CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 10 (2) (2025), Pages 495-506
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Clustering and Mixture Distribution Analysis of Average Years of Schooling in Papua (2010–2023)

Alvian Sroyer[1*], Henderina Morin[2], Felix Reba[1], Jonathan Wororomi[1], and Agustinus Languwuyo[1]

[1]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Cenderawasih University, Indonesia*
[2]*Department of Government Studies, Faculty of Social Sciences and Government Studies, Cenderawasih University, Indonesia*

Abstract

The purpose of this research is to analyze the distribution of the Human Development Index (HDI) in Papua based on the average years of schooling during the 2010–2023 period using the Gaussian Mixture-based Clustering approach. Data from 28 districts are grouped into five clusters according to their distributional characteristics. Each cluster is modeled using one of the four probability distributions: Inverse Gaussian, Rician, Weibull, or Nakagami. Parameter estimation was performed using the Maximum Likelihood Estimation (MLE) method, and the best distribution for each cluster was selected based on several information criteria (AIC, BIC, AICc, CAIC, and HQC) and validated through Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests. The analysis results show that the Inverse Gaussian distribution fits Cluster 1 and Cluster 3, which represent districts with lower HDI schooling patterns. Cluster 2 is best described by the Rician distribution, indicating moderate HDI variability. The Weibull distribution fits Cluster 4, representing areas with moderately improving education. Cluster 5, with the highest and most stable HDI levels, is best modeled using the Nakagami distribution. The resulting mixture model, combining these four distributions, accurately reflects the HDI distribution patterns across Papua. Policy implications from this study include the development of cluster-based educational strategies tailored to regional characteristics to improve educational equity and human development across the province.

Keywords: Gaussian Mixture Model; Goodness of Fit; HDI Papua; Probability Distribution; Mixture Model

1 Introduction

The Human Development Index (HDI) is a key indicator used to measure the level of human well-being in a region based on three dimensions: health, education, and a decent standard of living [1]–[3]. In Indonesia, the HDI is a crucial tool for both the government and international organizations in evaluating regional development, particularly in identifying inter-regional disparities [4]. Papua, one of Indonesia's most developmentally challenged provinces, demonstrates significant variation in the education dimension, especially in the average years of schooling. Despite improvements in recent years, educational inequality across districts remains a major obstacle to equitable development [5], [6].

[*] **Corresponding author**. E-mail: alvian.sroyer@fmipa.uncen.ac.id

Submitted: May 08, 2025 Reviewed: May 25, 2025 Accepted: June 06, 2025
DOI: https://doi.org/10.18860/cauchy.v10i2.32988

Previous studies have predominantly employed either single-distribution probability models or classical clustering methods such as k-means and hierarchical clustering [7]–[10]. These conventional techniques often fall short in capturing the inherent multimodal and heterogeneous characteristics of educational data. To overcome this limitation, this study introduces a more flexible, dual-layered analytical framework. The first layer applies Gaussian Mixture-based Clustering to uncover latent subgroup structures in the data. The second layer uses Finite Mixture Distribution Models—specifically incorporating the Inverse Gaussian, Rician, Weibull, and Nakagami distributions—to model the intra-cluster distribution of average schooling years [11]–[14].

Unlike previous work that relied on singular approaches, this two-tier method provides a more nuanced and robust understanding of educational disparities in Papua. Gaussian Mixture-based Clustering helps identify hidden district groupings with shared educational profiles, while the finite mixture models capture diverse statistical patterns within each group. These distributions were selected for their complementary strengths: Inverse Gaussian for positively skewed long-tailed data, Rician for location-influenced characteristics, Weibull for reliability-type behaviors, and Nakagami for irregular spread and heavy-tailed patterns [15]–[17].

The main objective of this study is not only to classify Papua's districts by schooling patterns, but also to determine the best-fitting probabilistic model for each group. The insights gained from this approach are expected to guide policymakers, educators, and researchers in developing more targeted and equitable educational policies across Papua's diverse regions.

2 Methods

2.1 Gaussian Mixture-based Clustering

The distribution of Human Development Index (HDI) data in Papua is analyzed using Gaussian Mixture-based Clustering [11]. This method models the heterogeneity of data by combining several Gaussian distributions, where each component corresponds to a specific cluster in the dataset. The formulation is given by:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k) \tag{1}$$

Here, $p(x)$ is the probability density function of the data x, composed of K Gaussian components. Each component has parameters π_k, μ_k, and Σ_k representing the mixing weight, mean, and covariance matrix, respectively. This approach provides high flexibility for describing multi-modal data characteristics.

To evaluate the clustering quality, the Calinski–Harabasz (CH) index is employed, defined as:

$$CH = \frac{\text{SSB}/(k-1)}{\text{SSW}/(n-k)} \tag{2}$$

where SSB and SSW denote the inter-cluster and intra-cluster sum of squares, respectively, and k is the number of clusters. A higher CH index indicates a better-defined clustering structure and assists in selecting the optimal cluster configuration [18].

2.2 Univariate Probability Distribution

This research employs several univariate probability distributions—namely Inverse Gaussian, Rician, Weibull, and Nakagami—to model the variation of average years of schooling in Papua over the 2010–2023 period. These distributions were selected based on their flexibility in representing skewed, heavy-tailed, and multimodal data, which are commonly observed in education-related variables. The mathematical forms of each distribution along with their respective parameters are summarized in Table 1, which serves as the theoretical foundation for subsequent fitting in the cluster analysis.

The Inverse Gaussian distribution is well-suited for modeling positively skewed data such as schooling years, particularly when values are bounded away from zero. The Rician distribution accommodates signals or measures with dominant modes and has been used in similar educational and environmental studies. The Weibull distribution has been widely applied due to its parameter adaptability, enabling accurate modeling of various shapes and scales, especially in wind energy and reliability contexts [12]–[14]. The Nakagami distribution is effective in capturing irregularities and tail behavior, making it useful for modeling educational disparities across regions [19].

Table 1: Probability Distribution Models and Their Parameters Used in Cluster Analysis

Distribution	**Distribution Model**	**Parameter**
Inverse Gaussian	$f(x)=\sqrt{\frac{\lambda}{2\pi x^3}}\exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right),\quad x>0$	$\mu=$ Mean, $\lambda=$ Shape
Rician	$f(x)=\frac{x}{\sigma^2}\exp\left(-\frac{x^2+s^2}{2\sigma^2}\right)I_0\left(\frac{xs}{\sigma^2}\right),\quad x>0$	$s=$ Location, $\sigma=$ Scale
Weibull	$f(x)=\frac{b}{a}\left(\frac{x}{a}\right)^{b-1}\exp\left(-\left(\frac{x}{a}\right)^{b}\right),\quad x\geq 0$	$a=$ Scale, $b=$ Shape
Nakagami	$f(x)=2\left(\frac{\mu}{\omega}\right)^{\mu}\frac{1}{\Gamma(\mu)}x^{2\mu-1}\exp\left(-\frac{\mu}{\omega}x^2\right),\quad x>0$	$\mu=$ Shape, $\omega=$ Scale

Each of these distributions plays a role in identifying the best-fitting probabilistic model for each cluster derived from the Gaussian Mixture-based Clustering. By integrating these distributions into the cluster analysis process, the study enhances its capacity to characterize distinct regional schooling patterns and identify underlying statistical behaviors of educational disparity in Papua. The application of Weibull and Nakagami in particular is reinforced by various prior studies focusing on their goodness-of-fit and flexibility in modeling regional variations and environmental metrics [15], [19], [20].

2.3 Goodness-of-Fit

Evaluation of the suitability of the distribution model with the observation data was carried out using the Goodness-of-Fit (GoF) tests, namely Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) [21], [22]. The KS test is used to calculate the maximum distance between the cumulative distribution function of sample data $F(x)$ and the assumed distribution model $H(x)$, as stated in the following formula [23], [24]:

$$D_{\text{count}}=\sup_{x}|F(x)-H(x)| \tag{3}$$

The above equation describes the calculation of the maximum distance between the cumulative distribution function of the sample data and the assumed distribution model. A smaller value of the KS statistic indicates a better fit between the model and the data. If this value is small enough, the hypothesis that the data follows the proposed distribution model is accepted [21], [22], [25].

$$A_n^2=-n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log F(x_i)+\log(1-F(x_{n+1-i}))\right] \tag{4}$$

The AD test is very useful in assessing the distribution pattern of years of schooling, especially for capturing the extremes of the data [21], [23].

2.4 Finite Mixture Distribution Models

The probability mixture model is constructed by combining several univariate probability distributions—Inverse Gaussian, Rician, Weibull, and Nakagami—with respective contribution weights denoted as π_1,π_2,π_3,π_4. These weights indicate the proportion of influence each distribution has on the overall model and must satisfy the constraint $\pi_1+\pi_2+\pi_3+\pi_4=1$ with $\pi_i\geq 0$ for all i.

The mathematical formulation of the mixture model is expressed as:

$$f(y|\Theta)=\pi_1 f_{\text{IG}}(y|\theta_1)+\pi_2 f_{\text{Rician}}(y|\theta_2)+\pi_3 f_{\text{Weibull}}(y|\theta_3)+\pi_4 f_{\text{Nakagami}}(y|\theta_4) \tag{5}$$

The mixture model components consist of four distinct distributions: the Inverse Gaussian distribution $f_{\mathrm{IG}}(y|\theta_1)$ with parameters $\theta_1 = (\mu, \lambda)$, the Rician distribution $f_{\mathrm{Rician}}(y|\theta_2)$ with parameters $\theta_2 = (s, \sigma)$, the Weibull distribution $f_{\mathrm{Weibull}}(y|\theta_3)$ with parameters $\theta_3 = (a, b)$, and the Nakagami distribution $f_{\mathrm{Nakagami}}(y|\theta_4)$ with parameters $\theta_4 = (\mu, \omega)$.

These mixture components are capable of representing a wide range of data patterns. The Inverse Gaussian distribution is particularly effective for modeling positively skewed data with long tails [16], while the Rician distribution handles location-dependent phenomena with shape control [23]. The Weibull distribution is widely recognized in reliability and lifetime modeling due to its flexible shape parameter [20], and the Nakagami distribution has shown effectiveness in modeling positively valued, skewed data with variable spread [19].

2.5 Distribution Selection Criteria

To determine the most optimal probability distribution model, a number of information criteria are used, including the Akaike Information Criterion (AIC) [26], Bayesian Information Criterion (BIC) [27], Corrected Akaike Information Criterion (AICc) [28], Consistent Akaike Information Criterion (CAIC) [29], and Hannan-Quinn Criterion (HQC) [30]. The model with the lowest criterion value is considered to be the model that best fits the data:

1. $$AIC = -2\ln(L) + 2k \tag{6}$$
2. $$BIC = -2\ln(L) + k\ln(n) \tag{7}$$
3. $$AICc = AIC + \left(\frac{2k(k+1)}{n-k-1}\right) \tag{8}$$
4. $$CAIC = -2\ln(L) + k\left(1 + \ln\left(\frac{n}{k}\right)\right) \tag{9}$$
5. $$HQC = -2\ln(L) + 2k\ln(\ln(n)) \tag{10}$$

The use of these criteria ensures that the selected distribution model not only fits the data, but also remains simple to avoid the risk of overfitting.

3 Results and Discussion

3.1 Clustering and Evaluation of HDI

This study adopts a dual-layer analytical approach to assess educational disparities across the 28 districts in Papua from 2010 to 2023. The first layer implements Gaussian Mixture-based Clustering [11], which is capable of modeling the multimodal and heterogeneous nature of the data, offering a more flexible alternative to classical clustering methods [7], [8].

To evaluate the effectiveness of the clustering, the Calinski–Harabasz Index (CHI) was employed [18], which quantifies the ratio between inter-cluster and intra-cluster dispersion. The results indicate that the optimal number of clusters is $k = 5$, as it yields the highest CHI value. Each of these five clusters is characterized by a distinct schooling pattern and is further analyzed using Finite Mixture Distribution Models, including Inverse Gaussian, Rician, Weibull, and Nakagami distributions [12], [13], [19], [20], selected based on their capability to represent skewed, heavy-tailed, and multimodal characteristics in the data.

Table 2: Gaussian Mixture Clustering Results and CHI Evaluation

(a) Calinski–Harabasz Index (CHI) for Different k

k	2	3	4	**5**	6	7	8	9	10	11	12	13	14	15
CHI	132.98	140.27	145.21	**174.00**	150.20	109.81	121.01	102.38	109.45	109.50	88.72	73.67	67.05	57.57

(b) Cluster Mean Values per Year (for $k = 5$)

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Cluster 1	4.46	4.59	4.62	4.70	4.79	4.98	5.13	5.23	5.32	5.46	5.62	5.76	4.46	4.59
Cluster 2	7.08	7.46	7.66	7.80	7.91	8.08	8.31	8.45	8.63	8.65	8.92	9.13	7.08	7.46
Cluster 3	1.61	1.67	1.80	1.89	1.99	2.07	2.17	2.28	2.48	2.68	2.82	2.92	1.61	1.67
Cluster 4	2.64	3.01	3.12	3.18	3.29	3.43	3.46	3.49	3.54	3.66	3.76	3.84	2.64	3.01
Cluster 5	8.71	8.82	9.16	9.25	9.31	9.32	9.47	9.67	9.87	9.88	10.04	10.13	8.71	8.82

Based on the results in Table 2, the clustering reveals diverse educational profiles among districts. Cluster 5 consistently exhibits the highest average years of schooling, indicating districts with the most favorable educational conditions. In contrast, Cluster 3 shows the lowest average, reflecting significant educational challenges. Each cluster was subsequently modeled using the most appropriate distribution: Cluster 1 with *Inverse Gaussian*, Cluster 2 with *Rician*, Cluster 3 again with *Inverse Gaussian*, Cluster 4 with *Weibull*, and Cluster 5 with *Nakagami* distribution. This integration of Gaussian-based clustering with finite mixture distribution modeling allows a comprehensive and statistically grounded understanding of regional education patterns in Papua.

3.2 Parameter Estimation and Goodness-of-Fit Test

Following the clustering analysis using the Gaussian Mixture-based Clustering and evaluation with the Calinski-Harabasz Index (CHI), the next step involves assessing the suitability of data in each cluster to specific probability distributions. Parameter estimation was conducted using the Maximum Likelihood Estimation (MLE) method, as defined in Equation (6). The distributions considered include Inverse Gaussian, Rician, Weibull, and Nakagami, selected due to their flexibility in modeling skewed, heavy-tailed, and location-sensitive data characteristics.

Each cluster's assigned distribution was fitted using MLE, and the results are shown in Table 3. These distributions are characterized by different parameter sets: shape, scale, and location (when applicable), which capture the unique statistical structure of each cluster.

Table 3: Parameter Estimation for Each Distribution

Distribution	**Cluster**	**Parameter**		
		Shape	**Scale**	**Location**
Inverse Gaussian	Cluster 1	756.6024	5.0544	–
Rician	Cluster 2	–	0.5911	8.1510
Inverse Gaussian	Cluster 3	58.8453	2.1968	–
Weibull	Cluster 4	3.5101	12.7884	–
Nakagami	Cluster 5	115.7179	89.8349	–

To validate the goodness-of-fit of these models, two statistical tests were applied: the Kolmogorov Smirnov (KS) and Anderson–Darling (AD) tests [20]. These tests assess whether the observed data distributions are statistically consistent with the fitted distributions under the null hypothesis H_0 (that the data follow the proposed distribution).

The KS test evaluates the maximum deviation between the empirical and theoretical cumulative distribution functions, while the AD test places more weight on the tails of the distribution, which is particularly useful for education data prone to extreme values.

Table 4 displays the p-values and test statistics for each distribution under both KS and AD tests, along with the corresponding critical values (CV).

Table 4: Goodness-of-Fit Test Results Using KS and AD for Each Cluster

Cluster	Distribution	KS H_0/H_1	p-val	KS Stat	KS CV	AD H_0/H_1	p-val	AD Stat
1	Inverse Gaussian	H_0	0.926	0.147	0.375	H_0	0.959	0.269
2	Rician	H_0	0.993	0.113	0.375	H_0	0.999	0.148
3	Inverse Gaussian	H_0	0.988	0.119	0.375	H_0	0.987	0.214
4	Weibull	H_0	0.996	0.108	0.375	H_0	1.000	0.134
5	Nakagami	H_0	0.909	0.151	0.375	H_0	0.961	0.267

The results in Table 3 confirm that all selected distributions passed the KS and AD tests at conventional significance levels, validating their appropriateness. Notably, Cluster 5 represented by the Nakagami distribution has both a high shape parameter (115.71) and scale (89.83), capturing districts with consistently high HDI schooling values. Cluster 3, with the Inverse Gaussian distribution, captures high variability in lower-HDI regions, while Cluster 4 (Weibull) describes moderate educational patterns. Rician and Weibull also provide excellent fits in Clusters 2 and 4 respectively, based on their strong *p*-values and low test statistics.

These findings reinforce the robustness of the model and provide a solid statistical foundation for further policy analysis and evidence-based planning at the regional level.

3.3 Best Distribution Model Information Criteria

To determine the best probability distribution model for each cluster, this study utilized several information-theoretic criteria, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Corrected Akaike Information Criterion (AICc), Consistent Akaike Information Criterion (CAIC), and the Hannan–Quinn Criterion (HQC). These criteria help identify the most appropriate distribution by balancing model fit and complexity.

Each distribution model was fitted using the Maximum Likelihood Estimation (MLE) method, and the corresponding values of these criteria were calculated. The distribution model with the lowest criterion value in each cluster is considered the most appropriate. Table 5 presents the calculated values for AIC, BIC, AICc, CAIC, and HQC for the four tested distributions (Inverse Gaussian, Rician, Weibull, and Nakagami) across five clusters.

Table 5: Information Criteria (AIC, BIC, AICc, CAIC, HQC) per Cluster and Distribution

Cluster	Distribution	AIC	BIC	AICc	CAIC	HQC
1	Inverse Gaussian	16.718	17.688	19.242	19.399	16.359
2	Rician	25.404	26.374	77.752	77.910	25.045
3	Inverse Gaussian	16.824	17.794	19.267	19.425	16.465
4	Weibull	10.035	11.004	11.368	11.526	9.676
5	Nakagami	18.363	19.333	19.717	19.874	18.004

Based on the results in Table 5, the distribution model with the lowest information criterion value in each cluster is identified as the best-fitting model:

- In **Cluster 1**, the **Inverse Gaussian** distribution has the lowest AIC and HQC values, making it the most suitable distribution for this cluster.
- In **Cluster 2**, the **Rician** distribution consistently yields the lowest values across all criteria, indicating it is the best-fitting distribution.
- In **Cluster 3**, the **Inverse Gaussian** distribution again shows the lowest AIC and HQC values, indicating its superiority.
- In **Cluster 4**, the **Weibull** distribution demonstrates the best performance with the lowest values for all criteria.
- In **Cluster 5**, the **Nakagami** distribution has the lowest criterion values and is therefore considered the most appropriate.

These results confirm that the information-criteria-based approach offers a systematic and objective method for selecting the most appropriate distribution model. This methodology ensures both statistical accuracy and parsimony, and serves as a strong basis for further analysis and data-driven policy development.

3.4 Mixture Model of HDI based on Regional Clusters

The distribution of average years of schooling in Papua during the 2010–2023 period was analyzed using the Gaussian Mixture-based Clustering approach [11], which resulted in five regional clusters based on the data distribution patterns. The evaluation using the Calinski–Harabasz Index (CHI) confirmed that the five-cluster solution was the most optimal, as shown in Table 2, while the parameter estimates for each cluster are presented in Table 3.

Cluster 1 (Inverse Gaussian) includes the districts of Jayawijaya, Paniai, Mappi, Asmat, Mamberamo Raya, and Dogiyai. These areas exhibit very low average years of schooling. This cluster is best modeled by the Inverse Gaussian distribution with a shape parameter of 756.6024 and a scale parameter of 5.0544. **Cluster 2 (Rician)** comprises the districts of Merauke, Boven Digoel, Sarmi, Keerom, and Supiori. This cluster displays a broad distribution pattern with a pronounced peak in the middle values and is best modeled by the Rician distribution, which has a scale parameter of 0.5911 and a location parameter of 8.1510. **Cluster 3 (Inverse Gaussian)** includes the districts of Pegunungan Bintang, Nduga, Mamberamo Tengah, Yalimo, Puncak, and Intan Jaya. These regions tend to exhibit a concentrated distribution at very low values and are best described by the Inverse Gaussian distribution with a shape parameter of 58.8453 and a scale parameter of 2.1968. **Cluster 4 (Weibull)** covers the districts of Puncak Jaya, Yahukimo, Tolikara, Lanny Jaya, and Deiyai. The Weibull distribution with a shape parameter of 3.5101 and a scale parameter of 12.7884 is suitable to represent the moderately increasing educational trend in this cluster. **Cluster 5 (Nakagami)** consists of Jayapura, Nabire, Yapen Islands, Biak Numfor, Mimika, and Waropen. These areas demonstrate high and stable HDI levels, and the Nakagami distribution with a shape parameter of 115.7179 and a scale parameter of 89.8349 provides the best fit.

The mixture model used in this study is a Finite Mixture Distribution Model, which combines the best probability distribution for each cluster. The overall model is expressed as follows:

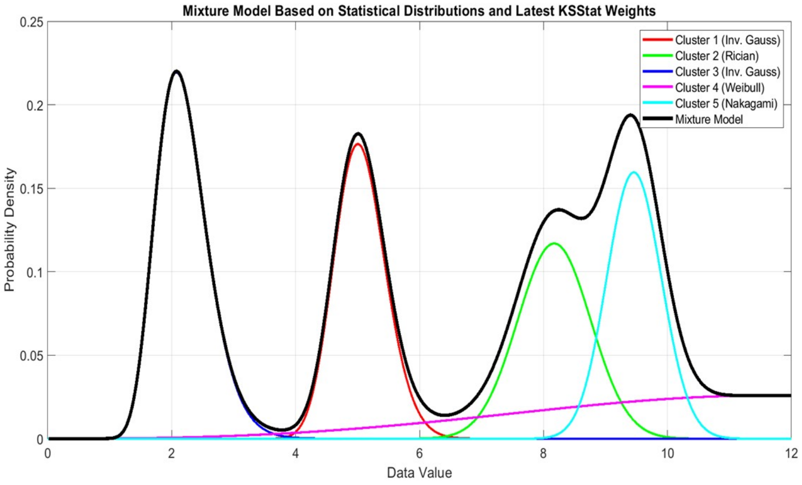
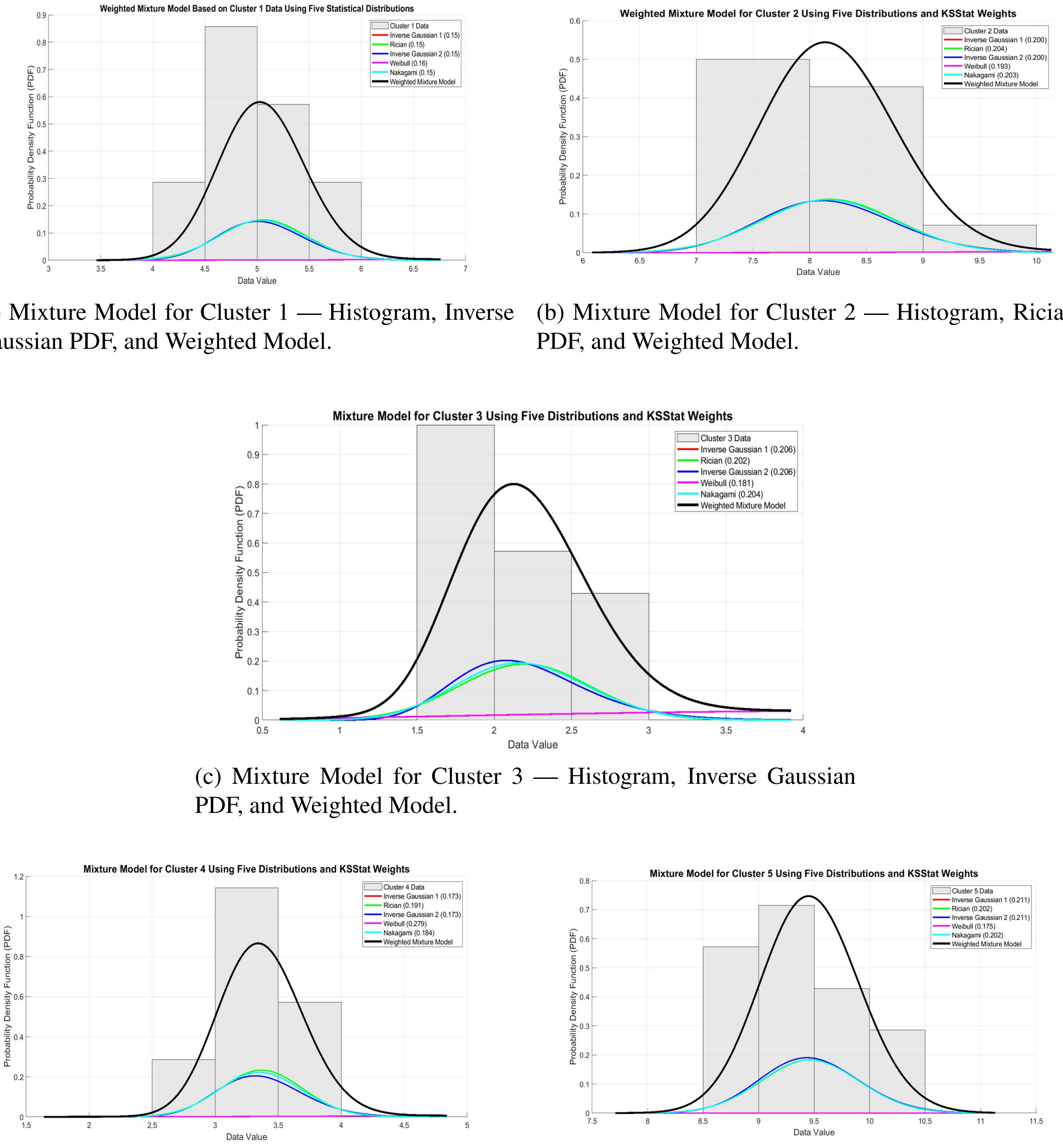


Figure 1: Mixture Model for HDI Data in Papua Province (2010–2023)

Figure 1 illustrates the mixture curve composed of five probability distributions derived from the clustering results using the Gaussian Mixture-based Clustering approach. Each curve represents the best-fitting distribution for each cluster, based on the parameter estimation results in Table 3 and the optimal cluster structure determined by the Calinski–Harabasz Index (CHI) in Table 2.

This mixture model is constructed using the Finite Mixture Distribution Model approach, where the mixing weights for each distribution are determined based on the Kolmogorov–Smirnov (KS) *Goodness-of-Fit* test values, indicating the closeness of the distribution to the actual data. Although the legend in Figure 1 does not explicitly display the weights of each distribution, KS-based weighting has been applied in the modeling process to generate the most optimal combined model curve.

For details on the weights and complete visualizations that include this information, refer to Figure 2, which presents the mixture model graphs for each cluster. These include the actual data histograms, the dominant probability distributions, and the explicit contribution of each weight in the combined model, as determined by the KS results.



(a) Mixture Model for Cluster 1 — Histogram, Inverse Gaussian PDF, and Weighted Model.

(b) Mixture Model for Cluster 2 — Histogram, Rician PDF, and Weighted Model.

(c) Mixture Model for Cluster 3 — Histogram, Inverse Gaussian PDF, and Weighted Model.

(d) Mixture Model for Cluster 4 — Histogram, Weibull PDF, and Weighted Model.

(e) Mixture Model for Cluster 5 — Histogram, Nakagami PDF, and Weighted Model.

Figure 2: Mixture model visualizations for each cluster. Each subfigure includes the cluster's histogram, the dominant fitted distribution, and the complete mixture model based on KSStat weighting.

This model effectively captures the HDI distribution patterns across Papua by reflecting the distinct statistical characteristics of each cluster. This approach enables more in-depth spatial analysis and can serve as a foundation for regionally tailored policy formulation.

3.5 Relevance and Implications

The results of the mixture model analysis on Papua's HDI data from 2010 to 2023 provide valuable insights into the diversity of human development across 28 districts. By applying the *Gaussian Mixture-based Clustering* approach, the districts are optimally divided into five distinct clusters, as supported by the Calinski–Harabasz Index (CHI). Each cluster is modeled using the best-fitting statistical distribution—namely Inverse Gaussian, Rician, Weibull, and Nakagami—as presented in Table 3.

This clustering approach allows for a deeper understanding of spatial disparities in education and HDI outcomes. Each cluster reflects distinct characteristics in average years of schooling, ranging from very low (Cluster 1 and 3 with Inverse Gaussian distribution), moderate (Cluster 2 with Rician), gradual improvement (Cluster 4 with Weibull), to high and stable education levels (Cluster 5 with Nakagami).

The implication is that policymakers and local governments can use this model to allocate education resources more effectively and equitably. It enables the design of tailored interventions for each region's specific context. Furthermore, this modeling framework demonstrates the adaptability of statistical approaches for heterogeneous development data and offers a replicable reference for regional analysis in other provinces or countries.

3.6 Policy Implications Based on Analysis Results

The findings from the finite mixture model analysis form a strong foundation for data-driven policy formulation in the education and human development sectors in Papua:

- **Cluster 1 (Inverse Gaussian)**: Includes Jayawijaya, Paniai, Mappi, Asmat, Mamberamo Raya, and Dogiyai. These districts require urgent interventions to improve access to basic education, reduce dropout rates, and expand school infrastructure in remote areas.
- **Cluster 2 (Rician)**: Comprising Merauke, Boven Digoel, Sarmi, Keerom, and Supiori, this cluster requires targeted teacher training, enrichment of local curricula, and support for students in regions with moderately variable HDI.
- **Cluster 3 (Inverse Gaussian)**: Includes Pegunungan Bintang, Nduga, Mamberamo Tengah, Yalimo, Puncak, and Intan Jaya. These areas show extremely low HDI concentration and should be prioritized for foundational education services, literacy programs, and community-based schooling initiatives.
- **Cluster 4 (Weibull)**: Covering Puncak Jaya, Yahukimo, Tolikara, Lanny Jaya, and Deiyai, this cluster benefits from moderately rising HDI trends and thus calls for continued investment in quality education and gradual scaling of digital learning programs.
- **Cluster 5 (Nakagami)**: Consists of Jayapura, Nabire, Yapen Islands, Biak Numfor, Mimika, and Waropen. With high and stable HDI, policies in this cluster can emphasize innovation, human capital development, and the integration of education with economic transformation.

This cluster-specific policy framework ensures targeted planning and more equitable development across all educational levels, supporting both short-term improvement and long-term sustainability in Papua's human development.

4 Conclusion

This study has introduced a comprehensive two-layer analytical approach to explore the distribution patterns of the Human Development Index (HDI) in Papua based on the average years of schooling from 2010 to 2023. Through the use of Gaussian Mixture-based Clustering, 28 districts were optimally grouped into five regional clusters with distinct educational profiles. Subsequently, each cluster was modeled using the most appropriate univariate distribution—Inverse Gaussian, Rician, Weibull, or Nakagami—based on multiple information criteria and validated with Goodness-of-Fit (GoF) tests.

The resulting Finite Mixture Distribution Model successfully captured the multimodal and heterogeneous nature of educational disparities across Papua. By leveraging statistical flexibility and robust distribution fitting, the model offers not only high accuracy but also interpretability in complex, real-world datasets. Each distribution component reflects a unique structural characteristic of its respective cluster, including tail behavior, skewness, and variability.

Unlike traditional methods, this mixture-based framework offers greater resolution in identifying latent patterns and subgroup characteristics in HDI-related data. The integration of model selection criteria (AIC, BIC, AICc, CAIC, HQC) with empirical validation (KS and AD tests) strengthens the scientific validity and replicability of the findings.

Beyond regional application, this model provides a scalable methodology that can be replicated in other provinces or countries facing similar developmental disparities. It bridges the gap between theoretical statistical modeling and practical policy planning by enabling more nuanced, data-driven, and regionally adapted interventions. Ultimately, the research contributes to the advancement of statistical tools for equitable human development strategy formulation in diverse sociogeographic contexts.

CRediT Authorship Contribution Statement

Alvian Sroyer: Conceptualization, Methodology, Formal analysis, Writing - original draft. **Henderina Morin:** Validation, Investigation, Resources, Writing - review & editing. **Felix Reba:** Data curation, Software, Visualization. **Jonathan Wororomi:** Supervision, Funding acquisition, Project administration. **Agustinus Languwuyo:** Writing - review & editing.

Declaration of Generative AI and AI-assisted technologies

During the preparation of this work, Generative AI tools (specifically ChatGPT) were used solely for language refinement and grammar editing purposes. No part of the analysis, interpretation, or core content was generated by AI. All substantive results and discussions were manually developed by the authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding and Acknowledgments

We would like to express our sincere appreciation to the Institute for Research and Community Service (LPPM) of Cenderawasih University (Uncen) for the 2024 Research permission. This support has been essential for the successful completion of this research project. Without this valuable support, this research would not have been possible.

Data Availability

The dataset used in this study, consisting of average years of schooling in 28 districts in Papua from 2010 to 2023, was obtained from the Central Statistics Agency (BPS) of Papua Province. The data is publicly available and can be accessed through the official publications and statistical databases of BPS Papua. Further details regarding the data are discussed in Section 3.1 of this paper.

References

[1] R. Karagiannis and G. Karagiannis, “Constructing composite indicators with shannon entropy: The case of human development index,” *Socio-Economic Planning Sciences*, vol. 70, p. 100 701, 2020. DOI: 10.1016/j.seps.2019.03.007.

[2] C. Türe and Y. Türe, “A model for the sustainability assessment based on the human development index in districts of megacity istanbul (turkey),” *Environment, Development and Sustainability*, vol. 23, pp. 3623–3637, 2021. DOI: 10.1007/s10668-020-00735-9.

[3] Y. Jiang, H. Liang, C. Shi, and H. Xia, “Estimating sustainability and regional inequalities using an enhanced sustainable development index in china,” *Sustainable Cities and Society*, vol. 99, p. 104 555, 2023. DOI: 10.1016/j.scs.2023.104555.

[4] S. Kwatra, A. Kumar, and P. Sharma, “A critical review of studies related to construction and computation of sustainable development indices,” *Ecological Indicators*, vol. 112, p. 106 061, 2020. DOI: 10.1016/j.ecolind.2019.106061.

[5] W. Afalia, I. Hamda, S. Adriana, A. Alamsyah, and N. Wafiroh, *Determinants of human development index in papua province 2012-2021. wiga: Jurnal penelitian ilmu ekonomi, 13 (2), 246–256*, 2023. DOI: 10.30741/wiga.v13i2.1076.

[6] D. Rahmawati, I. Budiantara, D. Prastyo, and M. Octavanny, “Modeling of human development index in papua province using spline smoothing estimator in nonparametric regression,” in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1752, 2021, p. 012 018. DOI: 10.1088/1742-6596/1752/1/012018.

[7] D. A. N. Sirodj, I. M. Sumertajaya, and A. Kurnia, “Analisis clustering time series untuk pengelompokan provinsi di indonesia berdasarkan indeks pembangunan manusia jenis kelamin perempuan,” *Statistika*, vol. 23, no. 1, pp. 29–37, 2023. DOI: 10.29313/statistika.v23i1.2181.

[8] A. M. Sikana and A. W. Wijayanto, “Analisis perbandingan pengelompokan indeks pembangunan manusia indonesia tahun 2019 dengan metode partitioning dan hierarchical clustering,” *Jurnal Ilmu Komputer*, vol. 14, no. 2, pp. 66–78, 2021. DOI: 10.24843/JIK.2021.v14.i02.p01.

[9] E. Luthfi and A. W. Wijayanto, “Analisis perbandingan metode hirearchical, k-means, dan k-medoids clustering dalam pengelompokkan indeks pembangunan manusia indonesia,” *Inovasi*, vol. 17, no. 4, pp. 761–773, 2021. DOI: 10.34010/komputika.v13i1.10812.

[10] F. Fajriani *et al.*, “K-means clustering analysis pada persebaran tingkat pengangguran kabupaten/kota di sulawesi selatan,” *Jurnal Varian*, vol. 3, no. 2, pp. 103–112, 2020. DOI: 10.14710/tataloka.25.2.121-132.

[11] S. Kageyama, N. Mori, S. Mugikura, H. Tokunaga, and K. Takase, “Gaussian mixture model-based cluster analysis of apparent diffusion coefficient values: A novel approach to evaluate uterine endometrioid carcinoma grade,” *European Radiology*, vol. 31, pp. 55–64, 2021. DOI: 10.1007/s00330-020-07047-6.

[12] M. Krit, O. Gaudoin, M. Xie, and E. Remy, “Simplified likelihood based goodness-of-fit tests for the weibull distribution,” *Communications in Statistics-Simulation and Computation*, vol. 45, no. 3, pp. 920–951, 2016. DOI: 10.1080/03610918.2013.879889.

[13] M. H. Ouahabi, H. Elkhachine, F. Benabdelouahab, and A. Khamlichi, “Comparative study of five different methods of adjustment by the weibull model to determine the most accurate method of analyzing annual variations of wind energy in tetouan-morocco.,” *Procedia Manufacturing*, vol. 46, pp. 698–707, 2020. DOI: 10.1016/j.promfg.2020.03.099.

[14] P. A. C. Rocha, R. C. de Sousa, C. F. de Andrade, and M. E. V. da Silva, “Comparison of seven numerical methods for determining weibull parameters for wind energy generation in the northeast region of brazil,” *Applied Energy*, vol. 89, no. 1, pp. 395–400, 2012. DOI: 10.1016/j.apenergy.2011.08.003.

[15] M. Nassar, A. Alzaatreh, M. Mead, and O. Abo-Kasem, “Alpha power weibull distribution: Properties and applications,” *Communications in Statistics-Theory and Methods*, vol. 46, no. 20, pp. 10 236–10 252, 2017. DOI: 10.1080/03610926.2016.1231816.

[16] A. M. Basheer, “Alpha power inverse weibull distribution with reliability application,” *Journal of Taibah University for Science*, vol. 13, no. 1, pp. 423–432, 2019. DOI: 10.1080/16583655.2019.1588488.

[17] Q. Han, S. Ma, T. Wang, and F. Chu, “Kernel density estimation model for wind speed probability distribution with applicability to wind energy assessment in china,” *Renewable and Sustainable Energy Reviews*, vol. 115, p. 109 387, 2019. DOI: 10.1016/j.rser.2019.109387.

[18] M. Kozak, ““a dendrite method for cluster analysis” by caliński and harabasz: A classical work that is far too often incorrectly cited,” *Communications in Statistics-theory and Methods*, vol. 41, no. 12, pp. 2279–2280, 2012. DOI: 10.1080/03610927408827101.

[19] M. Samuh and A. Salhab, “Distribution of squared sum of products of independent nakagami-m random variables,” *Communications in Statistics-Simulation and Computation*, vol. 53, no. 12, pp. 6457–6470, 2024. DOI: 10.1080/03610918.2023.2234668.

[20] M. Krit, O. Gaudoin, and E. Remy, “Goodness-of-fit tests for the weibull and extreme value distributions: A review and comparative study,” *Communications in Statistics-Simulation and Computation*, vol. 50, no. 7, pp. 1888–1911, 2021. DOI: 10.1080/03610918.2019.1594292.

[21] B. Yazici and S. Yolacan, “A comparison of various tests of normality,” *Journal of Statistical Computation and Simulation*, vol. 77, no. 2, pp. 175–183, 2007. DOI: 10.1080/10629360600678310.

[22] A. K. Mbah and A. Paothong, “Shapiro–francia test compared to other normality test using expected p-value,” *Journal of Statistical Computation and Simulation*, vol. 85, no. 15, pp. 3002–3016, 2015. DOI: 10.1080/00949655.2014.947986.

[23] M. M. Badr, “Goodness-of-fit tests for the compound rayleigh distribution with application to real data,” *Heliyon*, vol. 5, no. 8, 2019. DOI: 10.1016/j.heliyon.2019.e02225.

[24] S. Dey, D. Kumar, P. L. Ramos, and F. Louzada, “Exponentiated chen distribution: Properties and estimation,” *Communications in Statistics-Simulation and Computation*, vol. 46, no. 10, pp. 8118–8139, 2017. DOI: 10.1080/03610918.2016.1267752.

[25] B. Husam, “Tests of normality: New test and comparative study,” *Communications in Statistics-Simulation and Computation*, 2019. DOI: 10.1080/03610918.2019.1643883.

[26] S. Portet, “A primer on model selection using the akaike information criterion,” *Infectious Disease Modelling*, vol. 5, pp. 111–128, 2020. DOI: 10.1016/j.idm.2019.12.010.

[27] M. Drton and M. Plummer, “A bayesian information criterion for singular models,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 79, no. 2, pp. 323–380, 2017. DOI: 10.1111/rssb.12187.

[28] T. Matsuda, “Inadmissibility of the corrected akaike information criterion,” *Bernoulli*, vol. 30, no. 2, pp. 1416–1440, 2024. DOI: 10.3150/23-BEJ1638.

[29] D. Anderson, K. Burnham, and G. White, “Comparison of akaike information criterion and consistent akaike information criterion for model selection and statistical inference from capture-recapture studies,” *Journal of Applied Statistics*, vol. 25, no. 2, pp. 263–282, 1998. DOI: 10.1080/02664769823250.

[30] L. Lopez and S. Weber, “Testing for granger causality in panel data,” *The Stata Journal*, vol. 17, no. 4, pp. 972–984, 2017. DOI: 10.1177/1536867X1801700412.