ABSTRACT
Cluster analysis involves the methodical categorization of data based on the degree of similarity within each group to group data with similar characteristics. This study focuses on classifying poverty data across Indonesian provinces. The methodologies employed include the Fuzzy C-Means (FCM) and Fuzzy Probabilistic C-Means (FPCM) algorithms. The FCM algorithm is a clustering approach where membership values determine the presence of each data point in a cluster. On the other hand, the FPCM algorithm builds upon FCM and Probabilistic C (PCM) algorithms by incorporating probabilistic considerations. This research compares the FCM and FPCM algorithms using local poverty data from Indonesia, specifically examining the Partition Entropy (PE) index value. It aims to identify the optimal number of clusters for provincial-level poverty data in Indonesia. The findings indicate that the FPCM algorithm outperforms the FCM algorithm in categorizing poverty in Indonesia, as evidenced by the PE validity index. Furthermore, the study identifies that the ideal number of clusters for the data is 2.

Keywords:
Cluster Analysis; Fuzzy C-Means; Fuzzy Probabilistic C-Means; Partition Entropy.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.
1. INTRODUCTION

Poverty is a significant global issue affecting economic, cultural, geographical, and social aspects [1] [2]. The 2022 Multidimensional Poverty Index (MPI) report highlighted that over 1.2 billion people in 111 countries live in poverty. A recent reevaluation by Indonesia's Central Statistics Agency (BPS) in March 2022 found that 26.36 million people in Indonesia were living in poverty. Although there have been claims of a steady decline in poverty since 2011, World Bank data suggests otherwise. According to the World Bank, the number of impoverished people in Indonesia rose from 54 million to 67 million over the same period. This discrepancy arises from different poverty threshold standards used by the World Bank and Indonesia.

Consequently, the World Population Review (WPR) ranks Indonesia as the 73rd poorest country in the world.

Poverty represents a significant barrier to regional or national development as it often results in diminished human capital through limited access to education and healthcare [3]. That underscores the critical need for effective government policies aimed at poverty alleviation to foster progress within a region or country. However, poverty rates can vary significantly across regions due to various influencing factors. Therefore, the government must gain a comprehensive understanding of poverty across different provinces in Indonesia by categorizing them based on poverty indicators [4].

Cluster analysis, a statistical technique, involves grouping objects with similar characteristics [5]. The primary objective is to divide a dataset into clusters such that elements within each cluster are more alike to each other compared to elements in different clusters, using predefined criteria [6]. This method offers several advantages, including efficiently organizing extensive data sets with multiple variables and its applicability to data measured on interval and ratio scales [7].

Cluster analysis encompasses a variety of algorithms designed for different objectives. One well-known algorithm is K-means, valued for its straightforwardness and efficiency in clustering extensive datasets. However, it can introduce bias when determining the placement of data points relative to their cluster centres. Consequently, several new algorithms have been proposed to address this issue, such as Fuzzy C-Means (FCM) and Fuzzy Probabilistic C-Means (FPCM). FCM, widely applied in pattern recognition and image processing, has demonstrated superior performance compared to K-means. Nevertheless, this algorithm is notably sensitive to noise [8].

Alternatively, the FPCM algorithm extends both the FCM and PCM algorithms, integrating fuzzy and possibilistic principles to overcome their limitations by utilizing two distinct types of memberships: fuzzy degree of membership and possibilistic absolute specificity values. This approach offers the advantage of effectively handling data that may include errors or outliers [9].

Numerous studies have explored the application of FCM and FPCM algorithms, such as the investigation by Rajkumar et al. [10]. They analyzed and validated the partitioned cluster FCM using a subset of CiteScore data. The study employed fuzzy clustering and FCM algorithms on a portion of the CiteScore dataset to investigate data points with similar distances. Their findings suggest that FCM is superior in effectiveness and efficiency compared to the fuzzy clustering method. Wiharto and Suryani [11] examined the efficacy of K-means and FCM algorithms in segmenting retinal blood vessels. Their performance differs significantly, particularly in establishing thresholds based on mean or median values. The research indicates that the FCM algorithm outperforms the K-means algorithm.

Apsari et al. [12] evaluated and categorized students according to their performance using the FCM and FPCM algorithms. These methods effectively identify high-achieving students by minimizing the objective function. Rahakbaw et al. [13] utilized the FCM algorithm to determine scholarship recipients, considering multiple criteria such as GPA, semester, number of dependents, parental income, and transportation methods. Their findings suggest that this algorithm represents a viable option for evaluating scholarship eligibility. Jayasree and Selvakumari [14] introduced an enhanced FPCM clustering algorithm for predicting students' academic outcomes based on their health status. Their research demonstrates that the proposed FPCM algorithm outperforms other clustering techniques like K-Means, K-Medoids, and FCM, achieving an accuracy rate of up to 93%.

Based on the above description, government strategies and policies to alleviate poverty in Indonesia can be effectively implemented through equitable development. Moreover, the government needs to classify Indonesian provinces based on their specific poverty characteristics to understand poverty dynamics in each region comprehensively. We propose employing clustering analysis using the FCM and FPCM algorithms on various datasets of Indonesian poverty. The poverty variables considered include the poverty line,
percentage of the population living in poverty, poverty depth index, poverty severity index, and total number of individuals below the poverty line. This study will compare the two algorithms based on their accuracy and validity. It will utilize the Partition Entropy Index (PEI) to identify the most suitable algorithm for clustering regencies and cities across Indonesia.

2. RESEARCH METHODS

The study utilizes poverty data from Indonesia in 2021, consisting of 514 data points. The variables encompass the poverty line, percentage of the population classified as poor, poverty depth index, poverty severity index, and total number of people living in poverty.

a. The process begins with sourcing data from the Indonesian Central Bureau of Statistics (BPS), encompassing poverty data for each district/city in Indonesia. The variables utilized in this study are as follows: \( x_1 \) representing the poverty line, \( x_2 \) representing the percentage of the poor population, \( x_3 \) representing the poverty depth index, \( x_4 \) representing the poverty severity index and \( x_5 \) representing the number of poor population.

b. Standardizing poverty statistics involves utilizing z-scores. This process is necessary due to variations in the unit sizes of the variables within the dataset.

c. The clustering process using the FCM and FPCM algorithms involves setting a range for the number of clusters. For instance, the number of clusters considered ranges from two to a maximum relevant to Indonesia’s poverty context. Specifically, the number of clusters examined is between 2 and 5, with a weighting exponent of 2. This range of 2 to 5 clusters is chosen because it is considered sufficient to represent the variation in urban poverty data without adding unnecessary complexity to the analysis.

d. Evaluate the clusters formed using PEI.

e. Compare the results of both algorithms to determine which algorithm is most suitable for clustering.

2.1 Data Standardization

Data standardization, called data scaling or normalization, represents an essential preprocessing step in data analysis and machine learning. It involves converting data into a uniform format or scale to facilitate straightforward analysis, comparison, and modeling. The main objective of data standardization is to achieve uniform scaling across all variables or features within a dataset, typically aiming for a mean of 0 and a standard deviation of 1. This approach enables equitable comparison among diverse variables. It prevents any one feature from unduly influencing the analysis based solely on its scale [15].

In common standardization, each variable is typically transformed into standard values known as z-scores [16]. Equation (1) specifies the procedure for standardizing the data.

\[
    z = \frac{x_i - \mu}{\sigma}
\]  

(1)

2.2 Fuzzy C-Means (FCM)

The FCM clustering algorithm holds substantial prominence and is widely utilized across various domains, including pattern recognition, machine learning, and data mining [17]. Initially introduced by Dunn [18], the algorithm was subsequently refined by Bezdek et al. [19].

Shan et al. [20] categorized the FCM algorithm as a form of soft clustering addressed through iterative updates of membership matrices and cluster centroids. In mathematical terms, the centroid of a cluster in the FCM method can be computed using Equation (2):

\[
    V_{kj} = \frac{\sum_{i=1}^{n} (\mu_{ik})^{w} x_{ij}}{\sum_{k=1}^{n} (\mu_{ik})^{w}}
\]  

(2)
In this context, \( v_{kj} \) represents the \( j \)-th component of the centroid for cluster \( k \). The term \( n \) denotes the number of data points. The value \( \mu_{ik} \) is the membership degree of data point \( i \) in cluster \( k \), ranging from 0 to 1. The parameter \( w \) is the fuzziness factor, usually greater than 1. Lastly, \( x_{ij} \) refers to the \( j \)-th component of the \( i \)-th data point.

The value \( \mu_{ik} \) indicates the strength of membership of the \( i \)-th data point in cluster \( k \). It ranges from 0 to 1, with values closer to 1 signifying a solid membership and values closer to 0 indicating a weak membership. Equation (3) defines how to calculate this membership degree.

\[
\mu_{ik} = \left[ \frac{\left(\frac{1}{\sum_{j=1}^{m}(x_{ij} - v_{kj})^2} \right)^{w-1} \sum_{j=1}^{c} \left(\sum_{j=1}^{m}(x_{ij} - v_{kj})^2 \right)^{w-1}} \right]^{-1}
\]

(3)

The summation symbol, \( \sum_{j=1}^{m}(x_{ij} - v_{kj}) \) used to compute the overall difference between each data component and the centroid cluster component across all dimensions. This method is implemented on every individual data point inside the collection. The data components \( x_{ij} \) represent the values assigned to each dimension, while the centroid cluster components \( v_{kj} \) the values that indicate the centre of each cluster across all dimensions. The parameter \( w \) is a critical determinant in the FCM algorithm, controlling the level of fuzziness in clustering. Usually, the value is more than 1, and as the value increases, the boundaries between the resulting clusters get increasingly muddled. The FCM algorithm uses this Equation to iteratively compute the cluster centers and the degree of membership of each data point to each cluster. This information can generate more adaptable data segmentation than classical clustering approaches.

An iteration process to minimize the objective function is required to obtain accurate cluster centers and membership degrees. The formula for the objective function in the FCM technique can be derived using Equation (4).

\[
J_w(U, V) = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^w d_{ik}^2
\]

(4)

Equation (4) computes the objective function in the FCM algorithm by summing the weighted squared distances between data points and cluster centroids, considering membership degrees. Here, \( n \) represents the total data points, \( c \) is the total clusters, \( \mu_{ik} \) indicates data point \( i \) membership degree to cluster \( k \), \( d_{ik} \) signifies the Euclidean distance between data point \( i \) and cluster centroid \( k \), and \( w \) is the fuzziness parameter controlling clustering’s fuzziness degree. The aim is to minimize \( J_w(U, V) \) to attain precise cluster centres and membership degrees through iterative processes. Lesser \( J_w(U, V) \) values imply better cluster quality.

The Euclidean distance equation utilized in this study is defined by Equation (5):

\[
d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - v_{kj})^2}
\]

(5)

where \( m \) denotes the dimensions, \( x_{ij} \) represents data point \( i \)'s component \( j \), and \( v_{kj} \) indicates cluster \( k \)'s centroid component \( j \). It measures dissimilarity between data points and cluster centroids.

### 2.3 Fuzzy Possibilistic C-Means (FPCM)

The FPCM represents a progression from the algorithms utilized in FCM and PCM. It incorporates cluster centres, specificity, and objective function values [21]. The cluster’s centre in this method can be found using Equation (6).

\[
v_{kj} = \frac{\sum_{i=1}^{n}\left(\mu_{ik}^p + \eta_{ik}^q\right)x_{ij}}{\sum_{i=1}^{n}\mu_{ik}^p + \eta_{ik}^q}
\]

(6)
The matrix $T$, referred to as the absolute distinctiveness matrix, is derived from the ultimate combination of the membership degree ($\mu_{ik}$) and cluster centre ($v_{kj}$) values in the FCM algorithm. Equation (7) presents the formal mathematical formulation employed to calculate this absolute distinctiveness matrix within the FPCM methodology.

$$T = \begin{bmatrix} t_{11} & t_{12} & \ldots & t_{1k} \\ t_{21} & t_{22} & \ldots & t_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ t_{i1} & t_{i2} & \ldots & t_{ik} \end{bmatrix}$$  \hspace{1cm} (7)

In the equation provided above, there exist matrix elements denoted as $t_{11}, t_{12}, \ldots, t_{ik}$. Each element in the absolute distinctiveness matrix can be computed using Equation (8):

$$t_{ik} = \left[ \frac{\sum_{j=1}^{m}(x_{ij} - v_{kj})^{2}}{\sum_{j=1}^{m} \left[ \sum_{j=1}^{n}(x_{ij} - v_{kj})^{2} \right]^{\frac{1}{2}}} \right]^{-1}$$  \hspace{1cm} (8)

The objective function of the FPCM algorithm is defined by Equation (9):

$$J_{\omega}(U, T, V) = \sum_{k=1}^{c} \sum_{i=1}^{n} (\mu_{ik}^{w} + t_{ik}^{\eta}) \left[ \sum_{j=1}^{m} (x_{ij} - v_{kj})^{2} \right]^{\frac{1}{2}}$$  \hspace{1cm} (9)

3. RESULTS AND DISCUSSION

3.1 Data Input

This study aims to cluster each province in Indonesia based on 2021 poverty data at the district level in 2021, using a variety of relevant variables. Table 1 provides a detailed summary of the poverty statistics at the district level in Indonesia for 2021.

<table>
<thead>
<tr>
<th>District Name</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simeulue</td>
<td>458896</td>
<td>18.98</td>
<td>2.37</td>
<td>0.50</td>
<td>18.25</td>
</tr>
<tr>
<td>Aceh Singkil</td>
<td>487249</td>
<td>20.36</td>
<td>3.67</td>
<td>0.91</td>
<td>25.48</td>
</tr>
<tr>
<td>Aceh Selatan</td>
<td>418689</td>
<td>13.18</td>
<td>1.69</td>
<td>0.40</td>
<td>32.25</td>
</tr>
<tr>
<td>Aceh Tenggara</td>
<td>404725</td>
<td>13.41</td>
<td>2.51</td>
<td>0.71</td>
<td>29.31</td>
</tr>
<tr>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
</tr>
<tr>
<td>Kota Jayapura</td>
<td>1051297</td>
<td>11.39</td>
<td>2.53</td>
<td>0.79</td>
<td>34.79</td>
</tr>
</tbody>
</table>

3.2 Data Standardization

This data standardization phase aims to homogenize the units of measurement across different data sets. As described in Section 2, the research data includes variables with varying units of measurement. Therefore, it is essential to use the z-score method for data normalization. This normalization aims to achieve more accurate and reliable clustering results. Table 2 presents the standardized poverty data at the district level in Indonesia.

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.01528</td>
<td>0.89919</td>
<td>0.15204</td>
<td>-0.08261</td>
<td>-0.58672</td>
</tr>
<tr>
<td>0.24073</td>
<td>1.08421</td>
<td>0.82583</td>
<td>0.42850</td>
<td>-0.46667</td>
</tr>
</tbody>
</table>
3.3 Clustering Process

The next step involves performing clustering using the FCM and FPCM methods. In the FCM method, a weighting exponent (w) of 2 is applied, and the number of clusters varies from 2 to 5. The process includes 1000 iterations using the Euclidean distance metric. Table 3 presents the list of Indonesian provinces assigned to each cluster.

<table>
<thead>
<tr>
<th>X₁</th>
<th>X₂</th>
<th>X₃</th>
<th>X₄</th>
<th>X₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.37834</td>
<td>0.12158</td>
<td>-0.20040</td>
<td>-0.20727</td>
<td>-0.35426</td>
</tr>
<tr>
<td>.....</td>
<td>.....</td>
<td>.....</td>
<td>.....</td>
<td>.....</td>
</tr>
<tr>
<td>0.87064</td>
<td>2.21712</td>
<td>0.83102</td>
<td>0.26644</td>
<td>-0.41852</td>
</tr>
<tr>
<td>2.38850</td>
<td>3.93994</td>
<td>1.99202</td>
<td>0.60302</td>
<td>-0.53591</td>
</tr>
<tr>
<td>1.50934</td>
<td>3.79648</td>
<td>0.88803</td>
<td>0.12931</td>
<td>-0.37784</td>
</tr>
<tr>
<td>5.33391</td>
<td>-0.11841</td>
<td>0.23497</td>
<td>0.27890</td>
<td>-0.31209</td>
</tr>
</tbody>
</table>

Table 3. List of provinces in Indonesia

<table>
<thead>
<tr>
<th>No</th>
<th>Provinces</th>
<th>No</th>
<th>Provinces</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aceh</td>
<td>18</td>
<td>Nusa Tenggara Barat</td>
</tr>
<tr>
<td>2</td>
<td>Sumatera Utara</td>
<td>19</td>
<td>Nusa Tenggara Timur</td>
</tr>
<tr>
<td>3</td>
<td>Sumatera Barat</td>
<td>20</td>
<td>Kalimantan Barat</td>
</tr>
<tr>
<td>4</td>
<td>Riau</td>
<td>21</td>
<td>Kalimantan Tengah</td>
</tr>
<tr>
<td>5</td>
<td>Jambi</td>
<td>22</td>
<td>Kalimantan Selatan</td>
</tr>
<tr>
<td>6</td>
<td>Sumatera Selatan</td>
<td>23</td>
<td>Kalimantan Timur</td>
</tr>
<tr>
<td>7</td>
<td>Bengkulu</td>
<td>24</td>
<td>Kalimantan Utara</td>
</tr>
<tr>
<td>8</td>
<td>Bengkulu</td>
<td>24</td>
<td>Kalimantan Utara</td>
</tr>
<tr>
<td>7</td>
<td>Bengkulu</td>
<td>24</td>
<td>Kalimantan Utara</td>
</tr>
<tr>
<td>8</td>
<td>Lampung</td>
<td>25</td>
<td>Sulawesi Utara</td>
</tr>
<tr>
<td>9</td>
<td>Kep. Bangka Belitung</td>
<td>26</td>
<td>Sulawesi Tengah</td>
</tr>
<tr>
<td>10</td>
<td>Kep. Riau</td>
<td>27</td>
<td>Sulawesi Selatan</td>
</tr>
<tr>
<td>11</td>
<td>DKI Jakarta</td>
<td>28</td>
<td>Sulawesi Tenggara</td>
</tr>
<tr>
<td>12</td>
<td>Jawa Barat</td>
<td>29</td>
<td>Gorontalo</td>
</tr>
<tr>
<td>13</td>
<td>Jawa Tengah</td>
<td>30</td>
<td>Sulawesi Barat</td>
</tr>
<tr>
<td>14</td>
<td>DI Yogyakarta</td>
<td>31</td>
<td>Maluku</td>
</tr>
<tr>
<td>15</td>
<td>Jawa Timur</td>
<td>32</td>
<td>Maluku Utara</td>
</tr>
<tr>
<td>16</td>
<td>Banten</td>
<td>33</td>
<td>Papua Barat</td>
</tr>
<tr>
<td>17</td>
<td>Bali</td>
<td>34</td>
<td>Papua</td>
</tr>
</tbody>
</table>

Figure 1 to Figure 4 offers a detailed depiction of poverty clustering in Indonesia using the FCM method, presented through graphical plots. These visuals utilize principal component analysis to streamline dimensions or variables into a more succinct format.
Figure 1. FCM method plot result with 2 cluster

Figure 2. FCM method plot result with 3 cluster

Figure 3. FCM method plot result with 4 cluster
Above of Figure 1 to Figure 4 show that the results of clustering poverty data at the provincial level in Indonesia using the FCM method. Each figure displays points representing provinces grouped within each cluster, with their respective order detailed in Table 1. In Figure 1, Cluster 1 exhibits points closer together than Cluster 2, suggesting more significant similarity in characteristics within Cluster 1 than Cluster 2. Similarly, in Figure 2, Cluster 2 shows closer points than Clusters 1 and 3, indicating more similarity within Cluster 2. Figure 3 demonstrates that Clusters 1 and 3 points are closer together than Clusters 2 and 4, highlighting similarities between Clusters 1 and 3. In Figure 4, Clusters 2, 3, and 4 points are closer together than in Clusters 1 and 5, emphasizing similarities among Clusters 2, 3, and 4. Subsequently, further elaboration on poverty clustering in Indonesia using the FPCM method will be presented through visualization in Figures 5-8.
Figure 6. FPCM method plot results for 3 cluster

Figure 7. FPCM method plot results for 4 cluster

Figure 8. FPCM method plot results for 5 cluster
Cluster 1 in Figure 5 loser dots than Cluster 2, implying that Cluster 1 shares more similar characteristics among its members than Cluster 2. Figure 6 illustrates that the points within Cluster 1 are closely grouped in contrast to Clusters 2 and 3, suggesting greater homogeneity in the characteristics of Cluster 1 compared to Clusters 2 and 3. Figure 7 demonstrates that the points in Clusters 2 and 3 are closer together than in Clusters 1 and 4, indicating higher similarity in the characteristics of Clusters 2 and 3 relative to Clusters 1 and 4. Figure 8 reveals that Clusters 1, 3, and 4 points are closer together than in Clusters 2 and 5, implying that the properties of Clusters 1, 3, and 4 are more comparable than those of Clusters 2 and 5.

After generating clustering results using FCM and FPCM methods, cluster evaluation was performed to determine the optimal algorithm for poverty clustering based on the PE value. Table 4 presents the results of the poverty data cluster evaluation using PE.

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Partition Entropy (PE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FCM</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.4000031</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>0.6827615</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>0.7853068</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>0.8064033</td>
</tr>
</tbody>
</table>

Table 4 shows that the PE validity index value in the FCM method consistently exceeds that in the FPCM method across clusters 2 to 5. The FPCM approach outperformed the FCM method in classifying poverty in Indonesia.

After selecting the best method, the next step is to determine the optimal number of clusters using the FPCM method, which is the best method identified. This determination is based on the PE validity index value, where the lowest PE value indicates the most optimal cluster. The criterion for the PE validity index in determining the optimal number of clusters is that a smaller PE value, closer to zero, indicates better results.

In Indonesia’s poverty context, optimizing clusters is crucial for identifying groups of people with similar poverty characteristics. Based on Table 4, it can be concluded that the optimal number of clusters is 2. That is because, among all the clusters tested (ranging from 2 to 5), the two-cluster solution using the FPCM method has the lowest PE value, 0.1560257.

Therefore, using this method, two clusters can provide the most precise and most informative segmentation of the poor population in Indonesia. The first cluster might include individuals or families experiencing the most severe poverty. In contrast, the second cluster could consist of those on the brink of poverty who still require assistance and intervention. By identifying and understanding the characteristic differences between these two clusters, more effective and targeted programs and policies can be designed to reduce poverty in Indonesia.

4. CONCLUSIONS

After closely analyzing and comparing the results of the FCM and FPCM methods discussed earlier, it becomes evident that the FPCM approach proves significantly more effective for clustering provincial-level poverty data in Indonesia. Evaluating this method’s effectiveness entails examining the PE validity index. Specifically, in the FPCM method, the PE validity index is lower than that of the FCM method, indicating that cluster 2 exhibits the lowest validity index values among various clusters, specifically 0.46827 and 0.02074, respectively. This finding highlights that the FPCM method has achieved more accurate and reliable clustering results.

REFERENCES


IMPLEMENTATION OF FUZZY C-MEANS AND FUZZY POSSIBILISTIC C-MEANS ALGORITHM ON …