table I.

TABLE I
THE FIRST CENTROID

Centroid	Voters	Unused ballot papers	Abstainers
Centroid 1	81	2	7
Centroid 2	102	3	17
Centroid 3	108	6	13

Table I shows that the first centroid was taken randomly and has 60 rows of data. The first, second, and third centroids were taken from the data's 4th, 45th, and 60th rows. The next step was to determine the Euclidean distance using equation 2. Using the first center distance (*d4, d45,* and *D60*) would result in a Euclidean distance of all data. The distance was determined as shown in the following formulas.

$$d4 = \sqrt{(68-81)^2 + (3-2)^2 + (9-7)^2} = 13.15$$

$$d45 = \sqrt{(68-102)^2 + (3-3)^2 + (9-17)^2} = 34.15$$

$$d60 = \sqrt{(68-108)^2 + (3-6)^2 + (9-13)^2} = 40.62$$

Table II shows all the Euclidean distances in the first iteration.

TABLE II
THE FIRST EUCLIDEAN DISTANCE IN THE FIRST ITERATION

Attributes	C1	C2	C3
I1	13.151487	34.604138	40.628837
I2	13.779753	32.175544	38.504026
I3	2.6	24.843317	30.506681
…	…	…	…
I59	26.07681	3.6461487	5.5937108
I60	28.294339	7.2067746	0
Total	1057.0235	632.29754	853.67025

Table II indicates that the 60 rows of data have three clusters, each with several members of data rows. The experiment result has seven iterations. The process stops at the seven iterations when similar values are obtained. Table III presents the centroids in all iterations.

TABLE III
ALL CENTROID IN ALL ITERATION

Centroid	Voters	Unused ballot papers	Abstainers
1st iteration	81	2	7
	102	3	15
	108	6	13
2nd Iteration	85	2	26
	100	3	13
	106	3	11
3rd Iteration	82	1	13
	96	3	12
	105	3	11
4th Iteration	81	1	7
	95	3	12
	104	3	12
5th Iteration	78	1	4
	93	3	12
	104	3	12
6th Iteration	77	1	3
	92	3	12
	103	3	13
7th Iteration	76	1	2
	91	3	13
	103	3	12
8th Iteration	76	1	2
	91	3	13
	103	3	12

The centroid value in Table III was computed based on *D (i, j),* where the distance from data-*i* to the cluster center *j*; *Xki*; data-*i* to data-k attribute; and *Xkj* central-*j* to data-k attribute. Each iteration produces a Euclidean distance that varies with the arrangement of the centroids in each iteration. Table III shows the similarity of the iteration results between the 7th and 8th iterations, and the iteration stops in iteration 7. Table IV presents the results of the exclusion distance in the last iteration.

TABLE IV
THE LAST OF EUCLIDEAN DISTANCE OF THE LAST ITERATION

Attribute	C1	C2	C3
I1	11.22592685	23.61772146	35.53085105
I2	14.73989398	21.57737869	33.57207077
I3	5.430940041	13.96937899	25.37742808
…	…	…	…
I59	31.92466397	14.36004971	5.008897683
I60	33.69051759	17.30145914	5.67654261
Total	1368.661896	643.9760345	638.2480347

Table IV presents the centroid of each cluster in iteration one. This step was repeated until the last centroid, whose composition is defined in the last iteration. Table IV is the result of the ecludiance distance in the 7th iteration. Table V shows the result of the last iteration.

TABLE V
THE RESULT OF THE LAST ITERATION

C1	C2	C3	Voters	Unused ballot papers	Abstainers
1			68	3	9
1			70	4	15
1			78	3	7
1			81	2	7
1			79	3	10
1			84	2	7
1			76	3	15
	2		81	2	14
	2		94	3	0
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
	2		92	5	17
	2		87	2	22
	2		96	3	25
	2		96	2	25
		3	98	3	4
		3	102	5	9
…	…	…	…	…	…
		3	100	2	11
		3	105	5	16
		3	105	4	17
		3	108	6	13

Table V presents the members of each cluster and the election pattern shown in Fig 2.

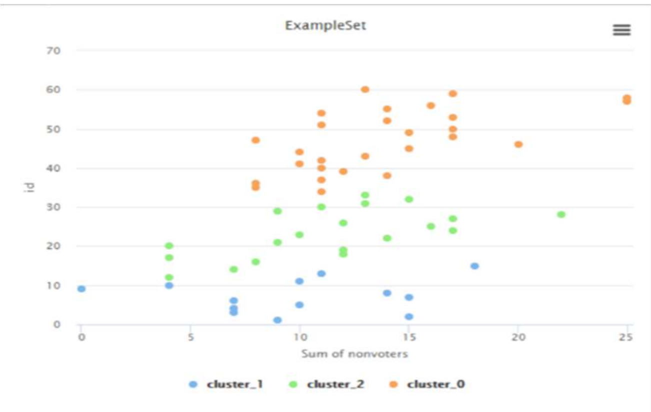

Fig. 2 The pattern of election

Fig 2 explains the cluster members of the election pattern. Clusters one, two, and three have 7, 22, and 31 members, respectively. Therefore, the study focused on cluster three, with the highest number of non-participants in election, implying reduced election budgets. The sum of non-voter letters is shown in Table VI.

TABLE VI
THE SUM OF THE ELECTION BREAKDOWN IS BASED ON CLUSTERING

Clusters	Voters	Unused ballot papers	Abstainers
C1	195	79	1570
C2	296	49	1546
C3	239	64	2512

Table VI summarizes each cluster's content and the variable composition. The second cluster has the highest number of abstainers. Therefore, the cluster was mapped as shown in Figure 3.

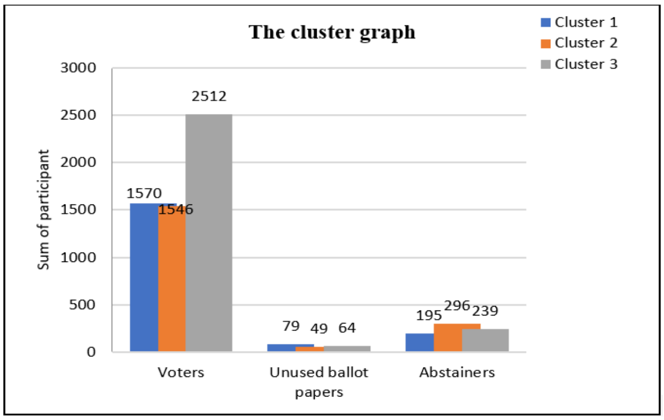

Fig. 3 The Cluster content graph

It can be seen in Fig 3 that cluster two has the highest number of abstainers compared to other clusters. Therefore, this area needs more attention and continued education about awareness of using voting or data checking in the next election. The data in Table II show that the sum of abstainers is 296 in cluster two.

B. *Evaluation Cluster Process using DBI*

The last step was evaluation cluster analysis using DBI Index. The analysis began by computing a centroid using the following formula and defining the SSW. Using the Tabel V result, the SSW will compute. For example, we used the cluster 1 members computation SSW shown in Table VII.

TABLE VII
THE MEMBER OF CLUSTER 1

Voters	Unused ballot papers	Abstainers	SSW of Sum of Voters	SSW of Unused ballot papers	SSW of abstainer
9	3	68	1.00	0.02	77.04
15	4	70	25.00	1.31	47.02
7	3	78	9.00	0.02	3.18
7	2	81	9.00	0.73	17.50
10	3	79	0.00	0.02	7.52
7	2	84	9.00	0.73	49.88
15	3	76	25.00	0.02	0.02
70	20	535	78.00	2.86	202.1595

Based on Table VII above, the example computation for the 1st-row data is

$$c1 = \frac{9+15+7+7+10+7+15}{7} = 10$$

And the SSW is

$$SSW = \sum \left(9 - \frac{68+70+78+81+79+84+76}{7}\right)^2$$

This computation is the same process for all clusters. Based on this step will get the result in Table VIII. Table VIII is the average of each cluster from Table VII.

TABLE VIII
THE CENTROID OF THE EACH CLUSTER FOR DBI

Cluster	Voters	Unused ballot papers	Abstainers
Cluster 1	10.00	2.86	76.46
Cluster 2	2.82	11.59	90.89
Cluster 3	16.65	3.10	99.78

Table VIII presents that the abstainers value in the third cluster is higher than others, which is 99.78. The next step was to compute the centroid distribution (*S*) using the following formula.

$$S = \left(\frac{1}{7} x 78\right)^2 = 3.34$$

The distribution of every centroid is shown in Table IX.

TABLE IX
THE SSB STRUCTURE

Clusters	Voters	Unused ballot papers	Abstainers
S1	3.34	0.64	5.37
S2	1.27	6.13	4.00
S3	22.06	1.67	5.52

Table IX presents the distribution of each centroid. The cluster 3 centroids distribution is bigger than others, especially the abstainers' distribution, which is 5.52. The next step evaluation process is, to compute the M value, M is the distance each centroid in each cluster. The results were computed using the following formula to compute the distance between the centroids.

$$M_{1-2} = |7.18 - 2.82| = 6.65$$

The results are presented in Table X.

TABLE X
THE CENTROID DISTANCE

Clusters	Voters	Unused ballot papers	Abstainers
M1-2	7.18	8.73	14.44
M1-3	6.65	0.24	23.32
M2-3	13.83	8.49	8.88

Table X shows the centroid distance in each parameter. These steps must be finished for all centroids in all clusters. The last step was cluster evaluation using Davied Bouldin Index. Equation 5 was used to compute the R-value shown in Table XI.

TABLE XI
RESULT OF R-VALUE

Value	Maximum value	DBI
D1	0.64	
D2	2.75	2.04
D3	2.75	

Table XI presents the R-value of each parameter. The data on R1,2 is generated from the sum of S1 and S2 divided by the value of M1,2. So it and the result 0.64, etc. The next, equation 6 was used to compute the DBI value shown in Table XII.

TABLE XII
RESULT OF DBI CALCULATION

Distance	Voters	Unused ballot papers	Abstainers
R 1,2	0.64	-0.78	-0.65
R 1-3	-3.82	-106.00	-1.09
R 2-3	-1.69	2.75	-2.63

Table XII presents the Davies Boulden Index calculation results. This value was obtained by dividing the sum of all maximum values of each centroid by the number of clusters. The value of D1 is taken from the comparison between R1,2 and R1,3. Meanwhile, the value of D2 is taken from the comparison value between R1,2 and R2,3, and the value of D3 is taken from the comparison of R1,3 and R2,3. While the DBI value is obtained from 1/3 x (D1+D2+D3), so the result is 2.04. This DBI value is close to 0, so this evaluation cluster can be used to approach the case in this study.

IV. CONCLUSION

The experiment result showed more than 11% of non-voters in cluster 1, 16% in Cluster 2, and 8% in cluster 3. The cluster evaluation value is 2.04, indicating that clustering using K-means is recommended for this case. The DBI value is close to 0, meaning the K-means model is recommended for use in clustering. The results showed the need to review two important things. First, it is important to review the election of participant data in 2024 based on the sum of non-voters. Second, there is a need for continuous socialization and education about election activities. This would help reduce the number of non-voters and the election budget. Several village areas in cluster one have many non-voters. Therefore, future studies should compare several algorithms to define the best clustering prediction. They should also use several parameters in clustering prediction experiments.

ACKNOWLEDGMENT

The authors thank Universitas Amikom Yogyakarta, Indonesia, for supporting this study.

REFERENCES

[1] S. Zahi and B. Achchab, "Clustering of the population benefiting from health insurance using K-means," in *Proceedings of the 4th International Conference on Smart City Applications*, 2019, pp. 1–6.

[2] S. A. Rizvi, M. Umair, and M. A. Cheema, "Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators," *Chaos Solitons Fractals*, vol. 151, p. 111240, 2021, doi: https://doi.org/10.1016/j.chaos.2021.111240.

[3] H. L. Nguyen, "Specific K-mean clustering-based perceptron for dengue prediction," *International Journal of Intelligent Information and Database Systems*, vol. 10, no. 3. pp. 269–288, 2017. doi: 10.1504/IJIIDS.2017.087242.

[4] P. Manivannan, "Dengue fever prediction using K-means clustering algorithm," *Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017*, vol. 2018. pp. 1–5, 2018. doi: 10.1109/ITCOSP.2017.8303126.

[5] T. Tada, K. Hitomi, Y. Wu, S. Y. Kim, H. Yamazaki, and K. Ishii, "K-mean clustering algorithm for processing signals from compound semiconductor detectors," *Nucl Instrum Methods Phys Res A*, vol. 659, no. 1, pp. 242–246, Dec. 2011, doi: 10.1016/J.NIMA.2011.09.007.

[6] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data Knowl Eng*, vol. 63, no. 2, pp. 503–527, Nov. 2007, doi: 10.1016/J.DATAK.2007.03.016.

[7] S. G. Khawaja, M. Usman Akram, S. A. Khan, A. Shaukat, and S. Rehman, "Network-on-Chip based MPSoC architecture for k-mean clustering algorithm," *Microprocess Microsyst*, vol. 46, pp. 1–10, Oct. 2016, doi: 10.1016/J.MICPRO.2016.08.006.

[8] S. R. Vadyala, "Prediction of the Number of COVID-19 Confirmed Cases Based on K-Means-LSTM (Preprint)." JMIR Publications Inc., 2020. doi: 10.2196/preprints.20798.

[9] S. W. Grant, G. L. Hickey, A. D. Grayson, D. C. Mitchell, and C. N. McCollum, "National risk prediction model for elective abdominal aortic aneurysm repair," *British Journal of Surgery*, vol. 100, no. 5. Wiley, pp. 645–653, 2013. doi: 10.1002/bjs.9047.

[10] P. Chévez, D. Barbero, I. Martini, and C. Discoli, "Application of the k-means clustering method for the detection and analysis of areas of homogeneous residential electricity consumption at the Great La Plata region, Buenos Aires, Argentina," *Sustain Cities Soc*, vol. 32, pp. 115–129, 2017.

[11] D. Debao, M. Yinxia, and Z. Min, "Analysis of big data job requirements based on K-means text clustering in China," *PLoS One*, vol. 16, no. 8 August, Aug. 2021, doi: 10.1371/JOURNAL.PONE.0255419.

[12] M. Mtshali, S. Dlamini, M. Adigun, and P. Mudali, "K-means based on resource clustering for smart farming problem in fog computing," in *2019 IEEE AFRICON*, 2019, pp. 1–6.

[13] P. Janrao, D. Mishra, and V. Bharadi, "Clustering approaches for management zone delineation in precision agriculture for small farms," in *Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India*, 2019.

[14] R. Trivedi and S. Khadem, "Peak Demand Management and Schedule Optimisation for Energy Storage through the Machine Learning Approaches," *IEEE EUROCON 2021-19th* ..., 2021, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9535559/

[15] T. N. Tete and S. Kamlu, "Detection of plant disease using threshold, k-mean cluster and ann algorithm," in *2017 2nd International Conference for Convergence in Technology (I2CT)*, 2017, pp. 523–526.

[16] I. Boz, "Measuring environmental, economic, and social sustainability index of tea farms in Rize Province, Turkey," *Environ Dev Sustain*, vol. 22, no. 3, pp. 2545–2567, 2020.

[17] F. Ghassemi Tari and Z. Hashemi, "Prioritized K-mean clustering hybrid GA for discounted fixed charge transportation problems," *Comput Ind Eng*, vol. 126, pp. 63–74, Dec. 2018, doi: 10.1016/J.CIE.2018.09.019.

[18] M. Tleis, R. Callieris, and R. Roma, "Segmenting the organic food market in Lebanon: an application of k-means cluster analysis," *British Food Journal*, 2017.

[19] S. C. Babu, S. N. Gajanan, and P. Sanyal, "Classifying Households On Food Security and Poverty Dimensions—Application of K-Mean Cluster Analysis," *Food Security, Poverty and Nutrition Policy Analysis*, pp. 417–439, 2014, doi: 10.1016/B978-0-12-405864-4.00013-2.

[20] T. Wei, X. Wang, X. Li, and S. Zhu, "Fuzzy subspace clustering noisy image segmentation algorithm with adaptive local variance & non-local information and mean membership linking," *Eng Appl Artif Intell*, vol. 110, Apr. 2022, doi: 10.1016/J.ENGAPPAI.2022.104672.

[21] R. Yu *et al.*, "Feature discretization-based deep clustering for thyroid ultrasound image feature extraction," *Comput Biol Med*, Jul. 2022, doi: 10.1016/J.COMPBIOMED.2022.105600.

[22] C. Wu and Z. Wang, "A modified fuzzy dual-local information c-mean clustering algorithm using quadratic surface as prototype for image clustering," *Expert Syst Appl*, vol. 201, Sep. 2022, doi: 10.1016/J.ESWA.2022.117019.

[23] K. Ren, Y. Ye, G. Gu, and Q. Chen, "Feature matching based on spatial clustering for aerial image registration with large view differences," *Optik (Stuttg)*, vol. 259, Jun. 2022, doi: 10.1016/J.IJLEO.2022.169033.

[24] A. Abernathy and M. E. Celebi, "The incremental online k-means clustering algorithm and its application to color quantization," *Expert Syst Appl*, vol. 207, p. 117927, Nov. 2022, doi: 10.1016/J.ESWA.2022.117927.

[25] R. W. Grant *et al.*, "Use of latent class analysis and k-means clustering to identify complex patient profiles," *JAMA Netw Open*, vol. 3, no. 12, pp. e2029068–e2029068, 2020.

[26] A. Risheh, P. Tavakolian, A. Melinkov, and A. Mandelis, "Infrared computer vision in non-destructive imaging: Sharp delineation of subsurface defect boundaries in enhanced truncated correlation photothermal coherence tomography images using K-means clustering," *NDT and E International*, vol. 125, Jan. 2022, doi: 10.1016/J.NDTEINT.2021.102568.

[27] S. Ilbeigipour, A. Albadvi, and E. Akhondzadeh Noughabi, "Cluster-based analysis of COVID-19 cases using self-organizing map neural network and K-means methods to improve medical decision-making," *Inform Med Unlocked*, vol. 32, p. 101005, 2022, doi: https://doi.org/10.1016/j.imu.2022.101005.

[28] J. A. L. Marques, F. N. B. Gois, J. Xavier-Neto, and S. J. Fong, "Predicting the geographic spread of the COVID-19 pandemic: a case study from Brazil," in *Predictive Models for Decision Support in the COVID-19 Crisis*, Springer, 2021, pp. 89–98.

[29] A. Sharma and S. Chauhan, "Sensor fusion for distributed detection of mobile intruders in surveillance wireless sensor networks," *IEEE Sens J*, vol. 20, no. 24, pp. 15224–15231, 2020.

[30] Y. Mekonnen, S. Namuduri, L. Burton, and ..., "Machine learning techniques in wireless sensor network based precision agriculture," *Journal of the ...*, 2019, doi: 10.1149/2.0222003JES.

[31] S. Mostafavi and V. Hakami, "A new rank-order clustering algorithm for prolonging the lifetime of wireless sensor networks," *International Journal of Communication Systems*, vol. 33, no. 7, p. e4313, 2020.

[32] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognit*, vol. 36, no. 2, pp. 451–461, 2003, doi: 10.1016/S0031-3203(02)00060-2.

[33] Y. Religia and A. S. Sunge, "Comparison of Distance Methods in K-Means Algorithm for Determining Village Status in Bekasi District," *Proceeding - 2019 International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, pp. 270–276, 2019, doi: 10.1109/ICAIIT.2019.8834604.

[34] V. Faber, "Clustering and the continuous k-means algorithm," *Los Alamos Sci*, vol. 22, no. 138144.21, p. 67, 1994.

[35] N. Shi, X. Liu, and Y. Guan, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," *3rd International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010*, pp. 63–67, 2010, doi: 10.1109/IITSI.2010.74.

[36] J. Carlos, R. Thomas, M. S. Peñas, and M. Mora, "New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance," no. 1, 2013, doi: 10.1109/SCCC.2013.29.

[37] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 725, no. 1, p. 012128.