BAREKENG: Journal of Mathematics and Its Applications
December 2024 Volume 18 Issue 4 Page 2231–2246
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss4pp2231-2246

IMPLEMENTATION OF K-MEANS AND FUZZY C-MEANS CLUSTERING FOR MAPPING TODDLER STUNTING CASES IN GUNUNGKIDUL DISTRICT

Bintang Wira Mahardika[1*], **Agus Maman Abadi**[2]

[1,2]*Department of Mathematics Education, Faculty of Mathematics and Natural Sciences,*
Universitas Negeri Yogyakarta
Jln. Colombo No.1, Yogyakarta, 55281, Indonesia

*Corresponding author's e-mail: ** **bintangwiram@gmail.com**

ABSTRACT

Article History:

Received: 19th January 2024
Revised: 15th May 2024
Accepted: 11th July 2024
Published: 14th October 2024

Keywords:

Fuzzy c-means;
K-means;
Gunungkidul Regency;
Mapping;
Toddler Stunting

Gunungkidul Regency has the highest prevalence of stunted toddlers in the Special Region of Yogyakarta. This study aims to describe the optimal clustering results of toddler stunting cases using the k-means and fuzzy c-means methods and to describe the characteristic of the mapping results of stunting-prone areas for toddlers in Gunungkidul Regency for the years 2020 – 2022. This study maps stunting-prone areas for toddlers across 30 community health centers in Gunungkidul Regency from 2020 to 2022, with variables including the percentage of babies with low birth weight, babies born stunted, babies receiving health services, stunted toddlers, toddlers receiving health services, babies given exclusive breastfeeding, poor couples of reproductive ages, and families with adequate drinking water. The k-means clustering method determines cluster membership using the distance between objects and centroids, while the fuzzy c-means method uses the degree of membership. Cluster evaluation uses the silhouette coefficient, Calinski-Harabasz index, Davies-Bouldin index, and Dunn index to obtain optimal clustering results. The mapping results are presented as a stunting vulnerability map. The findings indicate that the optimal number of clusters is two, with the fuzzy c-means method proving more optimal than the k-means method based on evaluation scores. In 2020, there were 23 community health centers in cluster 0 and 7 in cluster 1. In 2021, there were 21 community health centers in cluster 0 and 9 in cluster 1. In 2022, there were 18 community health centers in cluster 0 and 12 in cluster 1. Generally, community health centers in cluster 0 are less optimal in specific nutrition interventions, such as for infants and toddlers. In contrast, those in cluster 1 are less optimal in sensitive nutrition interventions, such as poverty and water adequacy.

How to cite this article:

B. W. Mahardika and A. M. Abadi., " IMPLEMENTATION OF K-MEANS AND FUZZY C-MEANS CLUSTERING FOR MAPPING TODDLER STUNTING CASES IN GUNUNGKIDUL DISTRICT," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2231-2246, December, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com;* barekeng.journal@mail.unpatti.ac.id

Research Article · Open Access

1. INTRODUCTION

Stunting in children under five is defined as children aged 0 to 59 months who have a Z-score of less than -2 standard deviations based on height-for-age or length-for-age indicators [1]. On a global average, height growth slows with age, such as after birth to the third year of life [2]. Stunted children are characterized as physically short or stunted compared to other children their age due to chronic malnutrition [3][4][5][6]. This malnutrition occurs from pregnancy to after birth, which is commonly referred to as the First 1,000 Days of Life (HPK) [7][8][9]. Meanwhile, stunting also has a long-term impact, affecting brain development which impacts future thinking, productivity, creativity, and immunity [9][10][11][12][13]. In addition, the child will be susceptible to degenerative diseases when they grow up, such as diabetes, heart disease, cancer, kidney disease, and other non-communicable diseases [14]. This causes stunted child development, which can affect the quality of human resources in the future.

Based on the results of the 2022 Indonesian Nutrition Status Survey (SSGI), the prevalence of stunting in Indonesia is 21.6%, while Gunungkidul Regency has a prevalence of 23.5%. The prevalence of stunting is still relatively high because it is still above the World Health Organization (WHO) standard, which is 20%, so it can be a big problem and cause concern [15]. The high prevalence of stunting is caused by direct causes, such as food consumption and infection status, as well as indirect causes, such as food security, social environment, health environment, and residential environment, where these two causes are influenced by various factors, one of which is economic conditions [8][16][17][18]. The Ministry of Health of the Republic of Indonesia seeks to reduce the prevalence of stunting with integrated nutrition interventions, which include nutrition-specific and nutrition-sensitive interventions [19]. Various indicators show the success of stunting interventions for children under five, some of which include the number of low birth weight babies, stunted babies, and toddlers, exclusive breastfeeding in infants, infant and toddler health services, poor childbearing age couples, and drinking water eligibility that can show the success rate of nutrition interventions.

Analysis and mapping in reducing the prevalence of stunting is necessary for effective and targeted interventions [20][21]. Technological advances in computing and data processing programs can be used to cluster stunting cases [22]. Clustering can make it easier to see the characteristics of the factors that cause stunting in an area [23]. Based on this, various indicators of stunting intervention can be used as research material in the process of mapping stunting-prone areas in an area.

Clustering is a process of grouping several objects into several clusters so that a cluster contains a collection of objects that have similar characteristics and is different from the collection of objects in other clusters [24][25]. Clustering can be done with various methods, some of which are fuzzy c-means and k-means. Fuzzy c-means is a grouping of objects based on the degree of membership (between 0 and 1) in each cluster [26]. Meanwhile, k-means is one of the clustering algorithms that serve to partition objects into one or more clusters without knowing in advance the target class [27]. The k-means and fuzzy c-means methods are sensitive to outliers or the value of an object that is very different from the data set [28]. The data used is made to have the same and comparable range to prevent outliers so that the clustering results are more optimal. The k-means and fuzzy c-means methods have the result or output of this clustering in the form of group data. These results can be mapped to clarify the visualization of areas or areas prone to stunting so that interventions are more optimal [29].

There is research in the field of health, especially toddler nutrition, by comparing the two methods. The research was conducted for clustering health centers based on toddler nutrition in Surabaya, which used a cluster evaluation silhouette coefficient worth 0.518 with the k-means method and 0.497 with the fuzzy c-means method, so in the study, the k-means algorithm was better because it had a slightly larger silhouette coefficient value [30]. In addition, many other studies in the health sector use the k-means or fuzzy c-means method with fairly optimal results, such as in the following studies [2][3][5][9][10][22][25].

Based on the description, this research aims to find out which clustering method is better between k-means and fuzzy c-means based on evaluation values and is used to map areas prone to toddler stunting cases in Gunungkidul Regency in 2020-2022. The research was conducted with the assistance of Python programming and QGIS software for mapping the clustering results. The mapping was carried out because no research showed the mapping of areas prone to stunting cases in Gunungkidul.

2. RESEARCH METHODS

2.1 Data Description

The data for this study were obtained from the Family Health Sector (Kesga) DIY website [31], the Gunungkidul Regency Health Office, the Gunungkidul Regency Regional Development Planning Agency, and the Gunungkidul Land and Spatial Planning Office. The data underwent preprocessing to form a one dataset with 90 data objects, each indexed by community health centre and year, with 8 variables namely:

a. Percentage of babies with low birth weight (X_1)

Birth weight is a fairly good indicator in determining the overall nutritional status of a baby and its well-being [32]. Low birth weight (LBW) babies are babies born with a weight of less than 2.5 kilograms due to short gestational age and/or stunted fetus growth, both of which are influenced by factors. risks, such as maternal, placental, fetus, and environmental factors [33]. This is caused by inadequate nutrition for the fetus during pregnancy.

b. Percentage of babies born stunted (X_2)

Stunting occurs when a child's height or body length is shorter compared to children his age. The body length index according to age or height according to age of these toddlers has a Z-score of less than -2 standard deviations (Std) [32]. The high number of stunted babies born shows that the interventions carried out have not been optimal.

c. Percentage of babies receiving health services (X_3)

Health services provide treatment for various health problems, one of which is related to malnutrition. This service can take the form of providing medication or certain products that are used to treat children experiencing malnutrition with various nutritional content that the body needs, such as vitamins, minerals, and protein, with steps starting from treatment of acute complications, healing, and recovery [32]. The quality of health services also depends on the quality of health workers. Poor health care for babies can be a factor in stunting. Thus, optimizing infant health services is in line with the aim of reducing the prevalence of stunting.

d. Percentage of stunted toddlers (X_4)

Just like stunted babies born, the high number of stunted toddlers shows that the interventions carried out have not been optimal.

e. Percentage of toddlers receiving health services (X_5)

Just like for babies, poor health services for toddlers can be a factor in stunting. Thus, optimizing toddler health services is in line with the aim of reducing the prevalence of stunting.

f. Percentage of babies given exclusive breastfeeding (X_6)

Breast milk is given exclusively to babies so that they can develop well. WHO recommends exclusive breastfeeding during the first six months of life so that it can provide various important substances for the body, for example lactose which is the main source of carbohydrates [32].

g. Percentage of poor couples of reproductive ages (X_7)

According to the Regulation of the National Population and Family Planning Agency of the Republic of Indonesia Number 1 of 2023, couples of childbearing ages are married couples whose wives are aged 15 – 49 years and are still menstruation, or a married couple whose wife is less than 15 years old, but is already menstruating [34]. Women with low welfare likely to marry before the age of 18 are four times more than women with high welfare, so these couples tend to remain poor [35].

h. Percentage of families with adequate drinking water (X_8).

Drinking water that is pure (not sea water or salt) and safe plays an important role in public health, one of which is in the food and beverage production process [32]. The suitability of drinking water from a house is closely related to the suitability of sanitation.

This dataset was used in the clustering process and served as a reference for the stunting vulnerability level of a community health centre area.

2.2 Theoretical Review

a. K-means Clustering

K-means is a clustering algorithm that functions to partition existing data into one or more clusters without first knowing the target class [27]. The following is the algorithm for k-means clustering [36][37]:

1) Determine the desired number of clusters, then calculate the average of each cluster. The initial selection of centroids is done randomly.
2) Calculate the distance between objects from each cluster with the following Euclidean equation.

$$d(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \qquad (1)$$

Where:
$d(x,y)$: distance between x and y
x_i : centroid at the $i-$th variable
y_i : data at the $i-$th variable
d : the number of dimensions (variables) of the data
i : index of variable
Then, proceed to assign each point to its nearest cluster.

3) Repeat the previous two steps when the last centroid value is equal to the previous centroid value.

b. Fuzzy C-Means Clustering

Fuzzy c-means is a fuzzy clustering algorithm which is a development of the k-means method. This method, which was first introduced by Dunn in 1973 and refined by Bezdek in 1981, will group objects with each object having a degree of membership (between 0 and 1) with each centroid of the cluster [26]. The steps in the fuzzy c-means algorithm are as follows [38], [39], [40]:

1) Determine the number of clusters $(\boldsymbol{c})$, power $(\boldsymbol{m})$, smallest error $(\boldsymbol{\xi})$, the initial objective function $(\boldsymbol{J_0})$, initial iteration $(\boldsymbol{t=0})$, and the maximum iteration (**MaxIter**).
2) Initiate the initial membership matrix with a random number $\boldsymbol{\mu_{ki}}$.
3) Calculate the $\boldsymbol{j-th}$ centroid, $\boldsymbol{V_{il}}$ with $\boldsymbol{j=1,2,\ldots,c}$ and $\boldsymbol{i=1,2,\ldots,d}$ as follows.

$$\boldsymbol{C_{ji} = \frac{\sum_{k=1}^{n}(\mu_{kj})^m x_{ki}}{\sum_{k=1}^{n}(\mu_{kj})^m}} \qquad (2)$$

Where:
C_{ji} : the $j-$th centroid for the $i-$th variable
μ_{kj} : membership degree of x_k in cluster-j
m : power
x_{ki} : the $k-$th data (x_k) on the $i-$th variable

4) Calculate the objective function at the t-th iteration, namely J_t, using the following equation.

$$J_{FCM} = \sum_{k=1}^{n}\sum_{j=0}^{c}(\mu_{kj})^m d^2(x_k, C_j) \qquad (3)$$

Where:
n : amount of data
c : number of clusters with value $2 \leq c < n$
m : power with value $m > 1$ (in general $1 < m < 3$)
$d^2(x_k, C_j)$: distance value between objects x_k with the centroid C_j
μ_{kj} : degree of membership of x_k in the $j-$th cluster with the equation

$$\mu_{kj} = \frac{\left(\frac{1}{d_{kj}}\right)^{\frac{2}{(m-1)}}}{\sum_{l=1}^{c}\left(\frac{1}{d_{kl}}\right)^{\frac{2}{(m-1)}}} \tag{4}$$

Where:

d_{kj} : Euclidean distance between x_k with the $j-$th centroid

d_{kl} : Euclidean distance between x_k with each centroid l (for all cluster, from 1 to c)

5) Iteration stops if $|J_{t+1} - J_t| < \xi$ or $t >$ MaxIter. Otherwise, iteration continues with $t = t + 1$, and restarts from step 4.

c. Cluster Evaluation

Cluster evaluation is carried out to show the feasibility of the clustering results of a method using a certain value. There are several evaluation measures used in this research, namely as follows [28][40][41].

1) Silhouette coefficient (SC)
SC is formulated by considering the distance of each object within the cluster and between clusters [28]. The following is the equation for determining SC.

$$\text{SC} = \frac{\sum_{k=1}^{n}\frac{b_k - a_k}{\max(a_k, b_k)}}{n} \tag{5}$$

Where:

a_k : the average distance of a particular object (M_i) to all objects in the same cluster

b_k : average object distance M_i to all objects in each cluster (other than the cluster containing the point M_i)

n : amount of data

2) Calinski-Harabasz index (CHI)
CHI is based on the relationship between the sum of the squares of the distance between the center of each cluster and the centroid of the data set and the sum of the squares of the distance between the center of each cluster and each point in the cluster [42]. The following is the equation for determining CHI.

$$\text{CHI} = \frac{BGSS}{WGSS} \cdot \frac{n-c}{c-1} \tag{6}$$

Where:

n : amount of data

c : number of clusters

$BGSS$: sum of squares of the distance between the center of each cluster and the centroid of the whole data

$WGSS$: sum of squared distances between the center of each cluster and each point in the cluster

3) Davies-Bouldin index (DBI)
DBI compares each cluster based on a function that measures the similarity of each pair of clusters, in the form of the average distance value of each point in the two clusters to each centroid [42]. The following is the equation for determining DBI.

$$\text{DBI} = \frac{1}{c}\sum_{j=1}^{c} R_j \tag{7}$$

Where:

c : number of clusters

R_j : cluster similarity size (maximum)

4) Dunn index (DI)

DI provides an evaluation value based on the square root of the minimum distance between two clusters (to measure differences between clusters) divided by the square root of the maximum distance between two points in a cluster (to measure the similarity between members of a cluster) [42]. The following is the equation to determine DI.

$$\mathrm{DI} = \frac{\min\limits_{j'\neq j}\left\{\min\limits_{\substack{k\epsilon I_j \\ k'\epsilon I_{j'}}}\left\{\left\|M_k^{\{j\}} - M_{k'}^{\{j'\}}\right\|\right\}\right\}}{\max\limits_{1\leq j\leq c}\left\{\max\limits_{\substack{k,k''\epsilon I_j \\ k\neq k''}}\left\{\left\|M_k^{\{j\}} - M_{k''}^{\{j\}}\right\|\right\}\right\}} \tag{8}$$

Where:

$M_k^{\{j\}}$: the $k-th$ data in the cluster j

$M_{k'}^{\{j'\}}$: the $k'-th$ data in the cluster j'

$M_{k''}^{\{j\}}$: the $k''-th$ data in the cluster j

c : number of clusters

Evaluation value rule [41] are shown in Table 1 below:

Table 1. Evaluation measure rules

Evaluation Measure	Rule
SC	Closest to 1
CHI	Maximum
DBI	Minimum
DI	Maximum

***Data source**: Desgraupes, 2017*

2.3 Research Steps

This research discusses the mapping of stunting-prone areas for toddlers at each community health center (Puskesmas) level in Gunungkidul Regency from 2020 to 2022, based on optimal clustering results using the k-means and fuzzy c-means methods. The study was conducted with the following steps:

a. Data preprocessing

All variables in the data are adjusted to a percentage to avoid noise and outliers to get optimal clustering results. Data preprocessing goes through the following steps:

1) Data preparation

Adjust the variables for the number of families with adequate drinking water and the number of families for each sub-district area into each community health center area based on the list of sub-districts covering the community health center area so that all data has the same index and amount of data. After that, all variables can be combined into a dataset with same object data and index.

2) Variable modification

Adjusting all variables in the form of numbers into percentages, such as the percentage of babies given exclusive breastfeeding based on the number of babies, the percentage of poor couples of childbearing ages based on the number of couples of childbearing ages, and the percentage of families with adequate drinking water based on the number of families.

3) Multicollinearity test

This stage checks the correlation values between variables with Pearson correlation to ensure that there is no multicollinearity in the data. The following is equation of the Pearson correlation [43].

$$r = \frac{n\sum_{k=1}^{n} X_k Y_k - \left(\sum_{k=1}^{n} X_k\right)\left(\sum_{k=1}^{n} Y_k\right)}{\sqrt{\left(n\sum_{k=1}^{n} X_k^2 - \left(\sum_{k=1}^{n} X_k\right)^2\right)\left(n\sum_{k=1}^{n} Y_k^2 - \left(\sum_{k=1}^{n} Y_k\right)^2\right)}} \tag{9}$$

Where:

n	: number of objects
X_k	: the $k-$th data of X variable
Y_k	: the $k-$th data of Y variable

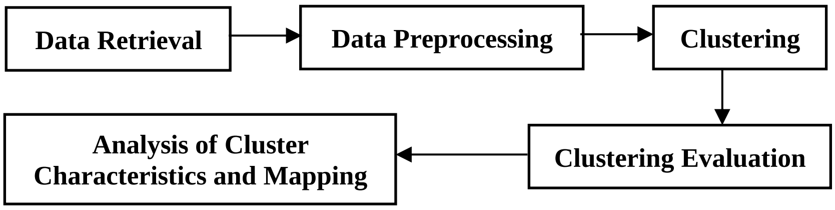
b. Clustering using k-means and fuzzy c-means

c. Clustering evaluation using SC, CHI, DBI, and DI

d. Analysis of cluster characteristics and mapping of each year from the clustering results with the optimal method and number of clusters using QGIS 3.28.8 software

The following is a diagram of the research steps in mapping areas prone to toddler stunting cases.

Figure 1. Diagram of research steps



3. RESULTS AND DISCUSSION

3.1 Results

The research data was adjusted at the preprocessing stage to obtain optimal results. In the data preparation, adjustments were made to the attributes of the **number of families with proper drinking water (NFPDW)** and the **number of families (NF)** which had a total of 144 objects (sub-district level) to 30 (health center level) to fit the dimensions of other attributes, so that Table 2 below is obtained:

Table 2. Suitability of drinking water

Health Center	2020		2021		2022	
	NFPDW	NF	NFPDW	NF	NFPDW	NF
Nglipar I	4173	4842	4242	4866	4242	4866
Nglipar II	5303	6327	5783	6346	5783	6346
Gedangsari I	5535	6898	5871	6828	5871	6828
Gedangsari II	4891	7031	5927	7037	5927	7037
Patuk I	5272	6107	5583	6111	5583	6111
Patuk II	3655	5135	4390	5188	4390	5188
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Semin II	6970	8880	7687	8880	7687	8880
Playen I	10706	11811	10957	11782	10957	11782
Playen II	8080	9101	8327	9133	8327	9133

After that, each variable was modified into a percentage, such as the variables mentioned in the variable modification stage in data preprocessing. Furthermore, all variables were combined into a dataset with the health center and year indexes, resulting in Table 3 below:

Table 3. Dataset

Health Center	Year	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Nglipar I	2020	2.6	22.73	84.42	12.6689	80.25	7.1429	70.1053	86.1834
Nglipar II	2020	7.62	16.19	73.33	14.1558	82.48	2.3810	70.5856	83.8154
Gedangsari I	2020	13.59	29.13	86.41	23.0942	86.36	29.6117	79.4128	80.2406

Health Center	Year	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Gedangsari II	2020	5.44	13.61	41.5	20.6755	43.15	40.4762	88.8150	69.5634
Patuk I	2020	7.8	14.15	81.95	14.5110	81.12	33.1707	43.0375	86.3272
Patuk II	2020	4.3	17.74	100	19.3122	87.64	5.3763	47.1749	71.1782
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Semin II	2022	6.67	11.56	74.56	10.0629	82.23	0	50.1085	86.5653
Playen I	2022	4.47	3.83	65.84	15.6692	73.95	0	61.8557	92.9978
Playen II	2022	2.9	10.51	96.86	15.4401	97.61	100	55.0183	91.1749

The final part of the data preprocessing stage was checking the correlation value between variables to ensure that the data did not occur in multicollinearity. This stage was done by creating a correlation matrix visualization using Python programming. The following is the visualization of the correlation matrix in Figure 2.

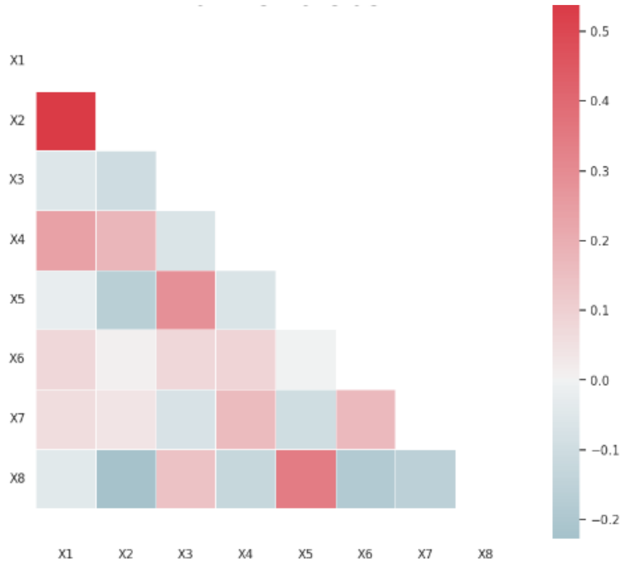


Figure 2. **Correlation matrix visualization**

The results of the correlation output between variables and intervals in the visualization show that there is no multicollinearity in the dataset because there is no correlation worth 0.7 to 1.0 or -0.7 to -1.0 (high or very high) [44]. Thus, the dataset is suitable for clustering. The following is a statistical description of the dataset.

Table 4. **Data description**

Descriptive Statistics	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Count	90	90	90	90	90	90	90	90
Mean	7.2268	16.2410	87.4370	16.4112	82.0068	30.7887	60.2705	85.5843
Std	2.8914	8.5538	13.6431	4.5274	14.8917	38.0692	18.0068	6.0514
Min	2.23	1.15	33.46	3.0211	34.31	0	11.4395	65.9735
Max	13.59	32.97	100	28.4753	100	100	100	95.6395

Table 4 shows that all variables have the same number of data objects, namely 90 data objects originating from 30 health centres in 2020 – 2022. Therefore, the clustering process is carried out on 1 dataset directly so that the clustering process is not carried out separately every year. Meanwhile, variable X_6 has the highest standard deviation value so it is likely to be the biggest influence in the clustering process. This is not a problem because the percentage value of each variable has the same influence according to the percentage value so it can be a guide in implementing stunting interventions. Thus, in reality there is still inequality in variable X_6.

After going through the preprocessing stage, the clustering stage was continued. Clustering using the k-means method was done using Python programming assistance, with a total of 2 clusters (clusters 0 and 1). Here is the Python script for the k-means method.

```
data = pd.read_excel('drive/MyDrive/filedata/data_all_new.xlsx', 'Sheet1', engine='openpyxl')
kmeans = KMeans(n_clusters=2, random_state=0)
cluster_labels = kmeans.fit_predict(data)
result_kmeans = pd.DataFrame(cluster_labels, index=data.index)
```

The following are the clustering results with the k-means method in **Table 5**.

Table 5. Clustering Results Using the K-Means Method (2 clusters)

Health Centre	2020	2021	2022
Nglipar I	0	0	1
Nglipar II	0	0	0
Gedangsari I	0	0	0
Gedangsari II	0	1	1
Patuk I	0	0	0
Patuk II	0	0	0
⋮	⋮	⋮	⋮
Semin II	0	0	0
Playen I	0	0	0
Playen II	0	1	1

Clustering using the fuzzy c-means method was also carried out using the help of Python programming, with the number of 2 clusters (clusters 0 and 1), power = 2, error = 0.005, maximum iterations = 1000. The following is a Python script for the fuzzy c-means method.

```
data = pd.read_excel('drive/MyDrive/filedata/data_all_new.xlsx', 'Sheet1', engine='openpyxl')
cntr, u, u0, d, jm, p, fpc = fuzz.cluster.means(data.T, c=2, m=2, error=0.005, maxiter=1000)
membership = u.T
pd.DataFrame(membership)
```

The following are the clustering results with the fuzzy c-means method in **Table 6**.

Table 6. Clustering Results Using the Fuzzy C-Means Method (2 Clusters)

Health Centre	2020	2021	2022
Nglipar I	0	0	1
Nglipar II	0	0	0
Gedangsari I	0	0	0
Gedangsari II	0	1	1
Patuk I	0	0	0
Patuk II	0	0	0
⋮	⋮	⋮	⋮
Semin II	0	0	0
Playen I	0	0	0
Playen II	0	1	1

The two methods' selection of the number of 2 clusters is in accordance with the results of clustering evaluation using the evaluation values of **silhouette coefficient (SC)**, **Calinski-Harabasz index (CHI)**, **Davies-Bouldin index (DBI),** and **Dunn index (DI)**. These evaluation values were determined using Python programming. **Tables 7** and **8** below show that the optimal **number of clusters (NC)** for both methods is 2 clusters.

Table 7. Value of all Evaluation Measures of the K-Means Method

NC	SC	CHI	DBI	DI
2	0.512593396	113.6774134	0.778100536	0.241207851
3	0.442552736	70.9720103	1.116769519	0.205170064
4	0.207097915	58.45622509	1.54912026	0.059961239
5	0.223414353	53.00000906	1.42213429	0.076751093

Table 8. Value of all Evaluation Measures of the Fuzzy C-Means Method

NC	SC	CHI	DBI	DI
2	0.512593396	113.6774134	0.778100536	0.305912211
3	0.257471995	69.42744554	1.553959435	0.06387343
4	0.206261256	57.78379477	1.600674762	0.059961239
5	0.207403495	50.14523611	1.591239141	0.070181182

Based on the clustering that has been done with the k-means and fuzzy c-means methods, in addition to having the same number of optimal clusters, it also has the same members of each cluster. This is also reinforced by the centroid of each cluster and similar t-SNE scatterplot visualizations, which are shown in **Table 9**, **Figure 3**, and **Figure 4** below.

Table 9. Centroid comparison

Variable	K-means		Fuzzy C-means	
	Centroid 0	Centroid 1	Centroid 0	Centroid 1
X_1	7.007742	7.71178571	7.017122181	7.576051272
X_2	15.90161	16.9925	16.29090048	16.49233683
X_3	86.40677	89.7182143	87.70590484	89.23991043
X_4	16.15307	16.9828913	16.0753475	17.07593608
X_5	82.14548	81.6996429	82.8869804	82.06913699
X_6	6.726609	84.0688853	6.027766843	81.24690836
X_7	58.47118	64.2547002	57.71658934	65.00583476
X_8	86.28483	84.0330361	86.57125662	84.07634962

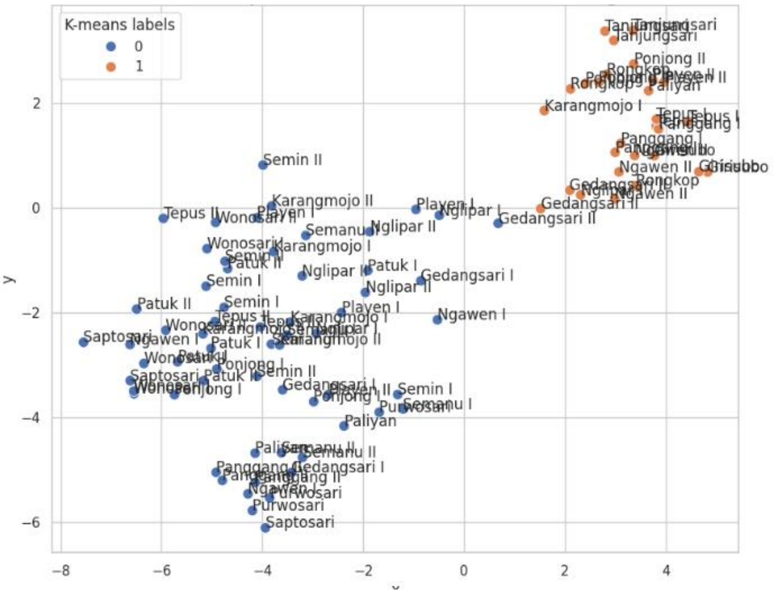


Figure 3. t-SNE Scatterplot of K-Means Clustering Results

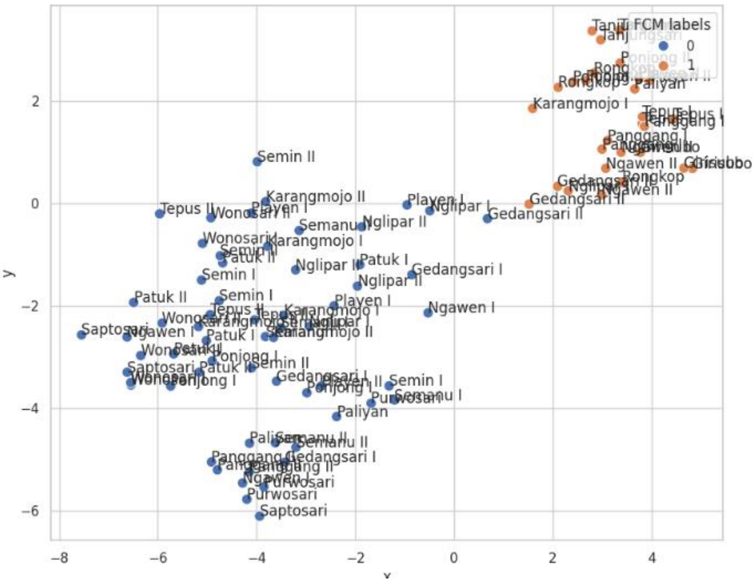


Figure 4. t-SNE Scatterplot of Fuzzy C-Means Clustering Results

Despite having similar cluster results, the optimal method used can still be determined based on Table 7 and Table 8. The tables show that **SC**, **CHI**, and **DBI** have similar evaluation values. The **SC** of the two methods shows decent cluster results $(\mathbf{0.5} < \mathbf{SC} \leq \mathbf{0.7})$ [45]. Meanwhile, the **DI** evaluation value in the fuzzy c-means method is greater than the k-means method. Therefore, in this study, the fuzzy c-means method is slightly better and optimal than the k-means method. The following is a graph of the centroid or average of each cluster resulting from fuzzy c-means clustering.

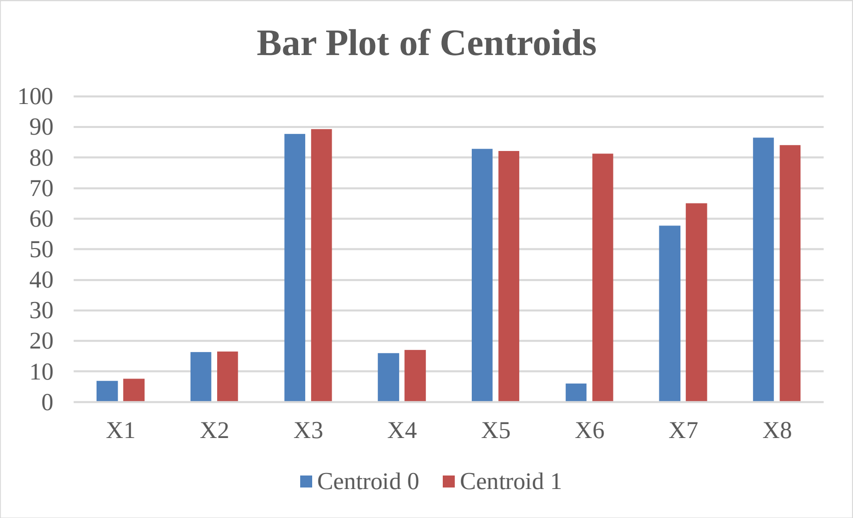


Figure 5. Bar Plot of Centroids

Based on Figure 5, it can be seen that cluster 0 has more variables that have a better percentage compared to cluster 1 because the values of $\boldsymbol{X_1, X_2, X_3, X_7}$ are slightly lower, and $\boldsymbol{X_5, X_8}$ are slightly higher. However, in the values of $\boldsymbol{X_3}$ and especially $\boldsymbol{X_6}$, cluster 1 is much better than cluster 0. Because of the significant difference between the differences between the centroids of each variable, it is necessary to analyze the characteristics of each variable from each cluster based on the centroids that have been obtained.

3.2 Discussion

Based on the clustering results that have been obtained, an analysis of cluster characteristics is needed to facilitate the interpretation of the results. This is done to know the stunting intervention indicators that need to be improved based on the characteristics of a cluster so that it can facilitate policymaking and give meaning to the mapping results. Cluster characteristics can be interpreted by comparing the profile of each cluster of variables that can be visualized using boxplots to facilitate interpretation in Figure 6 below.

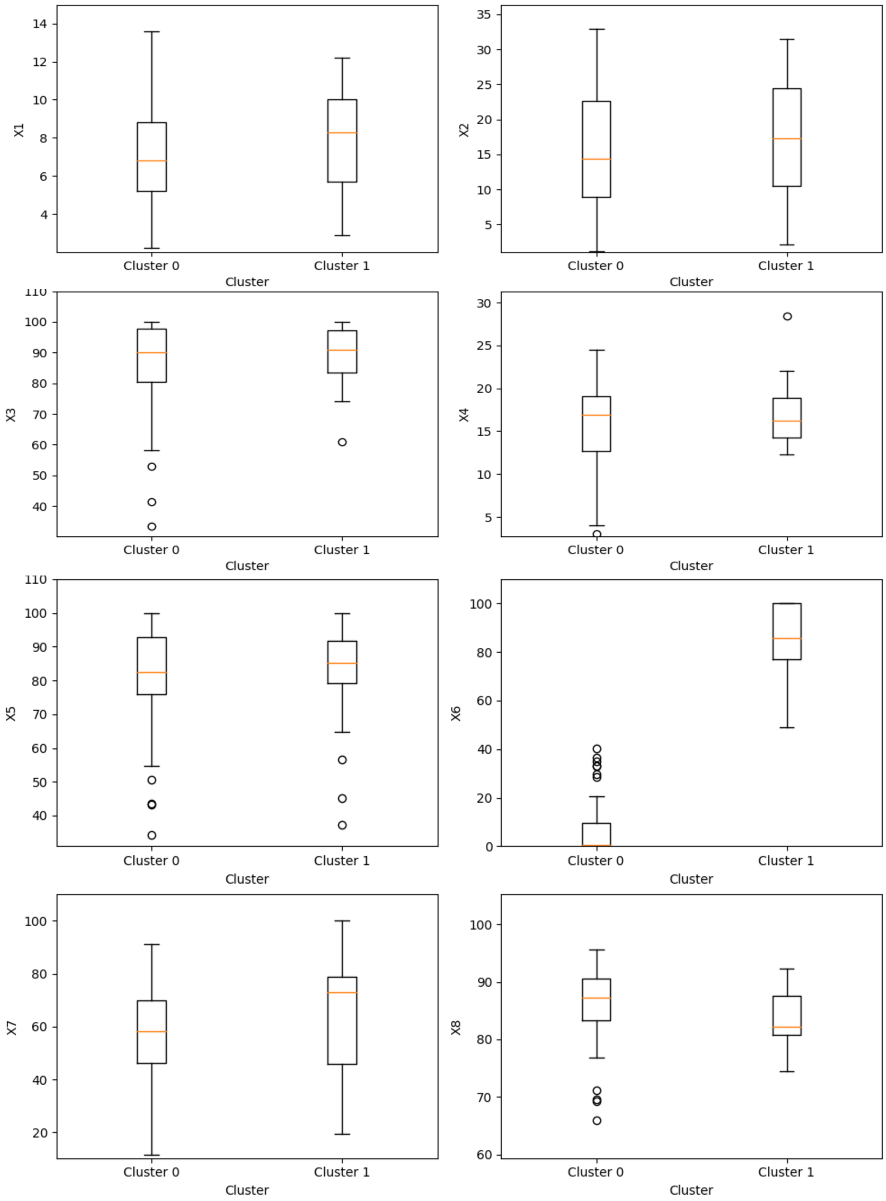


Figure 6. Boxplot of each variable

Based on **Figure 6**, the interventions carried out by community health centers in cluster 0 are pretty good regarding the indicators (variables) of the percentage of stunted toddlers and poor couples of reproductive ages. Therefore, there is a need to enhance interventions for other indicators, such as health services for infants and toddlers, especially monitoring to ensure exclusive breastfeeding, which is part of specific nutrition interventions. Meanwhile, the interventions performed by community health centers in cluster 1 are sufficiently good in terms of the percentage of infants receiving health services and toddlers receiving health services. Hence, there is a need to enhance interventions for other indicators, such as reducing stunted toddlers and poor couples of reproductive ages, which are part of sensitive nutrition interventions. As for the indicators of infants with low birth weight, stunted infants, and families with adequate drinking water, attention is still needed in both clusters, particularly in cluster 0, where it is more urgent to improve interventions for infants with low birth weight and stunted infants, while in cluster 1, it is more urgent to improve interventions for families with adequate drinking water. Thus, in general, for the future, community health centers in cluster 0 should focus more on specific nutrition interventions, such as for infants and toddlers. In contrast, those in cluster 1 should focus more on sensitive nutrition interventions, such as poverty and water adequacy.

Furthermore, mapping can be done every year with the assistance of QGIS 3.28.8 software. Before visualizing the map, the distribution of health centers in each cluster was determined based on the clustering results in **Table 10** below.

Table 10. Distribution of Health Centers in Each Cluster

Year	Cluster 0	Cluster 1
2020	Nglipar I, Nglipar II, Gedangsari I, Gedangsari II, Patuk I, Patuk II, Panggang II, Purwosari, Tepus II, Saptosari, Paliyan, Ponjong I, Wonosari I, Wonosari II, Semanu I, Semanu II, Ngawen I, Karangmojo I, Karangmojo II, Semin I, Semin II, Playen I, and Playen II.	Rongkop, Girisubo, Panggang I, Tepus I, Tanjungsari, Ponjong II, and Ngawen II.
2021	Nglipar I, Nglipar II, Gedangsari I, Patuk I, Patuk II, Panggang II, Purwosari, Tepus II, Saptosari, Paliyan, Ponjong I, Wonosari I, Wonosari II, Semanu I, Semanu II, Ngawen I, Karangmojo I, Karangmojo II, Semin I, Semin II, and Playen I.	Gedangsari II, Rongkop, Girisubo, Panggang I, Tepus I, Tanjungsari, Ponjong II, Ngawen II, and Playen II.
2022	Nglipar II, Gedangsari I, Patuk I, Patuk II, Panggang II, Purwosari, Tepus II, Saptosari, Ponjong I, Wonosari I, Wonosari II, Semanu I, Semanu II, Ngawen I, Karangmojo II, Semin I, Semin II, and Playen I.	Nglipar I, Gedangsari II, Rongkop, Girisubo, Panggang I, Tepus I, Tanjungsari, Paliyan, Ponjong II, Ngawen II, Karangmojo I, and Playen II.

Based on the distribution of health centers in each cluster, the following are the results of mapping stunting vulnerability in 2020 (Figure 7), 2021 (Figure 8), and 2022 (Figure 9) using QGIS 3.28.8 software.

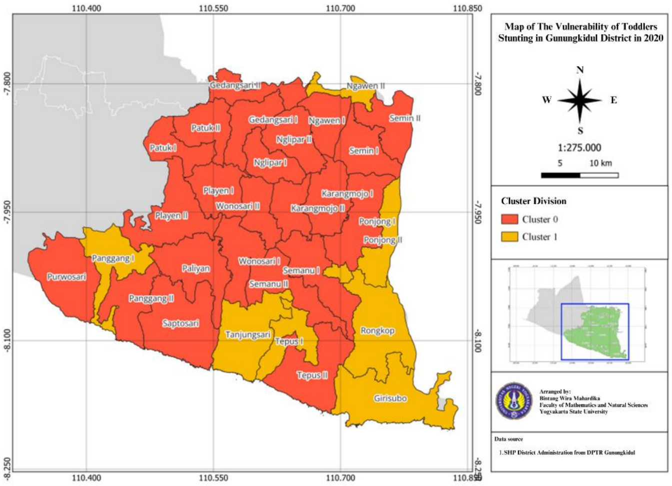


Figure 7. Stunting Vulnerability Map in 2020

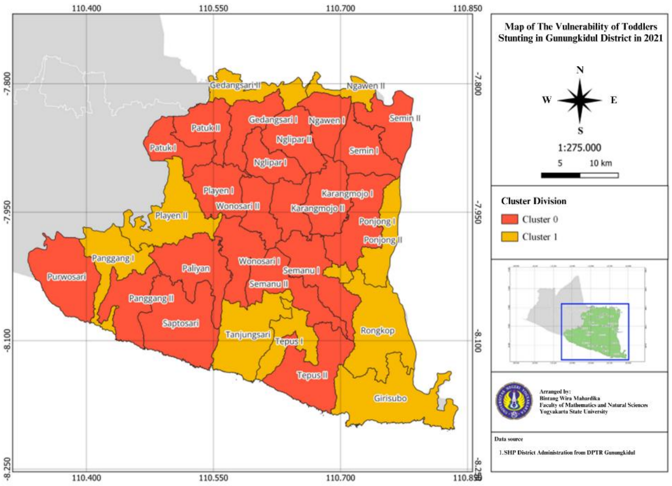


Figure 8. Stunting Vulnerability Map in 2021

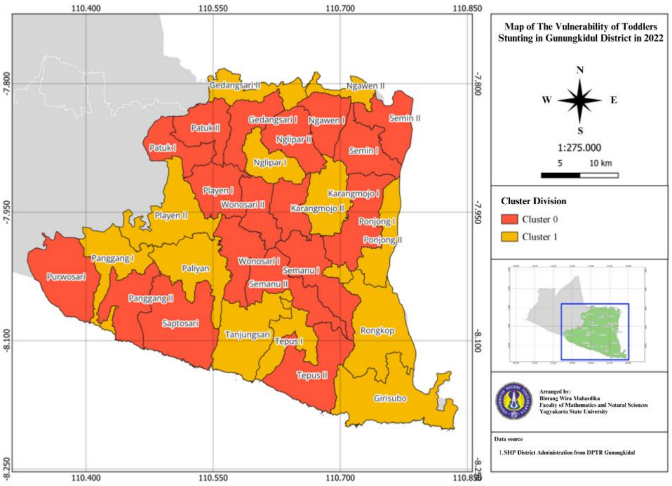


Figure 9. **Stunting Vulnerability Map in 2022**

In 2020, there were 23 health centers in cluster 0, shown in red, and 7 health centers in cluster 1, shown in yellow. Thus, in general, in 2020 many health centers have yet to be optimal in implementing specific nutrition interventions, such as on infants and toddlers, especially in increasing the percentage of infants given exclusive breastfeeding. While the 7 health centers in cluster 1 are not optimal in implementing nutrition-sensitive interventions, such as on poverty and water security, cluster 1 is more optimal in increasing the percentage of exclusively breastfed infants than cluster 0.

In 2021, there are 21 health centers in cluster 0, shown in red, and 9 health centers in cluster 1, shown in yellow. Thus, in 2021, there is considerable improvement in implementing interventions to increase the percentage of infants exclusively breastfed, which is part of the nutrition-specific interventions. However, on the other hand, the move of health centers from cluster 0 to cluster 1 may also be due to the deterioration of nutrition-sensitive interventions.

In 2022, there are 18 health centers in cluster 0, shown in red, and 12 health centers in cluster 1, shown in yellow. Thus, in 2022, there is considerable improvement in implementing interventions to increase the percentage of exclusively breastfed infants, which is part of the nutrition-specific interventions. However, just like before, the movement of health centers from cluster 0 to cluster 1 may also be due to the deterioration of nutrition-sensitive interventions.

4. CONCLUSIONS

This research discusses clustering and mapping of the stunting vulnerability of toddlers at the health center level in Gunungkidul Regency in 2020-2022 through the stages of data preprocessing, clustering, cluster evaluation, and analysis of cluster characteristics and mapping. The k-means and fuzzy c-means algorithms show that 2 clusters are the optimal number of clusters, with the fuzzy c-means method being slightly more optimal than the k-means method. The clustering results show that 28 objects are in Cluster 1, and 62 objects are in Cluster 0. In general, health centers in cluster 0 are less optimal in specific nutrition interventions, such as infants and toddlers. In contrast, health centers in cluster 1 are less optimal in nutrition-sensitive interventions, such as poverty and water eligibility. In 2020, there are 23 health centers in cluster 0 and 7 health centers in cluster 1. Meanwhile, in 2021 there are 21 health centers in cluster 0 and 9 health centers in cluster 1. In 2022 there are 18 health centers in cluster 0 and 12 health centers in cluster 1.

ACKNOWLEDGMENT

We are grateful to the agencies that helped with the research, including the Gunungkidul District Health Office and the Gunungkidul District Regional Development Planning Agency for providing data and the Gunungkidul Land and Spatial Planning Office for providing administrative maps.

REFERENCES

[1] K. A. Atalell, M. A. Techane, B. Terefe, and T. T. Tamir, “Mapping stunted children in Ethiopia using two decades of data between 2000 and 2019. A geospatial analysis through the Bayesian approach,” *J Health Popul Nutr*, vol. 42, no. 1, pp. 1–9, Dec. 2023, doi: 10.1186/s41043-023-00412-3.

[2] W. Sartika, S. Suryono, and A. Wibowo, “Information System for Evaluating Specific Interventions of Stunting Case Using K-means Clustering,” in *E3S Web of Conferences*, EDP Sciences, Nov. 2020, pp. 1–10. doi: 10.1051/e3sconf/202020213003.

[3] N. Izza, W. Purnomo, and IMahmudah, “Implementation of the K-means Clustering Method on Stunting Case in Indonesia,” *International Journal of Advances in Scientific Research and Engineering*, vol. 5, no. 6, pp. 103–107, 2019, doi: 10.31695/ijasre.2019.33258.

[4] M. Handayani and M. F. L. Sibuea, “Performance Analysis of Clustering Models Based on Machine Learning in Stunting Data Mapping,” *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 9, no. 4, pp. 715–720, Sep. 2023, doi: 10.33330/jurteksi.v9i4.2770.

[5] I. P. Sari, Al-Khowarizmi, O. K. Sulaiman, and D. Apdilah, “Implementation of Data Classification Using K-Means Algorithm in Clustering Stunting Cases,” *Journal of Computer Science, Information Technology and Telecommunication Engineering*, vol. 4, no. 2, pp. 402–412, Sep. 2023, doi: 10.30596/jcositte.v4i2.15765.

[6] M. Ula, A. F. Ulva, Mauliza, I. Sahputra, and Ridwan, “Implementation of Machine Learning in Determining Nutritional Status using the Complete Linkage Agglomerative Hierarchical Clustering Method,” *Jurnal Mantik*, vol. 5, no. 3, pp. 1910–1914, 2021.

[7] A. Aswi, B. Poerwanto, Sudarmin, and Nurwan, “Bayesian Spatial Modelling of Stunting Cases in South Sulawesi Province: Influential Factors and Relative Risk,” in *Proceedings of the 5th International Conference on Statistics, Mathematics, Teaching, and Research 2023 (ICSMTR 2023)*, 2023, pp. 87–96. doi: 10.2991/978-94-6463-332-0_11.

[8] N. Istiqomah, H. Wijayanto, and F. M. Afendi, “Clustering Districts Based on Influencing Factors of Child Undernutrition (Stunting),” in *Proceeding of International Conference On Research, Implementation And Education Of Mathematics And Sciences 2015*, 2015, pp. 239–244.

[9] H. Jamaludin and B. Y. Dharmahita, “K-Means Clustering Analysis on the Distribution of Stunting Cases In Mojokerto Regency in June 2022,” *Jurnal Media Pratama*, vol. 17, no. 1, pp. 33–44, 2023, [Online]. Available: https://data.go.id

[10] P. W. Sudarmadji and C. E. B. Bire, “Implementation of K-means Clustering Algoritm to Determine Stunted Status in Children Under Two Years Old,” in *ICESC 2019*, European Alliance for Innovation n.o., Dec. 2019. doi: 10.4108/eai.18-10-2019.2289979.

[11] B. Khura *et al.*, “Mapping Concurrent Wasting and Stunting Among Children Under Five in India: A Multilevel Analysis,” *Int J Public Health*, vol. 68, pp. 1–11, 2023, doi: 10.3389/ijph.2023.1605654.

[12] R. Hemalatha *et al.*, “Mapping of variations in child stunting, wasting and underweight within the states of India: the Global Burden of Disease Study 2000–2017,” *EClinicalMedicine*, vol. 22, pp. 1–16, May 2020, doi: 10.1016/j.eclinm.2020.100317.

[13] K. Y. Ahmed, K. E. Agho, A. Page, A. Arora, and F. A. Ogbo, “Mapping Geographical Differences and Examining the Determinants of Childhood Stunting in Ethiopia: A Bayesian Geostatistical Analysis,” *Nutrients*, vol. 13, no. 6, pp. 1–21, Jun. 2021, doi: 10.3390/nu13062104.

[14] L. E. Suranny and F. C. Maharani, “Mapping of Community Empowerment in Prevention Stunting in Kabupaten Wonogiri Through ‘Sego Sak Ceting,’” in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Nov. 2021, pp. 1–11. doi: 10.1088/1755-1315/887/1/012035.

[15] D. S. Effendy, P. Prangthip, N. Soonthornworasiri, P. Winichagoon, and K. Kwanbunjan, “Nutrition education in Southeast Sulawesi Province, Indonesia: A cluster randomized controlled study,” *Matern Child Nutr*, vol. 16, no. 4, pp. 1–14, Oct. 2020, doi: 10.1111/mcn.13030.

[16] K. Y. Ahmed, A. G. Ross, S. M. Hussien, K. E. Agho, B. O. Olusanya, and F. A. Ogbo, “Mapping Local Variations and the Determinants of Childhood Stunting in Nigeria,” *Int J Environ Res Public Health*, vol. 20, no. 4, pp. 1–16, Feb. 2023, doi: 10.3390/ijerph20043250.

[17] S. M. Rambe and Suendri, “Geographic Information System Mapping Risk Factors Stunting Using Methods Geographically Weighted Regression,” *Journal of Applied Geospatial Information*, vol. 7, no. 2, pp. 1075–1079, 2023, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAGI

[18] A. Iriany, W. Ngabu, D. Arianto, and A. Putra, “Classification of Stunting Using Geographically Weighted Regression-Kriging Case Study: Stunting in East Java,” *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 1, pp. 0495–0504, Apr. 2023, doi: 10.30598/barekengvol17iss1pp0495-0504.

[19] E. Selviyanti, M. C. Roziqin, D. S. H. Putra, and M. S. Noor, “Intelligent Application of Stunting Monitoring and Mapping Systems (Smart Ting) in Toddlers Based on Android in Jember,” in *2nd International Conference on Social Science, Humanity and Public Health (ICOSHIP 2021)*, 2022, pp. 147–157.

[20] F. A. Johnson, “Spatiotemporal clustering and correlates of childhood stunting in Ghana: Analysis of the fixed and nonlinear associative effects of socio-demographic and socio-ecological factors,” *PLoS One*, vol. 17, no. 2, pp. 1–22, 2022, doi: 10.1371/journal.pone.0263726.

[21] M. A. Ramdani and S. Abdullah, “Application of partitioning around medoids cluster for analysis of stunting in 100 priority regencies in Indonesia,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2021, pp. 1–10. doi: 10.1088/1742-6596/1722/1/012097.

[22] S. S. Nagari and L. Inayati, “Implementation of Clustering Using K-means Method to Determine Nutritional Status,” *Jurnal Biometrika dan Kependudukan*, vol. 9, no. 1, pp. 62–68, Jun. 2019, doi: 10.20473/jbk.v9i1.2020.62-68.

[23] I. K. Hasan, Nurwan, N. Falaq, and M. R. F. Payu, “Optimization Fuzzy Geographically Weighted Clustering with Gravitational Search Algorithm for Factors Analysis Associated with Stunting,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 1, pp. 120–128, Feb. 2023, doi: 10.29207/resti.v7i1.4508.

[24] S. P. Tamba, M. D. Batubara, W. Purba, M. Sihombing, V. M. M. Siregar, and J. Banjarnahor, "Book data grouping in libraries using the k-means clustering method," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Sep. 2019, pp. 1–7. doi: 10.1088/1742-6596/1230/1/012074.

[25] R. D. Christyanti, D. Sulaiman, A. P. Utomo, and M. Ayyub, "Implementation of Fuzzy C-Means in Clustering Stunting Prone Areas," *International Journal of Natural Science and Engineering*, vol. 6, no. 3, pp. 110–121, Oct. 2022, doi: 10.23887/ijnse.v6i3.53048.

[26] M. M. Saleck, A. El Moutaouakkil, M. Moucouf, M. Bouchaib, H. Samira, and J. Zineb, "Breast Mass Segmentation Using a Semi-automatic Procedure Based on Fuzzy C-means Clustering," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 16, no. 2, pp. 665–672, Apr. 2018, doi: 10.12928/TELKOMNIKA.v16i2.6193.

[27] D. P. Sari, D. Rosadi, A. R. Effendie, and Danardono, "K-means and bayesian networks to determine building damage levels," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 2, pp. 719–727, 2019, doi: 10.12928/TELKOMNIKA.V17I2.11756.

[28] C. C. Aggarwal and C. K. Reddy, *Data Clustering Algorithms and Applications*. 2014.

[29] S. H. Gebreyesus, D. H. Mariam, T. Woldehanna, and B. Lindtjørn, "Local spatial clustering of stunting and wasting among children under the age of 5 years: Implications for intervention strategies," *Public Health Nutr*, vol. 19, no. 8, pp. 1417–1427, Jun. 2015, doi: 10.1017/S1368980015003377.

[30] A. K. Rahmansyah, A. T. S. Aziz, N. Novianto, and D. Rolliawati, "Perbandingan Algoritma K-Means dan Fuzzy C-Means Untuk Clustering Puskesmas Berdasarkan Gizi Balita Surabaya," *Jurnal PROCESSOR*, vol. 18, no. 1, pp. 83–88, Apr. 2023, doi: 10.33998/processor.2023.18.1.696.

[31] "Sistem Informasi Komunikasi Data Kesehatan Keluarga." Accessed: Jun. 27, 2024. [Online]. Available: https://kesgadiy.web.id/lihat-data

[32] J. Mann and S. Truswell, *Essentials of Human Nutrition*. 2012.

[33] S. A. S. Mahayana, E. Chundrayetti, and Yulistini, "Faktor Risiko yang Berpengaruh terhadap Kejadian Berat Badan Lahir Rendah di RSUP Dr. M. Djamil Padang," *Jurnal Kesehatan Andalas*, vol. 4, no. 3, pp. 664–673, 2015.

[34] Badan Kependudukan dan Keluarga Berencana Nasional Republik Indonesia, "Peraturan Badan Kependudukan dan Keluarga Berencana Nasional Republik Indonesia Nomor 1 Tahun 2023, tentang Pemenuhan Kebutuhan Alat Dan Obat Kontrasepsi Bagi Pasangan Usia Subur Dalam Pelayanan Keluarga Berencana," 2023.

[35] Kementerian Kesehatan Republik Indonesia, *Buku Panduan Untuk Siswa Aksi Bergizi*. 2019.

[36] A. D. Mengistu, "The Effects of Segmentation Techniques in Digital Image Based Identification of Ethiopian Coffee Variety," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 16, no. 2, pp. 713–717, Apr. 2018, doi: 10.12928/TELKOMNIKA.v16i2.8419.

[37] W. W. Pribadi, A. Yunus, and A. S. Wiguna, "Perbandingan Metode K-Means Euclidean Distance dan Manhattan Distance pada Penentuan Zonasi COVID-19 di Kabupaten Malang," *Jurnal Mahasiswa Teknik Informatika*, vol. 6, no. 2, pp. 493–500, 2022.

[38] D. L. Rahakbauw, V. Y. I. Ilwaru, and M. H. Hahury, "Implementasi Fuzzy C-means Clustering dalam Penentuan Beasiswa," *BAREKENG Jurnal Ilmu Matematika dan Terapan*, vol. 11, no. 1, pp. 1–11, 2017.

[39] Y. Yang and S. Huang, "Image Segmentation by Fuzzy C-means Clustering Algorithm with a Novel Penalty Term," *Computing and Informatics*, vol. 26, pp. 17–31, 2007.

[40] G. R. Suraya and A. W. Wijayanto, "Comparison of Hierarchical Clustering, K-means, K-medoids, and Fuzzy C-means Methods in Grouping Provinces in Indonesia According to the Special Index for Handling Stunting," *Indonesian Journal of Statistics and Its Applications*, vol. 6, no. 2, pp. 180–201, Aug. 2022, doi: 10.29244/ijsa.v6i2p180-201.

[41] B. Desgraupes, "Clustering Indices," 2017, pp. 1–34.

[42] J. Baarsch and M. E. Celebi, "Investigation of Internal Validity Measures for K-Means Clustering," in *Proceedings of the International MultiConference of Engineers and Computer Scientist*, Newswood Ltd., 2012, pp. 1701–1706.

[43] T. Zulyanti and Noeryanti, "Perbandingan Pengelompokan Usaha Mikro Kecil dan Menengah di Kabupaten Klaten Tahun 2019 dengan Metode K-Means dan Clustering Large Application," *Jurnal Statistika Industri dan Komputasi*, vol. 7, no. 1, pp. 46–59, 2022.

[44] Mundir, *Statistik Pendidikan Pengantar Analisis Data Untuk Penulisan Skripsi dan Tesis*. 2012.

[45] M. R. Anggraeni, U. Yudatama, and Maimunah, "Clustering Prevalensi Stunting Balita Menggunakan Agglomerative Hierarchical Clustering," *JURNAL MEDIA INFORMATIKA BUDIDARMA* , vol. 7, no. 1, pp. 351–359, 2023, doi: 10.30865/mib.v7i1.5501.