The Gath–Geva Algorithm for Clustering Spatial Inequality of Stunting in East Nusa Tenggara Province

M R Nufus[1,*]

[1] Forest Management Study Program, State Agricultural Polytechnic of Kupang, Kupang, Indonesia

*Corresponding author's email: mhytha.nufus88@gmail.com

Abstract. Stunting remains a critical public health issue in Indonesia, particularly in East Nusa Tenggara (NTT), where prevalence rates are among the highest nationally. This study aims to classify districts and municipalities in East Nusa Tenggara Province based on socioeconomic and health-related indicators associated with stunting vulnerability. Using the Gath–Geva (Fuzzy K-Means Entropy) clustering algorithm, four key variables were analyzed, including poverty rate, access to proper housing, open unemployment rate, and number of health facilities. The results identified three distinct clusters with different regional characteristics. Cluster 1 consists of areas with low poverty and well-developed health infrastructure but relatively high unemployment rates. Cluster 2 represents the most vulnerable regions characterized by high poverty, poor housing access, and limited health facilities, while Cluster 3 comprises more stable areas with better housing, low unemployment, and adequate healthcare services. The silhouette coefficient value of 0.41 indicates that the three-cluster structure provides a reasonably good level of separation and internal consistency. These findings highlight that stunting vulnerability is strongly influenced by socioeconomic disparities and the distribution of health infrastructure. Therefore, intervention strategies should be tailored to the characteristics of each cluster, emphasizing integrated actions in high-risk regions and preventive measures in more stable areas to accelerate stunting reduction across East Nusa Tenggara Province.
Keyword: Cluster Analysis, East Nusa Tenggara, Gath-Geva Algorithm, Stunting.

1. Introduction

Stunting is a form of chronic nutritional problem that continues to pose a major challenge in the development of public health in Indonesia [8]. This condition is characterized by children's height being below the age-specific standard, resulting from prolonged nutritional deficiencies and recurrent infections starting from pregnancy until the first two years of life [20]. According to the World Health Organization (WHO), one of the most prevalent forms of malnutrition among children worldwide is linear growth impairment, a condition that affects both physical and cognitive development and impacts approximately 478 million children under five years of age [19]. The long-term consequences of stunting extend beyond impaired physical growth, encompassing delayed cognitive development, reduced productivity in adulthood, and a heightened susceptibility to non-communicable diseases [13].

East Nusa Tenggara (NTT) Province is among the regions with one of the highest stunting prevalence rates in Indonesia [2]. According to data from BPS – Statistics Indonesia of East Nusa Tenggara Province (NTT) in recent years, the prevalence of stunting in this province has shown a declining trend, decreasing from 21% in 2021 to 18% in 2022, and further dropping to 15% in 2023 in figure 1.

However, according to the most recent data from 2024, the prevalence of stunting in East Nusa Tenggara has risen again to 16.9% in figure 1, indicating a new challenge in sustaining the reduction of stunting rates [1].

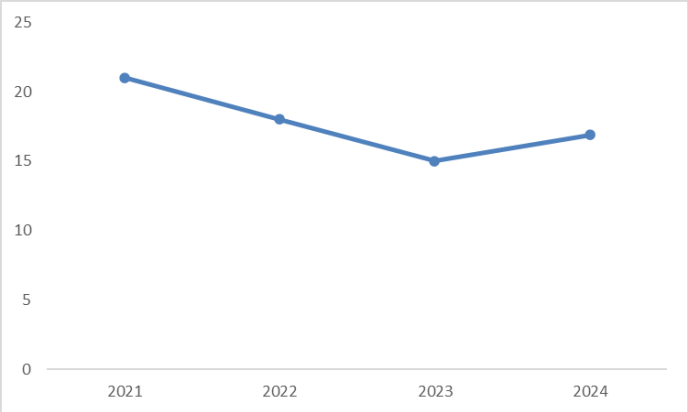

Figure 1. Stunting Rate in East Nusa Tenggara Province from 2021 to 2024.

(Source: BPS – Statistics Indonesia, 2024)

From the figure 1, we know that the increase of stunting indicates that the previous decline in stunting has not been evenly distributed nor consistently sustained. The underlying factors are highly complex, encompassing disparities in access to healthcare services, variations in poverty levels, maternal education, and differing environmental conditions across districts and municipalities. Furthermore, the geographic characteristics of East Nusa Tenggara, which consists of multiple islands with limited accessibility, reinforce the spatial disparities in stunting risk [20]. This spatial disparity necessitates an analytical approach capable of capturing local variations across regions. One of the methods that can be applied is clustering, a data analysis technique used to group stunting cases in East Nusa Tenggara (NTT) into clusters with similar characteristics. This method effectively simplifies complex data based on indicators such as nutritional status, poverty, sanitation, and consumption patterns. Being unsupervised, clustering does not require a target variable, making it suitable for stunting data that are unlabeled [15]. Furthermore, this method assists the government in identifying priority areas and designing policies that are more specific, differentiated, and contextually relevant. Thus, clustering not only provides an overview of stunting distribution but also offers strategic insights for targeted interventions.

One clustering approach that can be applied is the Gath–Geva (GG) algorithm, which is an extension of the Fuzzy C-Means method that accounts for cluster shapes and densities that are not necessarily spherical. This algorithm incorporates a fuzzy covariance matrix and employs maximum likelihood estimation to identify clusters with complex and overlapping characteristics [12]. With these advantages, the Gath–Geva method is considered more adaptive for analyzing spatial inequality of stunting across regions, as it can capture variations in data distribution and density among districts and municipalities with diverse socio-economic and geographical characteristics, such as those found in East Nusa Tenggara Province. Through this approach, the study seeks to examine the spatial disparities of stunting in East Nusa Tenggara Province and to identify the socioeconomic factors influencing its local prevalence. The findings are expected to provide scientific contributions as well as serve as a foundation for developing more targeted and region-based policies, thereby supporting a fairer and more sustainable acceleration of stunting reduction programs.

2. Research Method

2.1. Data and Variables

The data employed in this study are secondary data obtained from BPS – Statistics Indonesia of East Nusa Tenggara Province (NTT). Although the dataset consists of only 22 administrative units, such a sample size remains acceptable for clustering analysis, particularly when the cluster structure shows distinct separations or large effect sizes among groups. According to Dalmaijer, Nord, and Astle (2022), small-sample cluster analysis can still achieve reliable and interpretable results when the between-group variance is sufficiently high and the variable-to-observation ratio is moderate [4]. In this study, the variables were selected to maximize inter-district variability, ensuring that the clustering process could capture meaningful distinctions despite the relatively limited number of units. Similar approaches using fewer than 30 observations have also been applied effectively in regional-level studies of public health disparities [11]. Therefore, the number of units analyzed in this research is considered statistically adequate to generate valid and contextually relevant cluster structures for East Nusa Tenggara Province.

Table 1. Research Variables.

Notation	Variables	Description
X1	Poor Population	Ratio
X2	Household Percentage with Access to Proper Housing	Ratio
X3	Open Unemployment Rate	Ratio
X4	Number of Health Facilities	Ratio

The variable X1 (poor population) was included because economic deprivation directly affects household food security and access to health services, which are critical determinants of child nutritional status [17]. Household access to proper housing reflects the environmental and sanitation conditions that contribute to stunting through exposure to poor living standards [18]. Open unemployment rate captures the labor and income stability dimension, which influences household purchasing power and nutrition affordability [5]. Meanwhile, X4 (number of health facilities) represents the availability of healthcare infrastructure that supports early growth monitoring, maternal care, and nutrition interventions [9]. This theoretical framework aligns with the UNICEF Conceptual Framework for the Determinants of Child Undernutrition (2021), which categorizes stunting determinants into economic, environmental, and service-related factors. Therefore, even though the number of variables is limited, they are conceptually comprehensive and statistically sufficient to capture the multidimensional disparities that contribute to stunting vulnerability in East Nusa Tenggara Province.

2.2. Assumption Testing

Prior to conducting cluster analysis, it is essential to ensure that the research data satisfy several fundamental assumptions to be considered appropriate for the clustering method.

1. Normal Multivariate Assumption

   The multivariate normality assumption plays a crucial role in ensuring accurate parameter estimation through the Expectation–Maximization (EM) algorithm while also supporting more reliable interpretation of clustering outcomes. However, in real-world data, distributional patterns are often not perfectly normal, such as being skewed, heavy-tailed, or containing outliers. Therefore, it is necessary to test multivariate normality within each cluster prior to applying Gaussian Mixture Model (GMM)-based analysis. In general, the multivariate normality assumption not only determines the feasibility of implementing GMM but also guides the decision on whether the data should be transformed or analyzed using alternative distributional models. This aligns with recent studies which emphasize that non-Gaussian distributions, such as multivariate skew-normal, skew-t, and leptokurtic-normal, can generate clusters that are more adaptive and accurate when applied to empirical data [16]. The assumption of multivariate normality will be tested using the Mahalanobis distance approach. Each observation's Mahalanobis distance will be computed to identify potential multivariate outliers and to assess whether the data approximately follow a multivariate normal distribution. Furthermore, the Mahalanobis correlation plot will be examined to verify the linearity

and distributional characteristics of the data, providing additional evidence for the validity of the multivariate normality assumption prior to further analysis.

H_0 : the data follow a multivariate normal distribution

H_1 : the data are not follow a multivariate normal distribution

The procedures undertaken to assess the fulfillment of multivariate normality assumptions in the research data are outlined as follows.

a. Computing the value of $d_j^2 = \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)' \mathbf{S}^{-1} \left(\mathbf{X}_j - \bar{\mathbf{X}}\right);\ j = 1,2,3,\cdots,n$

Where it is known that,

d_j^2 : the distance value of the j-th data point

$\mathbf{X}_j$: the j-th variable vector

$\bar{\mathbf{X}}$: mean vector value

$\mathbf{S}^{-1}$: The inverse value of the variance-covariance matrix

b. Ordering values d_j^2 from all the observation carried out, it can be concluded that $d_1^2 \le d_2^2 \le \cdots \le d_j^2$

c. Generating a Q-Q plot based on the values d_j^2 as the X-axis and the upper quantile value as the Y-axis.

$$\left(d_j^2, \chi_p^2\left(\frac{j-0.5}{n}\right)\right) \quad (1)$$

The data can be concluded to follow a multivariate normal distribution if the visualized plot forms a straight line and the statistical value supports this assumption $d_j^2 \le \chi_{p;0.05}^2$ or p-value>0.05 [7].

2. Data Adequacy Assumption

The assumption of data adequacy in cluster analysis refers to the minimum number of samples required in each subgroup to ensure sufficient statistical power and valid interpretation of the clustering outcome. Recent studies suggest that when the separation between clusters is distinct (large effect size), cluster analysis can still perform effectively even with relatively small samples. In such cases, approximately 20 to 30 observations per cluster are considered sufficient to achieve high accuracy in identifying the true cluster structure [4].

H_0 : the data size is sufficient for analysis

H_1 : the data size is not sufficient for analysis

Statistics test,

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \ ;\ i = 1,2,3,\cdots,p \ ;\ j = 1,2,3,\cdots p \quad (2)$$

where,

r_{ij} : the correlation coefficient between variable i and variable j

a_{ij} : the partial coefficient between variable i and variable j

The KMO value ranges from 0 to 1, where a value closer to 1 indicates stronger inter-variable correlations, suggesting that the dataset is adequate for analysis. In practice, a KMO score above 0.50 is considered the minimum threshold for data adequacy, a score exceeding 0.70 reflects good sampling quality, and a value greater than 0.80 indicates very high adequacy of the data.

3. Multicollinearity Assumption

One of the essential assumptions in multivariate analysis is the presence of correlations among variables, which allows for the construction of an appropriate factor structure or analytical model. To assess this assumption, the Variance Inflation Factor (VIF) is employed to determine the extent

to which the variance of regression coefficient estimates is inflated due to multicollinearity among independent variables. This evaluation is carried out using specific test statistics [10].

$$VIF_j = \frac{1}{1 - R_j^2} \tag{3}$$

It is known that R_j^2 is the coefficient of determination of the j-th independent variable regressed on all the other independent variables. If the VIF exceeds 10, it is considered a serious multicollinearity problem that needs to be addressed [3] and an ideal regression model should be free from correlations among its independent variables [14].

2.3. Gath-Geva Clustering Algorithm

The Gath–Geva (GG) clustering algorithm is an extension of the Fuzzy C-Means method designed to address non-spherical cluster structures by incorporating a fuzzy covariance matrix and a distance norm derived from maximum likelihood estimation. Compared to Fuzzy C-Means, which relies solely on Euclidean distance, the Gath–Geva algorithm is capable of clustering data with varying densities and shapes, making it more adaptable in identifying complex data structures [12]. The procedure for applying the Gath–Geva algorithm can be outlined as follows.

a. Initializing the partition matrix U randomly, followed by calculating the centroid of each cluster.
b. The distance was calculated using the applied formula.

$$D_{kg} = \frac{(2\pi)^{\frac{\pi}{2}} \sqrt{\det(F_{wg})}}{\alpha_g} \exp\left(\frac{1}{2}(x_k - v_g)^T F_{wg}^{-1}(x_k - v_g)\right) \tag{4}$$

$$F_{wg} = \frac{\sum_{k=1}^{N}\left(\mu_{kg}^w (x_k - v_g)(x_k - v_g)^T\right)}{\sum_{k=1}^{N}(\mu_{kg})^w} \text{ and } \alpha_g = \frac{1}{N}\sum_{k=1}^{N}\mu_{kg} \tag{5}$$

With,
F_{wg} : The fuzzy covariance matrix of the *g*-th cluster
μ_{kg} : The posterior probability of the selected *g*-th cluster
x_k : The *k*-th data point in the cluster
v_g : the mean of data points in the *g*-th cluster
α_g : the prior probability of the selected *g*-th cluster

c. Calculating the values of the newly formed membership function U_{t+1}.

$$U_{kg} = \left[\sum_{j=1}^{c}\left(\frac{D(x_k, v_g)}{D(x_k, v_g)}\right)^{\frac{2}{m-1}}\right]^{-1} \tag{6}$$

d. If the change in the membership function value is still greater than the threshold value is still greater than the threshold value $\| U_{t+1} - U_t \| < \varepsilon$, then return to step b.

2.4. Validity Index

Cluster validity indices (CVI) serve as metrics to evaluate the quality of clustering results. The effectiveness of these indices is largely determined by the shape, size, and distribution of the data [11]. In this study, the validity index applied is the Silhouette coefficient. The Silhouette coefficient serves as a metric to evaluate how well the data are assigned to their respective clusters. Its values range from −1 to +1, where a higher score (closer to +1) indicates that the data points are well matched to their own cluster and clearly separated from other clusters, while a lower score (approaching −1) suggests that the data may be more appropriately placed in a different cluster [6].

$$s(c) = \frac{b_c - a_c}{\max\{a_c, b_c\}} \tag{7}$$

It is defined that $a_{(c)}$ represents the average distance from point *c* to all other points within the same cluster, while $b_{(c)}$ denotes the average distance from point *c* to all points in the nearest neighboring cluster. The optimal number of clusters is determined by the configuration that produces the highest Silhouette value.

3. Result and Discussion

3.1. Data Characteristics

To understand the distributional characteristics of the dataset used in this research, descriptive statistical measures such as the mean, minimum, maximum, and standard deviation were calculated. The results of these descriptive statistics are presented as follows,

Table 2. Data Characteristics.

Variable	Mean	Minimum	Maximum	Standard Deviation
X1	20.14	8.24	30.84	6.64
X2	46.2	23.2	69.99	13.59
X3	2.88	0.51	8.6	1.65
X4	89.64	29	185	36.95

The table 2 presents the descriptive statistical results for each research variable. The mean percentage of poor population (X1) is 20.14%, ranging from 8.24% to 30.84%, indicating regional disparities in poverty levels across East Nusa Tenggara. The variable X2, representing the household percentage with access to proper housing, has a mean value of 46.2%, with a range from 23.2% to 69.99%, suggesting inequality in access to adequate housing facilities. The open unemployment rate (X3) shows a mean of 2.88% and a standard deviation of 1.65, reflecting relatively small inter-district variation. Meanwhile, the number of health facilities (X4) has a mean of 89.64 units, ranging from 29 to 185 units, with a standard deviation of 36.95, indicating significant differences in healthcare service availability among districts. Overall, these descriptive statistics reveal notable regional disparities in socioeconomic and health-related conditions that may contribute to varying stunting risk levels across the province.

Subsequently, a graph is presented to illustrate the data characteristics of stunting in East Nusa Tenggara Province.

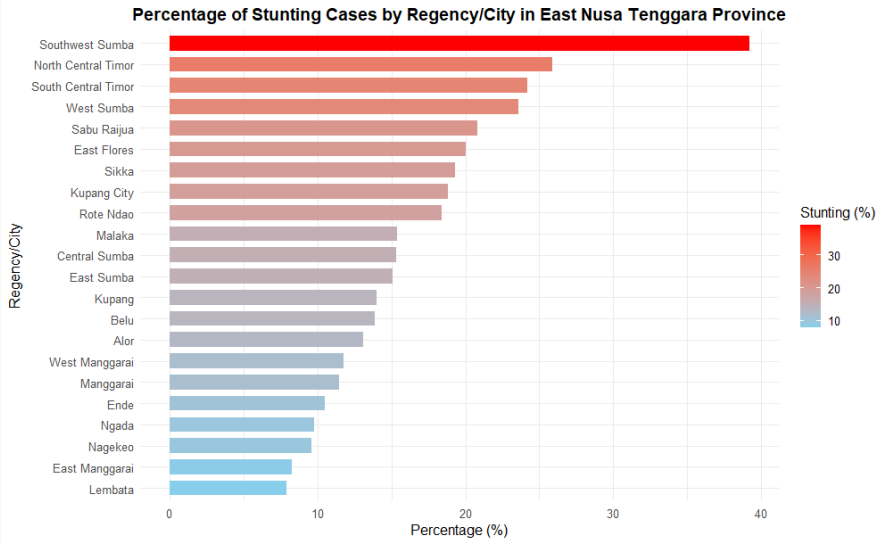

Figure 2. Data Characteristics.

Figure 2 illustrates the distribution of stunting prevalence across districts and municipalities in East Nusa Tenggara Province. The figure reveals considerable variation among regions. Southwest Sumba

recorded the highest prevalence, reaching approximately 39 percent, followed by North Central Timor and South Central Timor, each exceeding 20 percent. Other districts such as Sabu Raijua, East Flores, and Sikka also displayed relatively high stunting rates compared to most areas in the province. In contrast, Lembata reported the lowest prevalence at below 10 percent, with East Manggarai, Nagekeo, and Ngada also falling within the lower range. This pattern highlights notable disparities in child nutrition conditions across NTT. The observed variation may be influenced by differences in access to nutritious food, socioeconomic conditions, and the availability of health and sanitation services in each district. Overall, the figure underscores that stunting remains a critical public health challenge in NTT, particularly in regions with high prevalence, thereby calling for more targeted and area-based intervention strategies to reduce stunting rates equitably across all districts.

3.2. Clustering Assumption Testing

1. Multivariate Normality Assumption

Multivariate normality refers to the assumption that the joint distribution of research variables follows a multivariate normal distribution, rather than simply assuming normality for each variable individually. Therefore, even if every variable satisfies the univariate normality assumption, the overall dataset may not necessarily exhibit multivariate normality. A commonly applied approach to test this assumption is the Mahalanobis Distance, which measures the distance of each observation from the center of the data distribution. The following presents the visual results of the multivariate normality test using the Mahalanobis Distance approach.

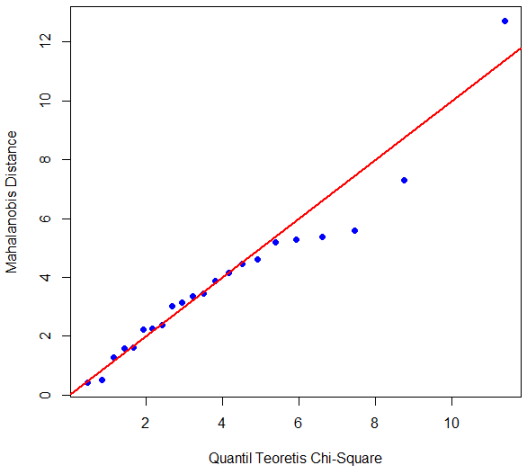

Figure 3. Q-Q Plot of Mahalanobis Distance.

Based on figure 3, the blue points represent the empirical data distribution, while the red line illustrates the theoretical distribution expected under the assumption of multivariate normality. In general, most of the data points lie relatively close to the diagonal line, indicating that the observed pattern aligns well with the Chi-Square distribution. This suggests that the assumption of multivariate normality is largely satisfied. Although a few points deviate slightly on the far-right tail of the plot, such variations remain within acceptable limits and do not indicate severe departures from normality. To further validate this result, the correlation between Mahalanobis distances and the chi-square values will be computed as follows.

Table 3. Mahalanobis Distance Correlation Results.

	Mahalanobis Distance	qi

Mahalanobis Distance	Pearson Correlation	1	0.971
	P-Value		<0.001
	N	22	22

Based on table 3, the correlation coefficient value of 0.971 indicates a very strong association between the variables. Furthermore, the p-value of 0.001 ($p < 0.05$) confirms that this relationship is statistically significant. These results demonstrate that the dataset satisfies the assumption of multivariate normality. Hence, it can be concluded that the overall data meet the multivariate normality requirement and are therefore appropriate for use in clustering analysis.

2. Data Adequacy Assumption

In multivariate analysis, data adequacy is commonly assessed using the Kaiser-Meyer-Olkin (KMO) measure. The KMO index serves to evaluate the suitability of the dataset for further analysis, particularly in factor analysis. The following section presents the results of the data adequacy test based on the KMO index.

Table 4. The result of the KMO test.

X1	X2	X3	X4	Overall MSA
0.53	0.54	0.70	0.54	0.54

Based on table 4, it can be seen that the KMO values for each independent variable and the overall MSA value are greater than 0.5, indicating that the research data are adequate for analysis.

3. Multicollinearity Assumption

The multicollinearity test in this study was conducted to ensure that the estimated regression coefficients can be interpreted accurately. The results of the multicollinearity test were obtained by examining the VIF values.

Table 5. The result of Multicollinearity test.

X1	X2	X3	X4
1.56	1.39	1.18	1.04

Based on table 5 above, it is known that the VIF values of each independent variable do not exceed 10. Therefore, it can be concluded that there is no multicollinearity among the independent variables, or the multicollinearity assumption has been satisfied, and the analysis can proceed to the next stage.

3.3. Clustering Using Gath-Geva Algorithm

This study employs clustering using the Gath–Geva algorithm to identify the optimal number of clusters for grouping the stunting percentages in East Nusa Tenggara Province in 2024, with the results of ANOVA as follows.

Table 6. ANOVA Result.

Variable	F-value	P-value
X1	5.48	0.0132
X2	10.49	0.000851

X3	4.056	0.0341
X4	3.406	0.0544

The one-way ANOVA test was conducted to examine whether the mean differences of each variable were statistically significant across the three clusters. The results show that the poor population (X1) variable yielded an F-value of 5.48 with p = 0.0132, indicating a significant difference among clusters. The household access to proper housing (X2) variable exhibited a highly significant difference (F = 10.49; p < 0.001). Similarly, the open unemployment rate (X3) showed a significant difference with F = 4.06 and p = 0.0341, confirming heterogeneity between clusters at the 95% confidence level. The number of health facilities (X4) variable presented an F-value of 3.41 with p = 0.0544, which is marginally significant, suggesting that differences in healthcare availability among clusters are still statistically meaningful, albeit weaker than other indicators.

Overall, the ANOVA results confirm that most variables (X1–X3) differ significantly among clusters, validating that the Gath–Geva algorithm successfully identified statistically distinct regional groups. These results reinforce the robustness of the clustering structure in representing the socioeconomic disparities related to stunting risk across East Nusa Tenggara Province.

Furthermore, to examine the distinctive characteristics of each cluster and to highlight the differences among them, the final cluster centers are presented in the following table.

Table 7. Final Cluster Center Result.

Cluster	X1	X2	X3	X4
1	15.215	46.547	4.318	116.833
2	24.930	34.789	2.388	69.875
3	19.045	57.341	2.301	89.000

Table 7 displays the final cluster centers obtained from the Gath–Geva algorithm, summarizing the average characteristics of each group. Cluster 1 represents regions with the lowest poverty rate (15.21%) and the highest availability of health facilities (116.83 units). However, these areas still show a moderate level of proper housing access (46.55%) and the highest unemployment rate (4.32%) among clusters. This suggests that even though healthcare infrastructure is relatively good, employment opportunities remain limited. Cluster 2, on the other hand, shows the highest poverty level (24.93%), the lowest access to proper housing (34.79%), and the lowest number of health facilities (69.88 units). This cluster clearly represents the most vulnerable areas, characterized by weak social and health conditions that potentially contribute to higher stunting risks. Cluster 3 is distinguished by the highest access to proper housing (57.34%), the lowest unemployment rate (2.30%), and a moderate number of health facilities (89 units). These regions can be considered the most socioeconomically stable, showing better living conditions compared to the other clusters. Overall, the results indicate that the Gath–Geva clustering effectively differentiated districts based on their socioeconomic disparities, particularly in poverty, housing, and health facility indicators. These distinct groupings can be used to guide targeted policy interventions in East Nusa Tenggara Province.

Table 8. Member of Each Cluster.

Area	Cluster
West Sumba	1
East Sumba	1
Kupang	3
South Central Timor	1
North Central Timor	2
Belu	1

Alor	2
Lembata	2
East Flores	3
Sikka	3
Ende	2
Ngada	3
Manggarai	2
Rote Ndao	2
West Manggarai	3
Central Sumba	1
Southwest Sumba	1
Nagekeo	2
East Manggarai	1
Sabu Raijua	1
Malaka	2
Kupang City	3

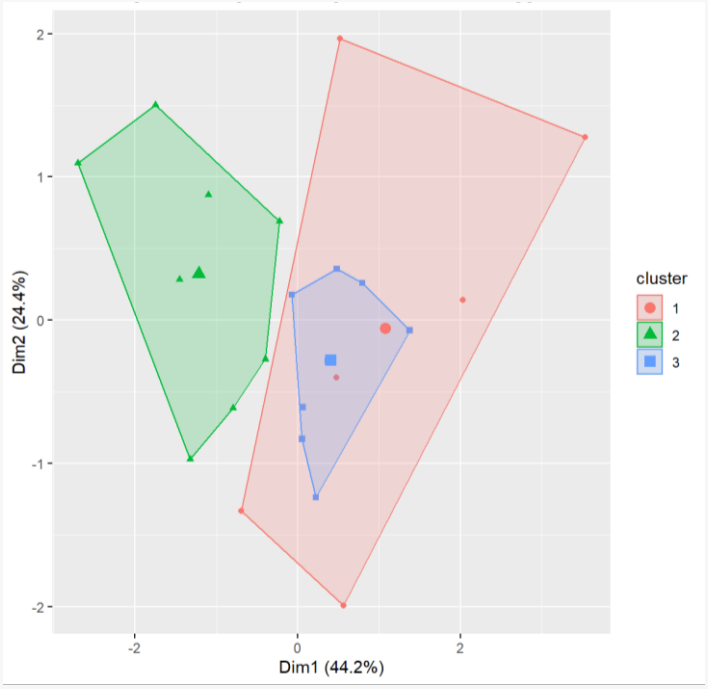

Figure 4. Cluster Plot of Stunting Percentage in East Nusa Tenggara.

Furthermore, based on figure 4 and table 8, the districts/cities belonging to each cluster obtained from the Gath–Geva algorithm can be visualized on the following map.

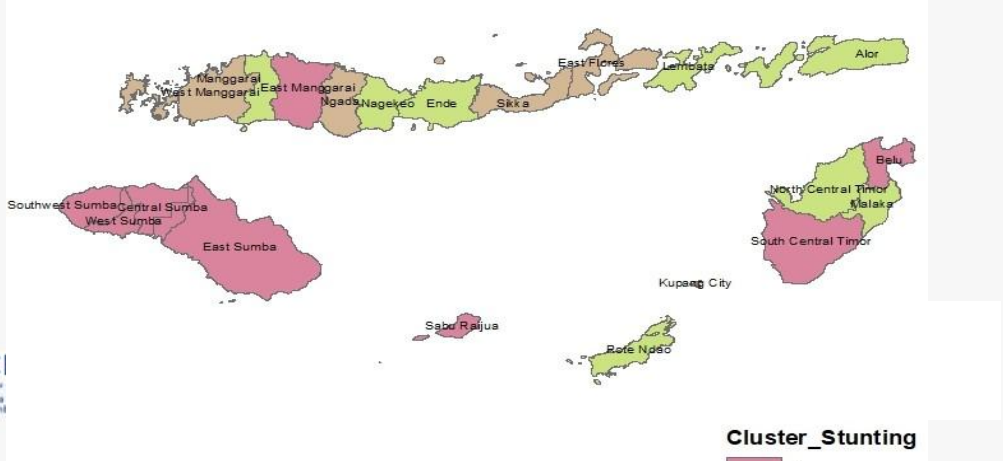

Figure 5. Regional Map Based on Clusters.

Cluster analysis classified the districts in East Nusa Tenggara Province into three main groups with distinct socioeconomic characteristics. The red cluster (Cluster 1) appears on the right side of the plot, representing regions with low poverty levels and a high number of health facilities, yet relatively high unemployment rates. These areas indicate adequate health infrastructure, but economic productivity still requires strengthening. The green cluster (Cluster 2) is clearly separated on the left side, representing the most vulnerable regions, characterized by high poverty, poor housing access, and limited health facilities. The distinct separation of this cluster highlights that its socioeconomic profile is markedly different from the other two groups. Meanwhile, the blue cluster (Cluster 3) is positioned near the center and partially overlaps with the red cluster, indicating transitional regions that share characteristics of both (relatively stable socioeconomic conditions and adequate healthcare access). This overlap reflects the fuzzy nature of the Gath–Geva model, where boundaries between clusters are not rigid but represent gradual transitions between similar regional profiles. Overall, the map demonstrates that the three-cluster solution provides a meaningful representation of regional disparities in East Nusa Tenggara Province, capturing both distinct group separations and natural overlaps among adjacent socioeconomic categories.

The overlapping region between the red (Cluster 1) and blue (Cluster 3) areas indicates that some observations share characteristics of both clusters. This overlap is expected in fuzzy clustering, where each observation can belong to multiple clusters with varying degrees of membership. The blue region appearing within the red area suggests that several districts in Cluster 3 exhibit mixed characteristics, for example, they may share similar health infrastructure levels with Cluster 1 but differ in employment and housing indicators. Overall, this overlap reflects the inherent fuzzy nature of the Gath–Geva model, where the cluster boundaries are soft rather than rigid, representing gradual transitions between regions with partially similar socioeconomic profiles.

Table 9. Comparison of Validity Indices for Each Cluster.

Number of Cluster	Silhouette
1	-0.18
2	0.15
3	0.41

Table 9 presents a comparison of Silhouette index values across different numbers of clusters to evaluate the result. The Silhouette coefficient was employed to evaluate the quality of the clustering results. The Silhouette value ranges from –1 to +1, where values closer to +1 indicate well-defined and clearly separated clusters, while values near 0 or negative suggest overlapping or poorly structured clusters. As shown in the table, the model with one cluster produced a Silhouette score of –0.18, indicating no meaningful cluster structure. The two-cluster model improved slightly to 0.15, suggesting

weak separation among data groups. However, the three-cluster model achieved the highest Silhouette value of 0.41, which falls within the range of a moderately good cluster structure, implying that the data are reasonably well separated into distinct groups. Therefore, the three-cluster solution was selected as the most appropriate, as it provides the best balance between within-cluster cohesion and between-cluster separation, effectively distinguishing regional variations across East Nusa Tenggara Province.

4. Conclusion

This study employed the Gath–Geva (Fuzzy K-Means Entropy) algorithm to classify 22 districts and municipalities in East Nusa Tenggara Province based on socioeconomic and health-related characteristics associated with stunting risk. The results identified three main clusters with distinct profiles. Cluster 1 represents regions with low poverty levels and high health facility availability, yet still facing relatively high unemployment. Cluster 2 corresponds to the most vulnerable areas, characterized by high poverty, limited access to proper housing, and low healthcare availability. Meanwhile, Cluster 3 includes regions with the most stable socioeconomic conditions, featuring better housing access, low unemployment, and adequate health facilities. These findings reveal that stunting vulnerability in East Nusa Tenggara is strongly influenced by socioeconomic disparities and the distribution of health infrastructure. Therefore, stunting intervention strategies should be adapted to the specific characteristics of each cluster. High-risk regions (Cluster 2) require intensive and integrated interventions, while low-risk areas (Cluster 3) should focus on preventive efforts and maintaining existing achievements. The findings of this study provide significant implications for policy formulation and regional planning in stunting reduction efforts. The classification of districts into three clusters offers an evidence-based framework for local governments to prioritize targeted interventions. High-risk regions should receive greater resource allocation and integrated programs focusing on community nutrition, sanitation, and household economic empowerment. Meanwhile, more stable regions should emphasize preventive strategies and the maintenance of existing nutritional achievements. Hence, this cluster-based approach can serve as a practical tool for evidence-driven policymaking, supporting the acceleration of stunting reduction initiatives across East Nusa Tenggara Province.

However, this study has several limitations. The number of analytical units (22 districts) is relatively small, making the clustering results exploratory in nature. The variables used are also limited to socioeconomic and health indicators, excluding environmental, behavioral, and cultural factors that may also contribute to stunting. Additionally, the analysis is cross-sectional, thus not capturing temporal dynamics or changes over time. Future research is recommended to incorporate longitudinal data to examine trends in stunting vulnerability, and to expand the variable scope by including environmental and educational indicators. Moreover, integrating spatial approaches and local policy data would enhance the practical relevance of clustering results, providing a stronger empirical basis for targeted intervention planning and regional stunting reduction strategies.

References

[1] Badan Pusat Statistik Provinsi NTT. 2025, Februari 24. *Jumlah dan persentase balita stunting menurut kabupaten/kota (jiwa).* Retrieved from https://ntt.bps.go.id/id/statistics-table/2/MTQ4OSMy/jumlah-balita-stunting-menurut-kabupaten-kota.html

[2] Badan Pusat Statistik Provinsi NTT. 2023. *Statistik kesehatan Provinsi Nusa Tenggara Timur tahun 2023.* Kupang: BPS.

[3] Choi, J., & Yun, J. I. 2025. Optimization of water chemistry to mitigate corrosion products in nuclear power plants using big data and multiple linear regression in machine. *Progress in Nuclear Energy, 183*, 1–8.

[4] Dalmaijer, E. S., Nord, C. L., & Astle, D. E. 2022. Statistical power for cluster analysis. *BMC Bioinformatics, 23*(205). https://doi.org/10.1186/s12859-022-04675-1

[5] Food and Agriculture Organization (FAO). 2022. *The state of food security and nutrition in the world 2022.* Rome: FAO.

[6] Guyeux, C., Chrétien, S., Tayeh, G. B., Demerjian, J., & Bahi, J. 2019. Introducing and comparing recent clustering methods for massive data management in the Internet of Things. *Journal of Sensor and Actuator Networks, 8*(4), 56. https://doi.org/10.3390/jsan8040056

[7] Cao, Y., Liang, J., Xu, L., & Kang, J. 2024. Testing Multivariate Normality Based on Beta-Representative Points. *Mathematics, 12*(11), 1711. https://doi.org/10.3390/math12111711.

[8] Kementerian Kesehatan Republik Indonesia. 2023. *Laporan nasional survei status gizi Indonesia (SSGI) 2022–2023.* Jakarta: Kementerian Kesehatan Republik Indonesia.
[9] Ministry of Health Republic of Indonesia. 2021. *Indonesia health profile 2021*. Jakarta: Ministry of Health.
[10] O'Brien, R. M. 2007. A caution regarding rules of thumb for variance inflation factor. *Quality & Quantity, 41*, 673–690.
[11] Pakgohar, N., Lengyel, A., & Botta-Dukát, Z. 2024. Quantitative evaluation of internal cluster validation indices using binary data sets. *Journal of Vegetation Science*, 1–13. https://doi.org/10.1111/jvs.13310
[12] Peso'a, N. G., Rais, & Gamayanti, N. F. 2023. Implementation of the Gath-Geva clustering algorithm in the clustering districts/cities in Central Sulawesi based on public health development indicators. *Proceedings of the 4th International Seminar on Science and Technology (ISST 2022)* (pp. 320–328).
[13] Picauly, I., Boeky, D., & Oematan, G. 2024. Factors affecting nutritional status of children under five in Rote Ndao District, Kupang, Nusa Tenggara Timur, Indonesia. *Journal of Maternal and Child Health, 9*(1), 34–45.
[14] Puspa, S. D., Riyono, J., & Puspitasari, F. 2021. Analisis faktor-faktor yang mempengaruhi pemahaman konsep matematis mahasiswa dalam pembelajaran jarak jauh pada masa pandemi Covid-19. *Jurnal Cendekia: Jurnal Pendidikan Matematika, 5*(1), 302–320.
[15] Shamsuddin, A. S., dkk. 2022. A review of spatial analysis application in childhood malnutrition studies. *Malaysian Journal of Medical Sciences.*
[16] Tomarchio, S. D., Luca, B., & Punzo, A. 2023. Model-based clustering using a new multivariate skew distribution. *Advances in Data Analysis and Classification, 18.* https://doi.org/10.1007/s11634-023-00552-8
[17] UNICEF. 2021. *Conceptual framework on the determinants of maternal and child nutrition.* New York: UNICEF.
[18] UNICEF & WHO. 2019. *Progress on household drinking water, sanitation and hygiene 2000–2017*. Geneva: WHO.
[19] World Health Organization. 2023. *Leadership dialogue on food systems for people's nutrition and health.* Retrieved from https://www.who.int/news/item/28-07-2023-leadership-dialogue-on-food-systems-for-people-s-nutrition-and-health
[20] Wulandary, W., & Sudiarti, T. 2024. Stunting on children aged 6–23 months in East Nusa Tenggara Province. *KEMAS: Jurnal Kesehatan Masyarakat, 19*(1), 88–96.