International Journal on Advanced Science Engineering Information Technology

Vol.13 (2023) No. 2
ISSN: 2088-5334

Mapping the Provincial Food Security Conditions in Indonesia Using Cluster Ensemble-Based Mixed Data Clustering-Robust Clustering with Links (CEBMDC-ROCK)

Vita Ratnasari [a,*], Andrea Tri Rian Dani [b]

[a] *Department of Statistics, Faculty of Science and Data Analytics, Sepuluh Nopember Institute of Technology (ITS), Surabaya, Indonesia*
[b] *Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Samarinda, Indonesia*

Corresponding author: [*]vita.statistikaits@gmail.com

***Abstract*—Problems related to food are indeed an issue that continues to be discussed by the government, both imports, self-sufficiency, the issue food security. Food security conditions have become one of the biggest problems in Indonesia, even though Indonesia is an agricultural country with abundant resources. The problem is not only the availability but also the affordability. It happens due to the social inequality between the rich and the poor, which means the rich can easily relish food. People with low incomes experience food insecurities. Thus, an appropriate strategy and policies can be done for each province in Indonesia to make it equal. Cluster analysis is used to map the provincial profile based on the condition of food security. However, the variable types in this research are numerical and categorical data, which makes general cluster analysis insufficient. This study used the Cluster Ensemble Based Mixed Data Clustering-Robust Clustering Using Links (CEBMDC-ROCK) method to cluster provinces in Indonesia based on food security conditions. The analysis process starts with numerical clustering data using Agglomerative Hierarchical Clustering (AHC) and then with categorical data using Robust Clustering Using Links (ROCK). The result shows that the province in Indonesia is divided into five groups based on the quality of food security, which is from very low to excellent. Based on the clustering results, which provinces need special attention from the government regarding food security can be seen.**

***Keywords*— Cluster analysis; food security; mixed data; robust clustering.**

Manuscript received 13 Oct. 2021; revised 8 Jan. 2023; accepted 5 Mar. 2023. Date of publication 30 Apr. 2023.

I. INTRODUCTION

Food is a basic human need, not only to fulfill the human rights of every people or as a moral obligation but also as an economic and social investment in forming a better generation in the future [1]–[3]. However, the availability of food that is smaller than the need can create economic instability. As a result, various social problems and political stability can occur if food security is disturbed [4]. Critical food conditions can disrupt economic and national stability [5]. Therefore, the government is continually working to improve society's food security through domestic production and additional imports.

The World Health Organization (WHO) defines three main components of food security. The first is access to food, the second is availability, and the last is food utilization. The Food and Agriculture Organization (FAO) has now added a fourth pillar, the three pillars of stability over the long term [5], [6]. From this definition, it can be seen that food security is a rather complex issue, and therefore, it can be seen that the participation of all sectors of society is necessary to make good food security [7]. Food security in Indonesia has been regulated in Law Number 18 Article 1 of 2012 [8].

The issue of food is indeed an issue that continues to be discussed by the government, both about imports, self-sufficiency, and food security [9]. Especially Indonesia, as an agricultural country with fertile soil and abundant resources, is still struggling with food supply problems [10]. Indonesia itself ranks 5th on food security scores from 10 countries in ASEAN. The rating of Indonesia is not good compared with other countries, considering Indonesia has enormous potential compared to other ASEAN countries.

Recently, the World Bank has also highlighted the food access problem in Indonesia, especially for vulnerable and poor groups [8]. The issue of food security in Indonesia is not only a matter of availability but also affordability. It causes

food imbalance in Indonesia because it is only enjoyed by financially capable groups and cannot be accessed by the poor. It means that people with low incomes experience food insecurities or food insecurity. Through these problems, it can be seen that there is inequality or disparity in food security in the territory of Indonesia [11]. It is certainly not easy to achieve equitable food security distribution in Indonesia. In overcoming the problem of food security in Indonesia, it is necessary to provide accurate, comprehensive, and well-organized food security information.

To be able to overcome inequality or food security gaps in Indonesia, as well as to be able to determine the right policy in formulating a food security strategy in Indonesia, it is necessary to map the provincial profile based on the condition of food security. Provinces can be considered the smallest unit of a government, with the hope that the policies set later become more targeted so that food security in Indonesia becomes more evenly distributed. Provincial mapping can be done through one of the statistical methods, namely cluster analysis.

Cluster analysis is an essential method in statistical data processing to perform data analysis [12]–[14]. Cluster analysis can be used to group objects or data into a group based on their similarity [15], [16]. There are two methods of grouping contained in cluster analysis, hierarchical and non-hierarchical clustering [17], [18]. Hierarchical clustering is a method that seeks to build a hierarchy until a dendrogram is formed [19], [20]. Unlike hierarchical clustering, non-hierarchical clustering starts with defining the desired number of clusters[21]. Then the clustering process is performed without following a hierarchical method. Finding the desired number of clusters in advance was problematic, so researchers are usually more interested in hierarchical clustering [22].

Several well-known cluster algorithms in hierarchical clustering include Single, Complete, and Average linkage [23]–[25]. Single, Complete, and Average linkage algorithms can be applied to numerical data [26]. As for categorical data, you can use Robust Clustering Using the Links (ROCK) method, which is a development of the hierarchical method for categorical data and has good accuracy [27], [28]. The ROCK method is robust for outlier data because it can handle outlier data quite well [18]. Using a link, the ROCK method measures the similarity between a pair of data points. Observations with a high link were combined into one cluster, while those with a low link were separated [27]. The number of links between observations highly depends on the specified threshold value (θ). In practice, errors in determining (θ) can result in errors in the clustering, where each observation was be in the same cluster or each observation was in a different cluster [29].

The reality that is developing at this time where the available data conditions are no longer in a single form (numeric or categorical) but is a combination of mixed scale data [30], [31]. The cluster method developed and used for mixed-scale data is Cluster Ensemble Based Mixed Data Clustering (CEBMDC), developed by [32]. CEBMDC is a clustering method to combine the grouping results from several clustering algorithms to get an optimal cluster [33]. Several previous studies have developed the CEBMDC method, including [16], [31], [34]–[36].

This study uses the Cluster Ensemble Based Mixed Data Clustering-Robust Clustering Using Links (CEBMDC-ROCK) method to cluster Provinces in Indonesia based on food security conditions. The CEBMDC-ROCK method was chosen because the research variables related to food security used in the study are a mixed data scale, which is numerical and categorical. The results of the clustering using the CEBMDC-ROCK are expected to describe the profile of the Provinces in Indonesia by the quality of their food security so that they can then become recommendations for the Central Government and Regional Governments in determining the strategy for developing Indonesian food security so that food security in Indonesia is more evenly distributed.

II. Material and Method

A. Cluster Analysis

Cluster analysis is a method in multivariate analysis to group n observations into C groups $(C \leq n)$ based on their characteristics. The purpose of cluster analysis is to cluster an object that is most similar to another object to be in the same cluster and have similarities [17].

1) Numerical Data Clustering: The Numerical data clustering method is based on the size of the dissimilarity or distance. The dissimilarity measure commonly used is the Euclidean distance [37].

$$d_{uv} = \sqrt{(\boldsymbol{x}_u - \boldsymbol{x}_v)^T(\boldsymbol{x}_u - \boldsymbol{x}_v)}; u, v = 1, \ldots, n \text{ and } u \neq v \quad (1)$$

with $\boldsymbol{x}_u{}^T = [x_{1u}, x_{2u}, \ldots, x_{nu}]$ and $\boldsymbol{x}_v{}^T = [x_{1v}, x_{2v}, \ldots, x_{nv}]$. The clustering method used in this study is a hierarchical grouping method called Agglomerative Hierarchical Clustering (AHC). The following are some algorithms for clustering using AHC [38].

- *Single Linkage.* Single linkage is a clustering algorithm that uses the minimum distance. In general, a single linkage algorithm is written in Equation 2.

$$d_{(UV)W} = min\{d_{UV}, d_{VW}\} \quad (2)$$

- *Complete Linkage:* Complete linkage is a clustering algorithm that uses the maximum distance. The complete linkage algorithm is shown in Equation 3.

$$d_{(UV)W} = max\{d_{UV}, d_{VW}\} \quad (3)$$

- *Average Linkage.* Average linkage is a clustering technique from the average distance between objects by the algorithm shown in Equation 4.

$$d_{(UV)W} = \frac{d_{UW} + d_{VW}}{n_{UV} n_W} \quad (4)$$

2) Categorical Data Clustering: The clustering of categorical data is carried out using similarity or distance measures for categorical data. Then, clustering can be carried out using hierarchical or non-hierarchical methods. However, these methods are not suitable for use with categorical data. Therefore, an appropriate method has been developed, namely the Robust Clustering Using Links (ROCK). A new concept is formed in the ROCK method, namely the link. Clustering using the ROCK method with the following stages:

- Measuring Similarity. The measure of similarity between the u-th and v-th observation pairs is calculated by the formula shown in Equation 5.

$$sim(X_u, X_v) = \frac{|X_u \cap X_v|}{|X_u \cup X_v|}, u \neq v \quad (5)$$

where
$u = 1,2,\dots,n$and $v = 1,2,\dots,n$
X_u : The u-th observation set with $X_u = \{x_{1u}, \dots, x_{m_{categorical}u}\}$
X_v : The v-th observation set with $X_v = \{x_{1v}, \dots, x_{m_{categorical}v}\}$

- Determine the Nearest Neighbor. Observations X_u with X_v can be considered as neighbors if the value $sim(X_u, X_v) \geq 0$. The threshold value (θ) used is at an interval of 0 to 1, according to the available data.
- Link. If there are observations of X_u, X_v, and X_w, with X_u is neighbor of X_v, and then X_v is a neighbor of X_w, it said that X_u has a link with X_w even though X_u is not a neighbor of X_w. The way to calculate links for all possible pairs of n objects can use a matrix $\boldsymbol{A}$. Matrix $\boldsymbol{A}$ is a matrix of size 1 if X_u and X_v are declared similar and 0 if X_u and X_v are declared dissimilar (not neighbors). The number of links between pairs of X_u and X_v is obtained from the product of the row to X_u and column to X_v of matrix $\boldsymbol{A}$. If the link from X_u and X_v is getting bigger, the more likely it is that X_u and X_v are in the one cluster.
- Local Heap. Local heap is the Goodness Measure value for each group with other groups if the link is not equal to zero. Goodness Measure is an equation that calculates the number of links divided by the possible links formed based on the size of thegroup. Goodness Measure can be calculated by the formula in Equation 6.

$$g(C_u, C_v) = \frac{link(C_u, C_v)}{(n_u + n_v)^{1+2f(\theta)} - n_u^{1+2f(\theta)} - n_v^{1+2f(\theta)}} \quad (6)$$

with
$\text{link}(C_u, C_v) = \sum_{X_u \in C_u, X_v \in C_v} link(X_u, X_v)$ is the number of links from all possible pairs of objects.
n_u and n_v is the number of members in group-u and v.
$f(\theta) = \frac{1-\theta}{1+\theta}$ with θ is the threshold value.

- Determine the Global Heap, which is the max value of Goodness Measure between columns in row-u.
- Perform steps (d) and (e) until the maximum value is obtained in the Local Heap and Global Heap.
- As long as the data size is more than k, with k being the specified number of classes, the cluster with the largest Local Heap value and the largest Global Heap is merged into one cluster.

3) Mixed Data Clustering: Suppose there are data with mixed-scale variables as much as m, where $m_{numeric}$ is the number of purely numerical-scale variables and $m_{categorical}$ is the number of pure variables with a categorical scale, so that: $m = m_{numeric} + m_{categorical}$. Furthermore, data clustering is carried out according to the data type separately. The clustering results are combined using the CEBMDC-ROCK method to obtain the final cluster.

4) CEBMDC-ROCK: The working concept of CEBMDC-ROCK is to cluster objects with a clustering algorithm that fits the data separately, which means that numerical data was grouped with the appropriate algorithm, and the same applies to categorical data. Furthermore, the clustering results are recombined using the ensemble clustering method based on ROCK.

B. Clustering Performance

Measuring the performance of clustering results is a step to determine the validity of a clustering result.

1) Performance of Numerical Data Clustering Results: The process of finding the best by a number of clusters is an important step [39]. This step is called cluster validation. Sum of Square Total (SST).

$$SST_{numeric} = \sum_{l=1}^{m_{numeric}} \sum_{u=1}^{n} (x_{ul} - \bar{x}_l)^2 \quad (7)$$

Sum of Square Within Group (SSW)

$$SSW_{numeric} = \sum_{c=1}^{C} \sum_{c=1}^{m_{numeric}} \sum_{u=1}^{n} (x_{ulc} - \bar{x}_{lc})^2 \quad (8)$$

Sum of Square Between Group (SSB)

$$SSB_{numeric} = SST_{numeric} - SSW_{numeric} \quad (9)$$

The new cluster R-square is the ratio of SSB and SST, meaning that R-square can be defined as a measure of the difference between clusters, with values ranging from 0 to 1.

$$R^2 = \frac{SSB_{numeric}}{SST_{numeric}} = \frac{SST_{numeric} - SSW_{numeric}}{SST_{numeric}} \quad (10)$$

The maximum value of the Pseudo-F can determine the best number of clusters. The formulation of Pseudo-F in Equation 11.

$$\text{Pseudo-F} = \frac{\left(\frac{R^2}{c-1}\right)}{\left(\frac{1-R^2}{n-c}\right)} \quad (11)$$

After getting the best number of clusters, the next step is to determine the best clustering algorithm for numerical data based on the ICD Rate value [40]. The ICD Rate formula is shown in Equation 12.

$$\text{ICD Rate} = 1 - \frac{SSB_{numeric}}{SST_{numeric}} = 1 - R^2 \quad (12)$$

2) Performance of Categorical Data Clustering Results: Performance measures for categorical data have been developed by [39]. If the performance measurement is on the data of n observations, then n_k is an observation with the k-th category where $k = 1,2,\dots,K$ and $\sum_{k=1}^{K} n_k = n$. Furthermore, n_{kc} is the number of observations with the k-th category and the c-group, where $c = 1,2,\dots,C$ is the number of clusters formed.

So that $n_k = \sum_{c=1}^{C} n_{kc}$ is the number of observations in the k-th category. The total number of observations can be written in Equation 13.

$$n = \sum_{c=1}^{C} n_c = \sum_{k=1}^{K} n_k = \sum_{k=1}^{K}\sum_{c=1}^{C} n_{kc} \quad (13)$$

To determine the optimal number of clusters by using the ratio of S_W with S_B. Sum of Square Total (SST)

$$SST_{categorical} = \frac{n}{2} - \frac{1}{2n}\sum_{k=1}^{K} n_k^2 \quad (14)$$

Sum of Square Within Group (SSW)

$$SSW_{categorical} = \frac{n}{2} - \frac{1}{2}\sum_{c=1}^{C}\frac{1}{n_c}\sum_{k=1}^{K} n_{kc}^2 \quad (15)$$

Sum of Square Between Group (SSB)

$$SSB_{categorical} = \frac{1}{2}\left(\sum_{c=1}^{C}\frac{1}{n_c}\sum_{k=1}^{K} n_{kc}^2\right) - \frac{1}{2n}\sum_{k=1}^{K} n_k^2 \quad (16)$$

Mean of Square Total (MST)

$$MST_{categorical} = \frac{SST_{categorical}}{(n-1)} \quad (17)$$

Mean of Square Within (MSW)

$$MSW_{categorical} = \frac{SSW_{categorical}}{(n-C)} \quad (18)$$

Mean of Square Between (MSB)

$$MSB_{categorical} = \frac{SSB_{categorical}}{(C-1)} \quad (19)$$

Standard Deviation in Group (S_W)

$$S_W = \sqrt{MSW_{categorical}} \quad (20)$$

Standard Deviation in Between Group (S_B)

$$S_B = \sqrt{MSB_{categorical}} \quad (21)$$

The performance of a categorical data clustering method is getting better if the ratio value of S_W and S_B is getting smaller.

C. Research Variables

Secondary data were sourced from the *Badan Ketahanan Pangan* (BKP) in 2020. The unit of observation is 34 Provinces in Indonesia. The variables used in this study consisted of 9 variables, which are presented in Table 1.

TABLE I
FOOD SECURITY RESEARCH VARIABLES

Numerical Variables	
Variables	**Description**
X_1	Percentage of the population living below the poverty line
X_2	Percentage of households with a proportion of expenditure on food more than 65%
X_3	Percentage of households without access to electricity
X_4	Percentage of households without access to clean water
X_5	The ratio of the population of health workers to the level of population density
X_6	The average length of schooling for girls is over 15 years
X_7	Open Unemployment Rate

Categorical Variables		
Variables	**Description**	**Category**
X_8	Human Development Index	0: Very High (≥80) 1: High (70 – 79,9) 2: Medium (60 – 69,9) 3: Low (< 60)
X_9	Percentage of toddlers with below standard height (stunting)	0: Above the national stunting prevalence (≥27,67%) 1: Below the national stunting prevalence (< 27,67%)

D. Research Methodology

The steps of analysis using CEBMDC-ROCK are described as follows:

- Divide data by numerical scale and categorical scale.
- Clustering numerical data scale variables using the AHC method with the distance used is the Euclidean distance.
- Calculating the Pseudo-F value and ICD Rate for each clustering algorithm. The optimal cluster is obtained by the largest Pseudo-F value and the smallest ICD Rate value, which is then expressed as the result of numerical grouping data.
- Clustering categorical data scale variables using the ROCK method with a predetermined threshold value (θ) of 0.05; 0.1; 0.2; 0.32; and 0.40.
- Calculating the value of S_W, S_B, and the ratio S_W with S_B for each threshold value (θ).
- Determining the smallest value of the ratio S_W with S_B, which is then expressed as the optimum number of clusters resulting from categorical grouping data.
- Combining the grouping results based on the output of the numerical and categorical data grouping using the ROCK method. The threshold value (θ) used is the same, namely 0.05; 0.1; 0.2; 0.32; and 0.40.
- Calculating the value of the ratio S_W with S_B for each combined cluster threshold value (θ). The smallest value of the ratio of S_W with S_B is the optimum group for mixed data, hereinafter referred to as the final cluster.

Based on point a-h, the stages of analysis in this study are visualized in Fig. 1.

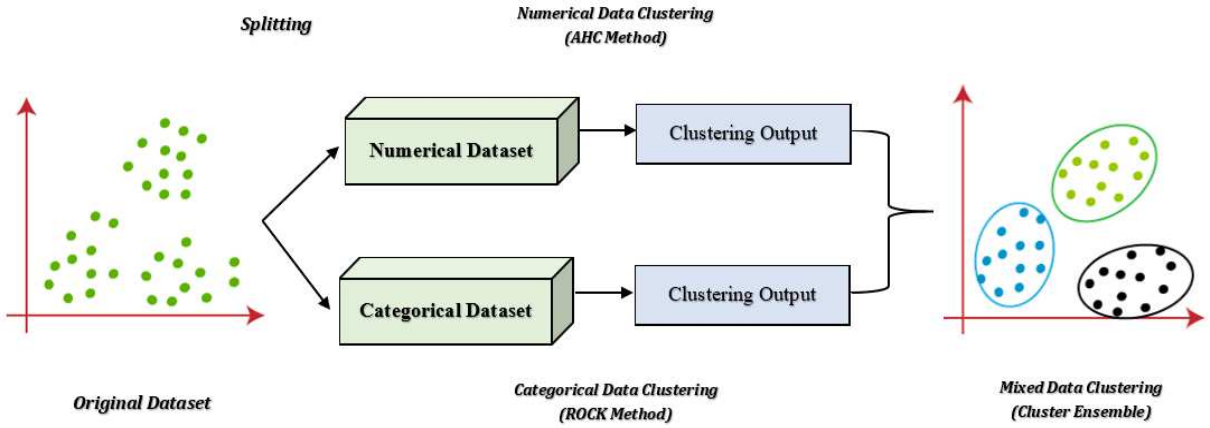

Fig. 1 Framework by CEBMDC-ROCK

III. Results and Discussion

Province clustering in Indonesia based on food security conditions begins with numerical clustering data, continues with categorical data, and ends with mixed data clustering.

A. Numerical Data Clustering

The variables used for numerical clustering data are variables $X_1 - X_7$ according to Table 1. The results of clustering variables with numerical data scale using the AHC method with several algorithms tried are shown in Table 2.

TABLE II
NUMERICAL DATA CLUSTERING RESULTS

Clustering Algorithm	Number of Clusters	Pseudo-F	ICD Rate
Single Linkage	**2**	**10.37**	**0.76**
	3	8.23	0.65
	4	8.53	0.54
Complete Linkage	**2**	**17.08**	**0.65**
	3	15.54	0.50
	4	13.31	0.43
Average Linkage	**2**	**10.37**	**0.76**
	3	8.23	0.65
	4	10.14	0.49

Based on Table 2, the largest Pseudo-F value in each clustering algorithm produces two optimal clusters. Furthermore, from each clustering algorithm used, it is found that the complete linkage algorithm is the best in clustering Provinces in Indonesia based on food security conditions, which is obtained from the smallest ICD rate value. The results of clustering Provinces in Indonesia for numerical data are divided into two clusters. Each group member contains 9 Provinces in Cluster 1 and 25 Provinces in Cluster 2, hereinafter referred to as the results of numerical grouping data.

B. Categorical Data Clustering

Clustering conditions of Indonesian food security based on categorical data using the ROCK method using variables $X_8 - X_9$ according to Table 1. Grouping with the ROCK method uses a threshold value (θ) set at the beginning, 0.05; 0.1; 0.2; 0.32; and 0.40. The threshold value (θ) is considered to have a large enough difference in the value of the ratio of S_W with S_B so that it is considered capable of providing a significant difference.

TABLE III
CATEGORICAL DATA CLUSTERING RESULTS

θ	Number of Clusters	S_W	S_B	The ratio of S_W with S_B
0.05	2	0.37	4.63	0.08
0.10	2	1.91×10^{-15}	5.82	3.29×10^{-16}
0.20	3	0.76	7.35	0.10
0.32	2	0.75	10.57	0.07
0.40	**4**	$\mathbf{3.45\times 10^{-16}}$	**5.15**	$\mathbf{6.69\times 10^{-17}}$

The information obtained from Table 3 shows that the threshold value (θ) that produces the smallest ratio S_W with S_B is $\theta = 0.4$, which has four clusters based on the condition of food security in Indonesia. The results of clustering Provinces in Indonesia for categorical data are divided into four clusters. The results are then referred to as categorical groups.

C. Mixed Data Clustering

After obtaining the results of the optimal grouping of numerical and categorical data, the next step is to combine the clustering results. The threshold value (θ) is 0.05; 0.1; 0.2; 0.32; and 0.40. And then the results of clustering are presented in Table 4.

TABLE IV
MIXED DATA CLUSTERING RESULTS

θ	Number of Clusters	S_W	S_B	The ratio of S_W with S_B
0.05	3	0.71	4.26	0.17
0.10	3	0.71	4.26	0.17
0.20	4	0.63	3.83	0.16
0.32	4	0.82	3.64	0.22
0.40	**5**	$\mathbf{5.80\times 10^{-16}}$	**8.47**	$\mathbf{6.85\times 10^{-17}}$

Table 4 provides information on what produces the smallest S_W to S_B ratio value is $\theta = 0.4$, with the number of clusters is five. We can conclude that the food security condition in Indonesia is divided into five clusters. Furthermore, Table 5 shows the details of provincial members in each cluster.

TABLE V
PROVINCE DETAILS OF EACH CLUSTER

Cluster	Province
1	Sulawesi Selatan, Kalimantan Selatan, Sumatera Utara, Kalimantan Tengah, Kalimantan Timur, Jawa Tengah, dan Sulawesi Tenggara
2	Kep. Bangka Belitung, Sumatera Barat, Riau, Kep. Riau, Jawa Barat, Jawa Timur, Bali, Kalimantan Utara, Sulawesi Utara, Yogyakarta, dan Banten
3	Maluku Utara, Gorontalo, Maluku, Nusa Tenggara Barat, dan Sulawesi Tengah
4	Papua, Nusa Tenggara Timur, Sulawesi Barat, dan Kalimantan Barat
5	DKI Jakarta, Aceh, Sumatera Selatan, Lampung, Papua Barat, Jambi, dan Bengkulu

In Cluster 1, there are seven provinces. Cluster 2 there are eleven provinces. Cluster 3 there are five provinces. In cluster 4, there are four provinces; in Cluster 5, there are seven. Cluster 1 is a group of provinces with good quality food security. It has a low percentage of poor people, a percentage of households without access to electricity, and a percentage of households without access to clean water, as well as the high HDI category. Meanwhile, it has a proportion of food expenditure of more than 65 percent, the lowest among other groups, and the ratio of health workers was down. The prevalence of stunting under five was above 27.67%.

Cluster 2 is a province group with very good quality food security. Because it has the lowest percentage of poor people, percentage of households without access to electricity, and percentage of households without access to clean water, as well as with a high ratio of health workers, the HDI group, which is included in the high category, and the prevalence of stunting under five is below 27.67%. But with a moderate proportion of food expenditure of more than 65%.

Cluster 3 is a group of provinces with low-quality food security. It has a high percentage of poor people and a percentage of households without access to electricity, a high percentage of households without access to clean water, an average ratio of health workers and the HDI category, and a prevalence of stunting under five children over 27. 67%, but has a low proportion of food expenditure of more than 65 percent.

Cluster 4 is a province group with very low-quality food security. It has a very high percentage of poor people, a percentage of households without access to electricity, a very high percentage of households without access to clean water, the lowest ratio of health workers among other groups, with the moderate HDI category, and the prevalence of stunting under five is above 27.67%. Still, it has a proportion of more than 65 percent of food expenditure, which is the highest among other groups.

Meanwhile, cluster 5 is a province group with moderate quality food security. It has an average percentage of poor people and households without access to electricity, households without access to clean water, and a high HDI category. But as a proportion of more than 65 percent of food expenditure is high, the ratio of health workers is the highest among other groups, and the prevalence of stunting is above 27.67%.

IV. CONCLUSION

The Cluster Ensemble Based Mixed Data Clustering-Robust Clustering Using Links (CEBMDC-ROCK) method is used to cluster Provinces in Indonesia based on food security conditions has been successfully carried out. Based on the analysis, five optimal clusters were obtained with the threshold value $\theta = 0.4$. Furthermore, the provinces in Indonesia are distributed into five clusters, where Cluster 1 is a group of Provinces with good quality food security. Cluster 2 with excellent food security quality. Cluster 3 is a group of provinces with low-quality food security. Cluster 4 is a cluster with a very low quality of food security. Moreover, lastly, Cluster 5 is a group of provinces with moderate-quality of food security. Thus, the strategy and new policies can be updated by considering the five clusters to make a balanced food security system in Indonesia.

ACKNOWLEDGMENT

The authors gratefully acknowledge financial support from the Department of Statistics, Faculty of Science and Data Analytics (FSAD), Sepuluh Nopember Institute of Technology, Surabaya, Indonesia.

REFERENCES

[1] R. Hakim, T. Haryanto, and D. W. Sari, "Technical efficiency among agricultural households and determinants of food security in East Java, Indonesia," *Sci. Rep.*, vol. 11, no. 1, pp. 1–9, 2021, doi: 10.1038/s41598-021-83670-7.

[2] M. F. F. Mardianto *et al.*, "Classification of Food Menu and Grouping of Food Potential To Support the Food Security and Nutrition Quality," *Commun. Math. Biol. Neurosci.*, vol. 2022, pp. 1–31, 2022, doi: 10.28919/cmbn/6801.

[3] V. Trivellone, E. P. Hoberg, W. A. Boeger, and D. R. Brooks, "Food security and emerging infectious disease: Risk assessment and risk management," *R. Soc. Open Sci.*, vol. 9, no. 2, 2022, doi: 10.1098/rsos.211687.

[4] Z. G. Dessie, T. Zewotir, and D. North, "The spatial modification effect of predictors on household level food insecurity in Ethiopia," *Sci. Rep.*, vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-23918-y.

[5] L. Bizikova, S. Jungcurt, K. McDougal, and S. Tyler, "How can agricultural interventions enhance contribution to food security and SDG 2.1?," *Glob. Food Sec.*, vol. 26, no. October, p. 100450, 2020, doi: 10.1016/j.gfs.2020.100450.

[6] M. Akbari *et al.*, "The Evolution of Food Security: Where Are We Now, Where Should We Go Next?," *Sustainability*, vol. 14, no. 6, pp. 1–27, 2022, doi: 10.3390/su14063634.

[7] K. Pawlak and M. Kołodziejczak, "The role of agriculture in ensuring food security in developing countries: Considerations in the context of the problem of sustainable food production," *Sustain.*, vol. 12, no. 13, 2020, doi: 10.3390/su12135488.

[8] H. Y. S. H. Nugroho *et al.*, "Toward Water, Energy, and Food Security in Rural Indonesia: A Review," *Water (Switzerland)*, vol. 14, no. 10, pp. 1–25, 2022, doi: 10.3390/w14101645.

[9] M. Campi, M. Dueñas, and G. Fagiolo, "Specialization in food production affects global food security and food systems sustainability," *World Dev.*, vol. 141, p. 105411, 2021, doi: 10.1016/j.worlddev.2021.105411.

[10] M. H. Montolalu, M. Ekananda, T. Dartanto, D. Widyawati, and M. Panennungi, "The Analysis of Trade Liberalization and Nutrition Intake for Improving Food Security across Districts in Indonesia," *Sustainability*, vol. 14, no. 6, 2022, doi: 10.3390/su14063291.

[11] H. Dharmawan, B. Sartono, A. Kurnia, A. F. Hadi, and E. Ramadhani, "A Study of Machine Learning Algorithms to Measure the Feature Importance In Class-Imbalance Data of Food Insecurity Cases in Indonesia," *Commun. Math. Biol. Neurosci.*, vol. 2022, pp. 1–25, 2022.

[12] A. Radovanovic, J. Li, J. V. Milanovic, N. Milosavljevic, and R. Storchi, "Application of agglomerative hierarchical clustering for clustering of time series data," *IEEE PES Innov. Smart Grid Technol. Conf. Eur.*, vol. 2020-Octob, pp. 640–644, 2020, doi: 10.1109/ISGT-Europe47291.2020.9248759.

[13] H. Nouraei, H. Nouraei, and S. W. Rabkin, "Comparison of Unsupervised Machine Learning Approaches for Cluster Analysis to

Define Subgroups of Heart Failure with Preserved Ejection Fraction with Different Outcomes," *Bioengineering*, vol. 9, no. 4, 2022, doi: 10.3390/bioengineering9040175.

[14] A. M. Jabbar, K. R. Ku-Mahamud, and R. Sagban, "Improved Self-Adaptive ACS Algorithm to Determine the Optimal Number of Clusters," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 11, no. 3, pp. 1092–1099, 2021, doi: 10.18517/ijaseit.11.3.11723.

[15] A. Munawar *et al.*, "Cluster Application with K-Means Algorithm on the Population of Trade and Accommodation Facilities in Indonesia," *J. Phys. Conf. Ser.*, vol. 1933, no. 1, 2021, doi: 10.1088/1742-6596/1933/1/012027.

[16] S. Sarumathi, P. Ranjetha, C. Saraswathy, M. Vaishnavi, and S. Geetha, "A Review and Comparative Analysis on Cluster Ensemble Methods," *Int. J. Comput. Inf. Eng.*, vol. 15, no. 6, pp. 385–394, 2021.

[17] J. Park, K. V. Park, S. Yoo, S. O. Choi, and S. W. Han, "Development of the WEEE grouping system in South Korea using the hierarchical and non-hierarchical clustering algorithms," *Resour. Conserv. Recycl.*, vol. 161, no. March 2020, p. 104884, 2020, doi: 10.1016/j.resconrec.2020.104884.

[18] W. B. Xie, Y. L. Lee, C. Wang, D. B. Chen, and T. Zhou, "Hierarchical clustering supported by reciprocal nearest neighbors," *Inf. Sci. (Ny).*, vol. 527, pp. 279–292, 2020, doi: 10.1016/j.ins.2020.04.016.

[19] L. Ramos Emmendorfer and A. M. de Paula Canuto, "A generalized average linkage criterion for Hierarchical Agglomerative Clustering," *Appl. Soft Comput.*, vol. 100, p. 106990, 2021, doi: 10.1016/j.asoc.2020.106990.

[20] E. Banjarnahor, A. Bustamam, T. Siswantining, and P. Tampubolon, "Analyzing Kinship in Severe Acute Respiratory Syndrome Coronavirus 2 DNA Sequences Based on Hierarchical and K-Means Clustering Methods Using Multiple Encoding Vector," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 12, no. 6, pp. 2237–2247, 2022, doi: 10.18517/ijaseit.12.6.15582.

[21] A. Sarica, M. G. Vaccaro, A. Quattrone, and A. Quattrone, "A novel approach for cognitive clustering of parkinsonisms through affinity propagation," *Algorithms*, vol. 14, no. 2, 2021, doi: 10.3390/a14020049.

[22] P. Vilas, L. Andreu, and J. L. Sarto, "Cluster analysis to validate the sustainability label of stock indices: An analysis of the inclusion and exclusion processes in terms of size and ESG ratings," *J. Clean. Prod.*, vol. 330, 2022, doi: 10.1016/j.jclepro.2021.129862.

[23] A. Dogan and D. Birant, "K-centroid link: a novel hierarchical clustering linkage method," *Appl. Intell.*, vol. 52, no. 5, pp. 5537–5560, 2022, doi: 10.1007/s10489-021-02624-8.

[24] F. Ros and S. Guillaume, "A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise," *Expert Syst. Appl.*, vol. 128, pp. 96–108, 2019, doi: 10.1016/j.eswa.2019.03.031.

[25] J. Senthilnath, P. B. Shreyas, R. Rajendra, S. Suresh, S. Kulkarni, and J. A. Benediktsson, "Hierarchical clustering approaches for flood assessment using multi-sensor satellite images," *Int. J. Image Data Fusion*, vol. 10, no. 1, pp. 28–44, 2019, doi: 10.1080/19479832.2018.1513956.

[26] M. Charikar, V. Chatziafratis, and R. Niazadeh, "Hierarchical clustering better than average-linkage," *Proc. Annu. ACM-SIAM Symp. Discret. Algorithms*, pp. 2291–2304, 2019, doi: 10.1137/1.9781611975482.139.

[27] R. Wang and A. Sun, "Research on user clustering algorithm based on improved rock algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 790, no. 1, 2020, doi: 10.1088/1757-899X/790/1/012065.

[28] H. Sofyan, M. Iqbal, M. Marzuki, and M. Muhammad, "The comparison of k-modes clustering and ROCK clustering to the poverty indicator in Samadua Subdistrict, South Aceh," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1087, no. 1, p. 012085, 2021, doi: 10.1088/1757-899x/1087/1/012085.

[29] R. Nooraeni, M. I. Arsa, and N. W. Kusumo Projo, "Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 677–684, 2021, doi: 10.1016/j.procs.2021.01.055.

[30] S. Yin, G. Gan, E. A. Valdez, and J. Vadiveloo, "Applications of clustering with mixed type data in life insurance," *Risks*, vol. 9, no. 3, pp. 1–19, 2021, doi: 10.3390/risks9030047.

[31] N. Yuvaraj and C. Suresh Ghana Dhas, "High-performance link-based cluster ensemble approach for categorical data clustering," *J. Supercomput.*, vol. 76, no. 6, pp. 4556–4579, 2020, doi: 10.1007/s11227-018-2526-z.

[32] Z. He, X. Xu, and S. Deng, "A cluster ensemble method for clustering categorical data," *Inf. Fusion*, vol. 6, no. 2, pp. 143–151, 2005, doi: 10.1016/j.inffus.2004.03.001.

[33] L. Wulandari, Y. Farida, A. Fanani, N. Ulinnuha, and P. K. Intan, "Evaluation of disadvantaged regions in east java based-on the 33 indicators of the ministry of villages, development of disadvantaged regions, and transmigration using the ensemble ROCK (Robust clustering using link) method," *Adv. Sci. Technol. Eng. Syst.*, vol. 5, no. 5, pp. 193–200, 2020, doi: 10.25046/aj050524.

[34] G. Caruso, S. A. Gattone, A. Balzanella, and T. Di Battista, *Cluster Analysis: An Application to a Real Mixed-Type Data Set*, vol. 179. Springer International Publishing, 2019. doi: 10.1007/978-3-030-00084-4_27.

[35] G. Caruso, S. A. Gattone, F. Fortuna, and T. Di Battista, "Cluster Analysis for mixed data: An application to credit risk evaluation," *Socioecon. Plann. Sci.*, vol. 73, no. February, p. 100850, 2021, doi: 10.1016/j.seps.2020.100850.

[36] G. Caruso and S. A. Gattone, "Waste management analysis in developing countries through unsupervised classification of mixed data," *Soc. Sci.*, vol. 8, no. 6, 2019, doi: 10.3390/socsci8060186.

[37] Vijaya, S. Aayushi, and R. Bateja, "A Review on Hierarchical Clustering Algorithms," *Journal of Engineering and Applied Sciences*, vol. 12, no. 24. pp. 7501–7507, 2017.

[38] R. S. Pontoh, F. Salsabila, F. C. Garini, R. A. Fatharani, S. Zahroh, and E. Supartini, "Clustering of fishery management areas based on the level of utilization in Indonesia," *Commun. Math. Biol. Neurosci.*, vol. 2021, pp. 1–19, 2021, doi: 10.28919/cmbn/6171.

[39] S. Zhou, F. Liu, and W. Song, "Estimating the Optimal Number of Clusters Via Internal Validity Index," *Neural Process. Lett.*, vol. 53, no. 2, pp. 1013–1034, 2021, doi: 10.1007/s11063-021-10427-8.

[40] F. A. Syaani, Irhamah, A. Mukarromah, and K. Fithriasari, "Incident Clustering in the Warehouse Workspaces by Using Text Mining," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1117, no. 1, p. 012023, 2021, doi: 10.1088/1757-899x/1117/1/012023.