Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2025, 2025:123

https://doi.org/10.28919/cmbn/9547

ISSN: 2052-2541

MAPPING DISTRICTS IN WEST JAVA BY UNDER-FIVE PNEUMONIA INDICATORS: AN AGGLOMERATIVE HIERARCHICAL CLUSTERING STUDY (OPEN DATA JABAR 2023)

DIANDA DESTIN[1], AISYA PUTRI SYANURLI[1], DINA SASKYA HUTAJULU[1], SRI WINARNI[2,*], RESTU ARISANTI[2], ANINDYA APRILIYANTI PRAVITASARI[2], TRIYANI HENDRAWATI[2], IRLANDIA GINANJAR[2]

[1]Bachelor Programme of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia

[2]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang 45363, Indonesia

Abstract: In West Java Province, Indonesia, pneumonia remains a major contributor to morbidity and mortality among children under five years of age. This study aims to employ an agglomerative hierarchical clustering approach to classify districts and cities based on indicators associated with under-five pneumonia. The analysis utilized secondary data from *Open Data Jabar* (2023), encompassing nine cities and eighteen districts. The variables considered included malnutrition prevalence, vitamin A supplementation coverage, incidence of low birth weight, DPT-HB-HIB 3 immunization coverage, the proportion of households practicing Clean and Healthy Living Behavior (PHBS), and the number of reported pneumonia cases. Euclidean distance was applied to compare five linkage methods, with the Single Linkage method selected for final clustering due to its highest cophenetic correlation coefficient (0.8681). The optimal clustering solution yielded four distinct profiles, ranging from regions with high PHBS coverage and immunization rates to those with low immunization coverage and high malnutrition prevalence. These findings provide an evidence-based framework for designing region-specific pneumonia prevention and control strategies targeting under-five populations in West Java.

[*]Corresponding author

E-mail address: sri.winarni@unpad.ac.id

Received August 09, 2025

Keywords: pneumonia; under-five; West Java; public health mapping; agglomerative hierarchical clustering; single linkage.

2020 AMS Subject Classification: 62P10.

1. INTRODUCTION

Pneumonia is a significant global health issue resulting in elevated morbidity and mortality rates across all age demographics, especially among at-risk populations including children under five, the elderly, and immunocompromised individuals. It is a severe respiratory infection of the lungs induced by bacteria, viruses, or fungi [1], with clinical severity varying from moderate to life-threatening based on patient age, comorbidities, and pathogen type. Pneumonia constitutes over 12.8% of all fatalities in children under five worldwide [2], resulting in over 700,000 deaths per year, or approximately 2,000 daily [3].

Pneumonia presents a significant risk to child health in low- and middle-income nations, where access to healthcare services and preventive measures like immunization is restricted. In Indonesia, pneumonia was the primary cause of death for children under five in 2022, accounting for 12.5% of overall mortality in this demographic [4].

Figure 1 illustrates the countrywide trend in reported pneumonia cases from 2012 to 2022. The figures reveal a significant rise from 29.5 thousand instances in 2014 to over 63 thousand cases in 2015, and 65.3 thousand cases in 2016. Following this peak, variations occurred, with instances decreasing to 31.4 thousand in 2021 before increasing again to 38.78 thousand in 2022. These fluctuations may be affected by factors including alterations in surveillance coverage, public health efforts, immunization initiatives, and the indirect consequences of the COVID-19 pandemic.

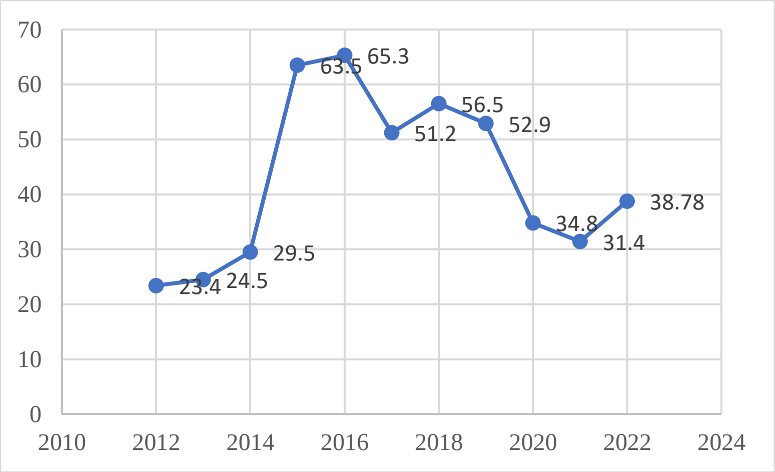

Figure 1. Pneumonia cases in Indonesia 2012-2022

West Java Province has a notably elevated burden, documenting the highest incidence of pneumonia fatalities in the country [4]. In 2023, the province reported 97,171 pneumonia cases in children under five, reflecting a 2.63% increase over the prior year. This load is affected by various factors, including viral and bacterial infections [5], dietary status [6], lifestyle, and healthcare availability [7]. Prior research has shown vaccination coverage (DPT-HB-HIB 3), vitamin A supplementation, family adoption of Clean and Healthy Living Behaviors (PHBS), and the prevalence of low birth weight (LBW) as significant drivers of pneumonia risk in early children [8], [9], [10]. These variables can impair immune function and elevate vulnerability to infection.

Immunization is essential for the prevention of pneumonia. Toddlers with insufficient DPT-HB-HIB immunization were nearly five times more likely to acquire pneumonia than those who were properly immunized [11], [12], [13]. Proper vitamin A consumption bolsters immune defense, whereas malnutrition, especially in moderate to severe cases, heightens pneumonia-related mortality [14]. Promoting public health behaviors, such as ensuring adequate air circulation and prohibiting indoor smoking, can diminish respiratory irritants and curtail the transmission of pathogens [15] [16] Low birth weight, particularly in very low birth weight neonates, has been associated with an increased susceptibility to pneumonia [17].

Pneumonia prevention and control are intimately associated with the United Nations Sustainable Development Goals (SDGs), specifically SDG 3, from a broader development viewpoint. Guarantee healthy lifestyles and foster well-being for all individuals across all age groups. Target 3.2 specifically seeks to eliminate preventable fatalities among newborns and children under five by 2030, with all nations striving to decrease under-five mortality to a maximum of 25 per 1,000 live births. Confronting pneumonia, the predominant infectious cause of mortality in this demographic, is crucial for attaining this objective. Furthermore, initiatives aimed at enhancing immunization rates, nutrition, and healthy living conditions support further SDG 3 objectives concerning universal health coverage and the mitigation of communicable illnesses.

Notwithstanding its significance, research on pneumonia in children under five in West Java is scarce, and to our knowledge, no studies have utilized the latest provincial data to delineate regions based on several risk markers. Prior research frequently depended on insufficient datasets or utilized less rigorous analytical methods, constraining the precision and relevance of their conclusions.

This work fills the gap by utilizing an agglomerative hierarchical clustering method on 2023 data from West Java Province to map districts and cities according to pneumonia-related indicators. This geographical clustering establishes a framework for focused preventative and intervention measures that correspond with SDG 3 and its related targets.

This study offers four key contributions: (a) increasing public awareness of pneumonia risk distribution in West Java, (b) aiding the development of region-specific prevention and treatment programs, (c) enhancing the quality and efficiency of local health services, and (d) enriching the literature on spatial health risk profiling in under-five populations to support global child health and development objectives.

2. Materials and Methods

2.1. Multicollinearity

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can lead to unstable and unreliable coefficient estimates [18]. In the context of mapping factors influencing pneumonia in West Java using hierarchical cluster analysis, addressing multicollinearity is crucial to ensure accurate factor identification. High multicollinearity inflates the variance of regression coefficients, making it difficult to determine the individual effect of each predictor on the outcome. This can result in misleading significance tests, where important variables appear insignificant, and vice versa. To detect multicollinearity, the Variance Inflation Factor (VIF) is commonly used, with a VIF value exceeding 10 indicating potential multicollinearity [19].

$$VIF_i = \frac{1}{1 - R_i^2} \tag{1}$$

where R_i^2 is the R^2 value obtained by regressing the i-th predictor on all other predictors. Reducing multicollinearity may involve removing highly correlated variables or applying regularization techniques, such as Ridge Regression, to stabilize the model. Addressing this issue is essential in hierarchical clustering, as it helps reduce Type 1 errors when analyzing correlated regressors through common factors [20].

2.2. Hierarchical Clustering Analysis

Before starting the hierarchical cluster analysis, it is necessary to measure the distance using the Euclidean method with the following formula.

$$\boldsymbol{d_{ij} = \sqrt{\sum_{k=1}^{p} (y_{ik} - y_{jk})^2}} \tag{2}$$

y_{ik} = Value of variable y_k for object i

y_{jk} = Value of variable y_k for object j

The main reason for using euclidean distance is that the algorithm follows the triangle inequality, an important property in metric spaces, which ensures that the distance between two points is always less than or equal to the sum of their distances from a third point. This characteristic is important for maintaining the integrity of distance-based clustering methods [21]. In addition, the Euclidean distance is a member of the Minkowski distance family (where $p = 2$), making it a natural choice for measuring dissimilarity in continuous data, especially when the variables are standardized to ensure comparability across dimensions.

Hierarchical cluster analysis is a method used to group variables into classes that share similar characteristics [22]. This grouping is done by dividing clusters based on proximity, ultimately forming a hierarchical tree-like structure, commonly known as a dendrogram. Such an analysis is useful for identifying patterns, relationships, and similarities within the data, enabling the classification of entities into distinct groups. Moreover, hierarchical clustering calculations typically involve distance measures like Euclidean distance to quantify how similar or dissimilar entities are [23].

At the heart of hierarchical clustering is the agglomerative method, which begins with the assumption that each variable is initially in its own cluster. Then, based on a distance algorithm, the clusters are iteratively merged, forming new, larger clusters until the optimal grouping is achieved. There are five popular agglomerative algorithms that can be applied, namely single linkage, complete linkage, average linkage, the centroid method, and Ward's method. Each of these methods varies in how distances between clusters are calculated and, therefore, in the kinds of clusters they form.

Single Linkage Clustering, or SLINK, is a hierarchical clustering algorithm that focuses on finding clusters based on the minimum distance between sample points in different clusters. It is sensitive to outliers and noise, as it prioritizes the shortest distance between any two points in separate clusters. The formula for calculating the single linkage value is:

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

d_{UW} = closest distance from cluster U to W

d_{VW} = closest distance from cluster V to W

Complete linkage is another hierarchical clustering algorithm. Unlike single linkage, it focuses on forming well-separated clusters by considering the maximum distance between data points in different clusters, which allows it to be more robust to outliers. The ability to form narrow clusters makes it suitable for scenarios where distinct separation between clusters is essential. The formula for complete linkage is:

$$d_{(UV)W} = max\{d_{UW}, d_{VW}\} \tag{4}$$

d_{UW} = closest distance from cluster U to W
d_{VW} = closest distance from cluster V to W

Similarly, the average linkage method is employed in hierarchical clustering analysis to form clusters based on minimizing the average distance between pairs of objects in different clusters. This method strikes a balance between single and complete linkage approaches, aiming to create well-organized clusters by minimizing the average distance. The formula for calculating the average linkage is:

$$d_{(UV)W} = \frac{\sum_{i=1}^{N_{UV}} \sum_{j=1}^{N_{VW}} d_{ij}}{N_{UV} N_{VW}} \tag{5}$$

d_{ij} = distance between object i in cluster UW and object j in cluster VW
N_{UW} = number of objects in cluster UW
N_{VW} = number of objects in cluster VW

Additionally, the centroid method calculates distances by averaging the values of the objects in clusters. This approach focuses on the central point or centroid of clusters to determine how they should be grouped. The formula for the centroid method is:

$$d_{(UV)W} = \frac{N_U}{N_U + N_V} d_{UW} + \frac{N_U}{N_U + N_V} d_{VW} + \frac{N_U}{N_U + N_V} d_{UV} \tag{6}$$

$d_{(UV)W}$ = distance between newly merged clusters (UV) and (W)

d_{UW} = distance between objects in cluster U and objects in cluster W
d_{VW} = distance between objects in cluster V and objects in cluster W
d_{UV} = distance between objects in cluster U and objects in cluster V
N_U = number of objects in cluster U
N_V = number of objects in cluster V

Finally, the Ward method is an agglomerative algorithm that uses the sum of squares within clusters across all variables. It tends to favor merging clusters with fewer objects, resulting in clusters of approximately equal size. The formula for Ward's method is:

$$JKG = \sum_{J=1}^{k}\left(\sum_{i=1}^{nJ} X_{iJ}^{2} - \frac{1}{nJ}\left(\sum_{i=1}^{nJ} X_{iJ}\right)\right) \tag{7}$$

X_{iJ} = value of the i-th object in the J-th cluster

k = number of clusters at each stage

nJ = number of objects in the J-th cluster

Based on the five hierarchical cluster methods, the best method can be selected based on the Cophenetic Correlation Coefficient (CPCC) value. The closer to 1 the cophenetic coefficient value of a method, the better the method represents the true distance. Cophenetic coefficient can be calculated with the following formula.

$$c = \frac{\sum_{i<j}(x_{ij} - \underline{x})(t_{ij} - \underline{t})}{\sqrt{[\sum_{i<j}(x_{ij} - \underline{x})^2][\sum_{i<j}(t_{ij} - \underline{t})^2]}} \tag{8}$$

x_{ij} = Ordinary Euclidean distance between i-th and j-th observations

t_{ij} = The dendrogrammatic distance between model points t_i and t_j. This distance is the height of the node where these two points first joined together.

2.2. Cluster Validation

Evaluating clustering results is essential to determine how well the generated clusters represent the underlying data structure. Several methods can be employed for this purpose, including internal validation and cluster stability analysis.

One commonly used internal validation method is the Silhouette Score. This score measures how similar an object is to its own cluster compared to other clusters [24]. A higher Silhouette value indicates better clustering, signifying that the object fits well within its own cluster and less so in other clusters. The Silhouette Score can be calculated using the following formula:

$$sil(c) = sil(k)\frac{1}{|k|}\sum_{i=1}^{k} sil(c_i) \tag{9}$$

$sil(c)$ = Silhouette value of all clusters

$|k|$ = Number of clusters k

c_i = Average silhouette value

In addition to internal validation, cluster stability analysis is crucial for assessing how stable and predictive the clustering results are. The goal of stability analysis is to ensure that the clusters remain consistent when small changes are made to the data, such as removing a column. Cluster stability analysis can be determined based on:

a. The average proportion of non-overlap (APN): the average proportion of observations that are not placed in the same cluster by clustering based on complete data and clustering based on data with one column removed.

$$APN = \frac{1}{n}\sum_{i=1}^{n} \frac{|C_i - C_i'|}{n} \tag{10}$$

where C_i is the cluster assignment of observation *i* in the full data, C_i' is the cluster assignment when a column is removed, and *n* is the total number of observations. This represents the average proportion of non-overlapping cluster memberships.

b. The average distance (AD): the average distance between observations placed in the same cluster in both cases (complete data and one column removal).

$$AD = \frac{1}{n}\sum_{i=1}^{n} d(x_i, x_i') \tag{11}$$

where $d(x_i, x_i')$ is the distance (usually the Euclidean distance) between the i-th observation in both data conditions. The smaller the AD value, the more consistent the cluster, as observations remain close to each other in the same cluster despite changes in the data.

c. The average distance between means (ADM): the average distance between cluster centres for observations placed in the same cluster in both cases.

$$\boldsymbol{ADM = \frac{1}{k}\sum_{i=1}^{k} d(\mu_i, \mu_i')} \tag{12}$$

where μ_i and μ_i' are the centers of cluster *i* in the full and transformed data, respectively, and *k* is the number of clusters. A lower ADM value indicates that the cluster centers do not shift significantly despite changes in the data, thus signifying better cluster stability.

d. The figure of merit (FOM): the average intra-cluster variance of the deleted columns, where clustering is based on the remaining (non-deleted) columns.

$$FOM = \frac{1}{k}\sum_{i=1}^{k} \frac{1}{|c_i|} \sum_{x_i{}^{(del)} \in c_i} (x_i{}^{(del)} - \bar{x}_i{}^{(rem)})^2 \tag{13}$$

where $x_i{}^{(del)}$ is the value of the observation in the deleted column, while $\bar{x}_i{}^{(rem)}$ is the average of the observations for the remaining columns in the i-th cluster. FOM assesses how well clusters remain compact even if certain attributes are removed from the data. Lower FOM values indicate clusters that are more stable and less susceptible to feature changes.

e. Connectivity: connectivity is a metric that measures how well adjacent observations in the feature space are placed in the same cluster.

$$Connectivity = \sum_{i=1}^{n} \sum_{j \in N_i} \frac{1}{j}(1 - \delta(c_i, c_j)) \tag{14}$$

Where N_i is the nearest neighbor of the i-th observation, and $\delta(c_i, c_j)$ is an indicator function that takes value 1 if observations i and j are in the same cluster, and 0 otherwise. This metric is desirable to have a lower value as it indicates that observations that are close to each other tend to be in the same cluster, which reflects better cluster quality.

f. Dunn Index: the dunn index is used to evaluate cluster quality by calculating the ratio between the minimum distance between clusters and the maximum diameter of the cluster.

$$Dunn = \frac{min_{1 \le i < j \le k} d(c_i, c_j)}{max_{1 \le l \le k} diameter(c_l)} \tag{15}$$

Where $d(c_i, c_j)$ is the closest distance between two clusters i and j, and the cluster diameter is the maximum distance between two observations in the cluster. A higher Dunn Index value indicates that the clusters are more clearly separated from each other and more internally compact, thus signaling good clustering quality.

g. Silhouette Score: besides being used for internal validation, silhouette score can also be used to assess data stability. Silhouette Score will assess how well an observation is placed in a particular cluster by comparing the average distance of the observation to the cluster it occupies and the average distance to other nearby clusters.

$$Silhouette = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{16}$$

Where $a(i)$ is the average distance between the i-th observation and all other points in the same cluster, while $b(i)$ is the average distance between that observation and all points in other nearby clusters. Silhouette values range from -1 to 1, where values close to 1 indicate a well-clustered observation, values close to 0 indicate that the observation is on the border between two clusters, and negative values indicate that the observation may have been placed in the wrong cluster.

The values of APN, ADM, FOM, Connectivity, Dunn Index, and Silhouette Score of the above stability measures have a range of 0 – 1. The smaller the value, the more consistent the clustering result. Meanwhile, the value of AD is infinitely positive. The smaller the value, the better the clustering result. Once clustering consistency is evaluated, the next crucial step is cluster profiling, which allows for a detailed understanding of the characteristics of each cluster. By identifying distinct patterns within the clusters, we can observe how objects vary significantly across different dimensions.

Profiling per cluster is a crucial step in clustering analysis, enabling a detailed description of the characteristics of each cluster. The aim of cluster profiling is to identify distinct patterns in the clusters formed, showing how objects differ significantly across various dimensions. This process involves interpreting the clusters based on their centroid or average value for each member within the cluster, providing a clear profile of each group.

A dendrogram is a visual representation used to illustrate the hierarchical relationships between variables in clustering analysis. It generates a tree diagram that clearly displays the levels between objects, helping to visually understand the structure of the data. In a dendrogram, the closer one object is to another, the more similar they are. The closer the connection between object entities, the greater their similarity. This visualization technique is especially useful for determining how clusters evolve and merge during the hierarchical clustering process.

The divisive coefficient is another important measure in clustering. It quantifies the degree of separation between clusters, aiming to assess how distinct the clusters are after hierarchical clustering has been performed. A divisive coefficient value close to 1 indicates strong separation between clusters, signifying well-defined clusters with clear distinctions. Conversely, a lower divisive coefficient suggests that the clusters are less distinct, potentially overlapping, and not well separated. This metric is essential for evaluating the quality of the clustering solution, helping to ensure that the clusters are meaningful and distinct from one another.

3. Main Results

3.1. Statistics Descriptive

The data used in this study used secondary data derived from the official website of Open Data Jabar. The data relates to pneumonia with observation units of 18 districts and 9 cities in West Java Province in 2023. Table 1 shows the variables used in this study.

Tabel 1. Variable and Name

Variable	Name
X_1	DPT-HB-HIB 3 Immunization
X_2	Households with PHBS (Clean and Healthy Living Behaviour)
X_3	Malnutrition
X_4	Vitamin A
X_5	Low Birth Weight

Data source: Open Data Jabar

Before starting the analysis, the characteristics of the data are known with descriptive statistics as follows.

Tabel 2. Statistics Descriptive

Variabel	X_1	X_2	X_3	X_4	X_5
Min	3120	25860	10447	527	118
1st Qu.	16476	107606	62246	2064	298.5
Median	27165	206354	95685	3095	637
Mean	30661	239630	117053	4214	782.6
3rd Qu	41550	365754	145740	5206	1123.5
Max	101756	583003	338398	16991	1965

From Table 2, the five variables (X_1 to X_5) show various measures of central tendency and spread. For variable X_1, the minimum value is 3,120, with a 1st quartile of 16,476, indicating that 25% of the data falls below this value. The median or middle value is 27.165, while the mean is 30.661, which is slightly higher than the median, indicating the possibility of a slightly right-skewed distribution. The 3rd quartile value of 41,550 indicates that 75% of the data falls below this value, while the maximum value of X_1 reaches 101,756.

For the X_2 variable, the minimum value is 25,860 and the maximum reaches 583,003. The average of this variable is 239,630, which is higher than the median of 206,354. Variable X_3 has a range from 10,447 to 338,398, with a mean of 117,053, also indicating a right-skewed distribution. For variable X_4, the minimum value is 527 and the maximum is 16,991, with a mean of 4,214 which is higher than the median of 3,095. Finally, variable X_5 shows a data distribution that also involves a wide range, ranging from 118 to 1,965, with a mean of 782.6 and a median of 637, indicating a fairly varied distribution of data.

Given these variations in scale across the variables, it becomes necessary to standardize the data before measuring the distances because clustering is very sensitive to differences in scale between variables. Then the method to measure the distance used is Euclidean so that the following heatmap is produced.

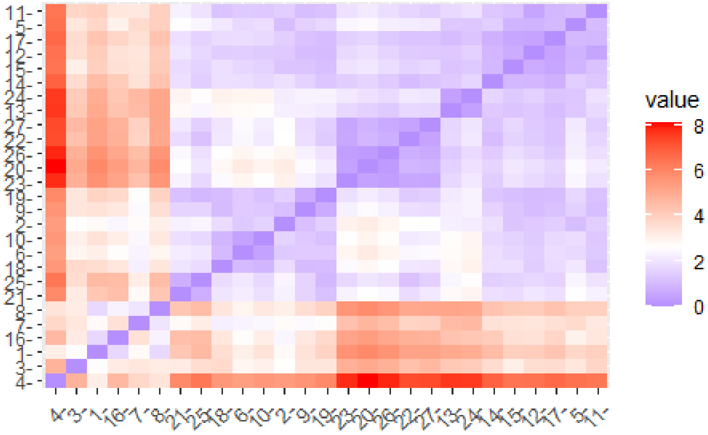

Figure 2. Heatmap for Euclidean Distance

The heatmap shows the relationship or correlation between several variables, with the color scale on the right side indicating values. Red color indicates higher value, the greater the distance between the two objects . While purple color indicates lower value, the closer the distance between the objects. Cimahi City (23) and Sukabumi City (26) in purple color indicate close distance. Meanwhile, Banjar City (20) and Bogor Regency (4) are colored red, indicating a long distance.

After evaluating these spatial relationships, it is also important to check for multicollinearity among the variables, as it can affect the clustering results Multicollinearity test on each variable will be detected using the VIF (Variance Inflation Factor) value. The variable will have a multicollinearity problem if the VIF value > 10.

Tabel 3. VIF Value

Variable	VIF
X_1	5.091128
X_2	1.805066
X_3	5.913942
X_4	5.473429
X_5	3.821124

Based on these calculations, there is no VIF value > 10 so there is no multicollinearity problem. Therefore, the analysis can proceed directly.

3.2. Best Method Selection

The best method based on the five methods can be selected by comparing the largest Cophenetic coefficient values.

Tabel 4. Cophenetic Coefficient Value

Variable	VIF
Single Linkage	0.8680988
Complete Linkage	0.8302522
Average Linkage	0.8414448
Centroid Linkage	0.8480795
Ward Linkage	0.8156354

In Table 4, the Cophenetic Coefficient values are compared across five different clustering methods. Among these, the Single Linkage method has the highest Cophenetic Coefficient value of 0.8680988, suggesting that this method provides the most accurate representation of the dissimilarities between the data points when clustered hierarchically. Therefore, the Single Linkage method is selected as the best clustering technique for this analysis, as it yields the highest correlation between the original data distances and the cluster-derived distances.

By maximizing the Cophenetic Coefficient, the Single Linkage method enhances the reliability of the clusters formed, thus ensuring that any subsequent analysis or policy recommendations based on these clusters are grounded in a robust representation of the data. This selection allows for more accurate interpretations and decisions regarding the grouping of influential factors, ultimately supporting targeted and effective interventions to address pneumonia in West Java.

This result implies that the Single Linkage method better preserves the data's natural groupings compared to other methods, making it the most suitable approach for further analysis in the context of mapping factors influencing pneumonia in West Java.

3.3. Cluster Validation

Silhouette Score

In looking at the quality and strength of clusters, the Silhouette Score value is used to measure how well an object is positioned in a cluster.

Tabel 5. Silhouette Score

Variable	VIF
3	0.3088369
4	0.4420914

The silhouette score provides a measure of how well-defined and distinct clusters are, with values ranging from -1 to 1. Based on Table 5, the clustering result with 3 clusters has a silhouette score of 0.3088, indicating that while the clusters are somewhat defined, there may be some overlap or ambiguity between them. This suggests that the separation between clusters is not particularly strong. However, when the number of clusters is increased to 4, the silhouette score improves to 0.4421, reflecting a better-defined clustering structure with more distinct groupings and less overlap. Although the 4-cluster solution appears to be a better fit than the 3-cluster model, the relatively moderate silhouette scores suggest that there is still some room for improvement, and additional refinement or an alternative clustering approach could enhance the clarity of the clusters.

Cluster Stability Analysis

Several stability cluster sizes will be measured so as to obtain the best number of clusters in each size. The greater number of clusters will be selected.

Tabel 6. Cluster Stability Score

Method	Score	Cluster
APN	0.0498	3
AD	1.7514	4
ADM	0.1898	3
FOM	0.7762	4
Connectivity	6.7548	3
Dunn	0.6492	4
Silhouette	0.4421	4

As shown in Table 6, the scores for different cluster sizes indicate that most measures favor 4 clusters as the optimal solution. Specifically, metrics such as AD, FOM, the Dunn index, and the Silhouette score support the use of 4 clusters, with the Silhouette score of 0.4421 suggesting moderate clustering quality. On the other hand, APN, ADM, and Connectivity slightly favor a solution with 3 clusters, but their differences are less pronounced. Given that the majority of the stability measures, especially the Silhouette score, point to 4 clusters, k = 4 is determined to be the most optimal and stable number of clusters. This ensures a well-balanced and interpretable clustering solution for analyzing factors influencing pneumonia in West Java.

3.4. Profiling

The validation results show that the best number of clusters is four clusters, then profiling is carried out to determine the characteristics of each cluster using the average value of each variable.

Tabel 7. Profiling

Cluster	X_1	X_2	X_3	X_4	X_5
Cluster 1	47708	427293	250533	8568	2024
Cluster 2	21857	182726	76030	2800	533,5
Cluster 3	76249	534689	228346	3706	445
Cluster 4	101756	388891	333332	16991	2200

Based on the profiling results, it can be observed that cluster 4 has the highest average for variables such as DPT-HB-HIB 3 immunization, malnutrition cases, vitamin A distribution, and the number of babies with low birth weight (BBLR). Meanwhile, cluster 3 has the highest average for the number of households practicing Clean and Healthy Living Behavior (PHBS). This distribution of variables across clusters can be better understood through the dendrogram, which visually represents the hierarchical relationships between these clusters and the data points within them. The dendrogram helps illustrate how clusters are formed based on the similarity of these variables, indicating the distinct characteristics of each cluster, such as cluster 4's higher health-related concerns and cluster 3's stronger focus on healthy household practices. So, the dendrogram above shows the result of single linkage clustering, a hierarchical clustering method that merges the two closest clusters based on the minimum distance between points in the cluster.

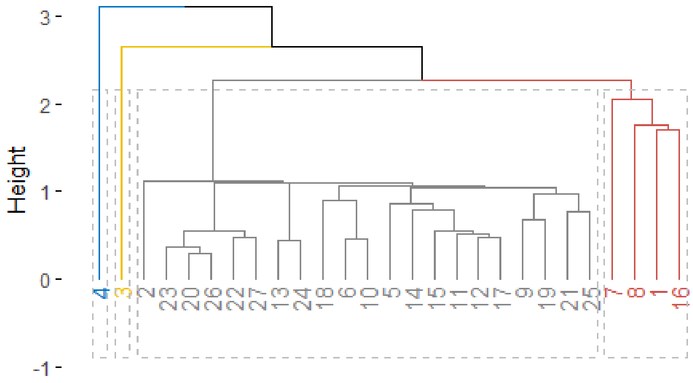

Figure 3. Single Linkage Clustering Dendogram

In this dendrogram, the vertical axis (Height) represents the level of dissimilarity or distance between the clusters when they are merged. The higher the merge, the greater the dissimilarity between the clusters. At the bottom of the dendrogram, there are data points identified by numbers, and as we move upwards, clusters start to form and merge into larger groups. For example, data points 7, 8 and 16 form a fairly tight cluster at a height of about 2.4, indicating that they are more similar to each other compared to other points. Meanwhile, points 4 and 3 are clustered at a lower height, indicating that they are very similar. Lines of different colors (such as blue, yellow, black, and red) mark the stages of clustering that occur at different distances. Overall, this dendrogram shows how the data points are gradually clustered, with the merging height reflecting the degree of similarity between cluster. The following is the distribution of cities/regencies in each cluster:

Tabel 8. Cluster Distribution

Cluster	City/Regency
Cluster 1	Bandung Regency, Cirebon Regency, Garut Regency, Sukabumi Regency
Cluster 2	West Bandung Regency, Ciamis Regency, Cianjur Regency, Indramayu Regency, Karawang Regency, Kuningan Regency, Majalengka Regency, Pangandaran Regency, Purwakarta Regency, Subang Regency, Sumedang Regency, Tasikmalaya Regency, Bandung City, Banjar City, Bekasi City, Bogor City, Cimahi City, Cirebon City, Depok City, Sukabumi City, Tasikmalaya City
Cluster 3	Bekasi Regency
Cluster 4	Bogor Regency

Based on this division, it was found that cluster one contained 4 districts, cluster two contained 21 cities/regencies, and clusters three and four contained 1 regency. In addition to using dendrograms, cluster plots can be used to map the clusters of districts/cities in West Java Province.

3.5. Cluster Plot

In addition to using dendrograms, cluster plots can be used to map regency/city clusters in West Java Province.

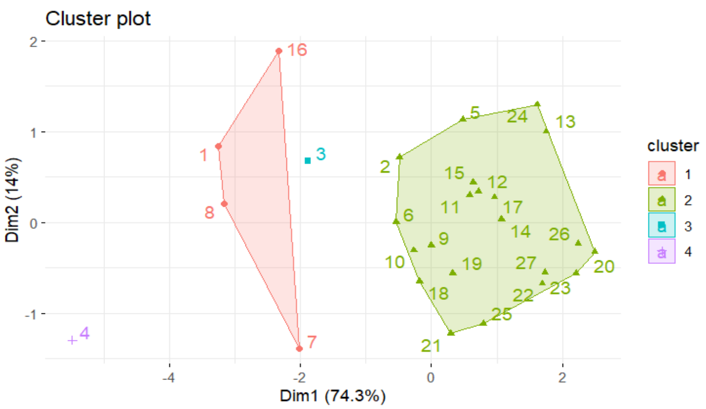

Figure 4. Cluster Plot

The plot shows the visualization of each cluster based on the previously obtained results. The red color indicates the first cluster, the green color the second cluster, the blue color indicates the third cluster, and the purple color indicates the fourth cluster.

3.6. Divisive Coefficient

After the hierarchical process, how well the cluster separates itself from other clusters can be measured by the Divisive Coefficient (DC). The closer to one the coefficient value is, the stronger the cluster differentiation is. Whereas the lower the DC value, the cluster is not well separated and overlap may occur.

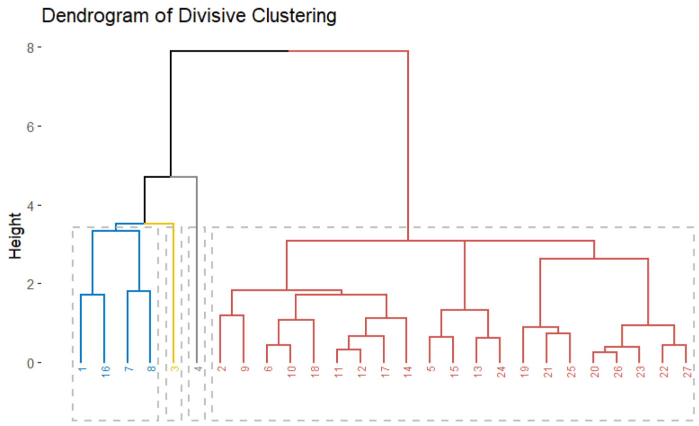

Figure 5. Divisive Clustering Dendrogram

Based on the calculation, the divisive coefficient (DC) value of 0,8655337 was obtained. This value shows that the agglomerative approach provides better results.

4. CONCLUSIONS

The investigation determined that the Single Linkage method is the most appropriate technique for mapping pneumonia-related indicators in West Java Province, as it attained the highest cophenetic correlation coefficient (0.8681) among the five assessed linkage methods. Four clusters were established using this strategy. The evaluation of cluster validity revealed that the four-cluster solution was best, achieving a silhouette score of 0.4421, in contrast to 0.3088 for the three-cluster option, and supported by cluster stability metrics. Cluster profiling revealed distinct characteristics for each group:

a. Cluster 1 (Bandung Regency, Cirebon Regency, Garut Regency, Sukabumi Regency): low immunization coverage, moderate PHBS adoption, high malnutrition prevalence, high vitamin A supplementation, and very low birth weight incidence.

b. Cluster 2 (West Bandung Regency, Ciamis Regency, Cianjur Regency, Indramayu Regency, Karawang Regency, Kuningan Regency, Majalengka Regency, Pangandaran Regency, Purwakarta Regency, Subang Regency, Sumedang Regency, Tasikmalaya Regency, Bandung City, Banjar City, Bekasi City, Bogor City, Cimahi City, Cirebon City, Depok City, Sukabumi City, Tasikmalaya City): very low immunization coverage, very low PHBS adoption, low malnutrition prevalence, low vitamin A supplementation, and adequate average birth weight.
c. Cluster 3 (Bekasi Regency): relatively high immunization coverage, high PHBS adoption, moderately high malnutrition prevalence, low vitamin A supplementation, and very low birth weight incidence.
d. Cluster 4 (Bogor Regency): relatively high immunization coverage, moderate PHBS adoption, high malnutrition prevalence, high vitamin A supplementation, and very low birth weight incidence.

The unique health profiles of each cluster underscore the necessity for region-specific pneumonia prevention and control initiatives in West Java Province. Districts in Cluster 2 necessitate immediate interventions to enhance immunization and promote PHBS adoption, whereas Clusters 1, 3, and 4 demand focused nutrition enhancement programs to combat malnutrition and the incidence of low birth weight. Incorporating these strategies into provincial health planning can enhance resource allocation and amplify effects on child health outcomes.

This study supports SDG 3: Ensure healthy lives and promote well-being for all, particularly Target 3.2 to end preventable deaths of newborns and children under five by 2030. Identifying regional variations in pneumonia-related risk factors provides actionable insights for reducing under-five mortality. Addressing immunization, nutrition, and PHBS adoption also advances other SDG 3 objectives on universal health coverage, communicable disease reduction, and health system strengthening.

ACKNOWLEDGMENT

This research was funded through the Internal Matching Funds (IMF) grant from Universitas Padjadjaran, Indonesia, under the project titled *"Classification Model of Rice Plant Diseases Using Deep Learning and Gaussian Copula to Promote Sustainable Precision Agriculture"* (Contract No. 4356/UN6.D/PT.00/2025).

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] P. Manohar, B. Loh, R. Nachimuthu, X. Hua, S.C. Welburn, S. Leptihn, Secondary Bacterial Infections in Patients with Viral Pneumonia, Front. Med. 7 (2020), 420. https://doi.org/10.3389/fmed.2020.00420.

[2] D. Kulkarni, X. Wang, E. Sharland, D. Stansfield, H. Campbell, H. Nair, The Global Burden of Hospitalisation Due to Pneumonia Caused by Staphylococcus Aureus in the Under-5 Years Children: A Systematic Review and Meta-Analysis, EClinicalMedicine 44 (2022), 101267. https://doi.org/10.1016/j.eclinm.2021.101267.

[3] O.V. Olatunde, O.S. Adewale, P. Thulasiraman, O.A. Daramola, Beyond Binary Diagnostics of Pneumonia Detection with Deep Learning, Int. J. Appl. Inf. Syst. 12 (2024), 22-28.

[4] Kementerian Kesehatan Republik Indonesia, Profil Kesehatan Indonesia 2022, 2023. https://kemkes.go.id/id/profil-kesehatan-indonesia-2022.

[5] A. Torres, C. Cilloniz, M.S. Niederman, et al. Pneumonia, Nat. Rev. Dis. Prim. 7 (2021), 25. https://doi.org/10.1038/s41572-021-00259-0.

[6] Y. Sun, X. Zheng, H. Zhang, X. Zhou, X. Lin, Z. Zheng, J. Zhang, Y. Su, Y. Zhou, Epidemiology of Respiratory Pathogens Among Children Hospitalized for Pneumonia in Xiamen: A Retrospective Study, Infect. Dis. Ther. 10 (2021), 1567-1578. https://doi.org/10.1007/s40121-021-00472-0.

[7] M. Vignari, Non-ventilator Health Care-Associated Pneumonia (NV-HAP): NV-HAP Risk Factors, Am. J. Infect. Control. 48 (2020), A10-A13. https://doi.org/10.1016/j.ajic.2020.03.010.

[8] A.S.M.S.B. Shahid, A.E. Rahman, K.M. Shahunja, et al. Vaccination Following the Expanded Programme on Immunization Schedule Could Help to Reduce Deaths in Children Under Five Hospitalized for Pneumonia and Severe Pneumonia in a Developing Country, Front. Pediatr. 11 (2023), 1054335. https://doi.org/10.3389/fped.2023.1054335.

[9] R. Li, W. Zhao, H. Wang, M. Toshiyoshi, Y. Zhao, H. Bu, Vitamin a in Children's Pneumonia for a COVID-19 Perspective: A Systematic Review and Meta-Analysis of 15 Trials, Medicine 101 (2022), e31289. https://doi.org/10.1097/md.0000000000031289.

[10] T.K. Setyarini, A. Lahdji, I.Z.N. Fajri, Pneumonia Degree Correlation in Children with Clean and Healthy Behavior (CHB), Qanun Med. - Med. J. Fac. Med. Muhammadiyah Surabaya 4 (2020), 217. https://doi.org/10.30651/jqm.v4i2.4269.

[11] S.N. Budihardjo, I.W.B. Suryawan, Faktor-Faktor Resiko Kejadian Pneumonia Pada Pasien Pneumonia Usia 12-59 Bulan di Rsud Wangaya, Intisari Sains Medis 11 (2020), 398-404. https://doi.org/10.15562/ism.v11i1.645.

[12] B.M. Iswari, I. Nurhidayah, S. Hendrawati, Correlation between Immunization Status of DPT-HB-HIB and Pneumonia in Toddler Aged 12-24 Months Old at Babakan Sari Community Health Center Bandung, J. Keperawatan 8 (2017), 101-115.

[13] E. Sidabutar, Ansariadi, Wahiduddin, et al. Analysis of Risk Factor for Pneumonia in Children Less Than Five Years in Makassar, J. Educ. Health Promot. 13 (2024), 16. https://doi.org/10.4103/jehp.jehp_727_23.

[14] A. Kirolos, R.M. Blacow, A. Parajuli, et al. The Impact of Childhood Malnutrition on Mortality from Pneumonia: A Systematic Review and Network Meta-Analysis, BMJ Glob. Health 6 (2021), e007411. https://doi.org/10.1136/bmjgh-2021-007411.

[15] S.M. Simkovich, L.J. Underhill, M.A. Kirby, et al. Design and Conduct of Facility-Based Surveillance for Severe Childhood Pneumonia in the Household Air Pollution Intervention Network (HAPIN) Trial, ERJ Open Res. 6 (2020), 00308-2019. https://doi.org/10.1183/23120541.00308-2019.

[16] C. Lu, W. Yang, Z. Liu, H. Liao, Q. Li, Q. Liu, Effect of Preconceptional, Prenatal and Postnatal Exposure to Home Environmental Factors on Childhood Pneumonia: A Key Role in Early Life Exposure, SSRN (2022). https://doi.org/10.2139/ssrn.4149391.

[17] J. Liu, R. Qiu, Lung Ultrasound Monitoring of Legionella Ventilator-Associated Pneumonia in an Extremely Low-Birth-Weight Infant, Diagnostics 12 (2022), 2253. https://doi.org/10.3390/diagnostics12092253.

[18] T. Kyriazos, M. Poga, Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions, Open J. Stat. 13 (2023), 404-424. https://doi.org/10.4236/ojs.2023.133020.

[19] C.G. Thompson, R.S. Kim, A.M. Aloe, B.J. Becker, Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from Typical Regression Results, Basic Appl. Soc. Psychol. 39 (2017), 81-90. https://doi.org/10.1080/01973533.2016.1277529.

[20] A.G. Assaf, M. Tsionas, A Bayesian Solution to Multicollinearity Through Unobserved Common Factors, Tour. Manag. 84 (2021), 104277. https://doi.org/10.1016/j.tourman.2020.104277.

[21] T.F. Havel, Distance Geometry: Theory, Algorithms, and Chemical Applications, in: Encyclopedia of Computational Chemistry, Wiley, New York, 1995.

[22] C. Essary, L.M. Fischer, E. Irlbeck, A Statistical Approach to Classification: A Guide to Hierarchical Cluster Analysis in Agricultural Communications Research, J. Appl. Commun. 106 (2022), 3. https://doi.org/10.4148/1051-0834.2431.

[23] P. Praveen, M. Ranjith Kumar, M.A. Shaik, R. Ravikumar, R. Kiran, The Comparative Study on Agglomerative Hierarchical Clustering Using Numerical Data, IOP Conf. Ser.: Mater. Sci. Eng. 981 (2020), 022071. https://doi.org/10.1088/1757-899x/981/2/022071.

[24] K.R. Shahapure, C. Nicholas, Cluster Quality Analysis Using Silhouette Score, in: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2020, pp. 747-748. https://doi.org/10.1109/DSAA49011.2020.00096.