CLUSTER ANALYSIS OF WARD AND K-MEANS METHODS TO GROUP REGENCY-CITY IN WEST JAVA PROVINCE BASED ON PARTICIPATION OF WOMEN IN DEVELOPMENT

(Thesis)

By

RAHMA NURKHOLIZ
NPM 2117031076

FACULTY OF MATHEMATICS AND NATURAL SCIENCES
UNIVERSITY OF LAMPUNG
BANDAR LAMPUNG
2025

ABSTRACT

CLUSTER ANALYSIS OF WARD AND K-MEANS METHODS TO GROUP REGENCY-CITY IN WEST JAVA PROVINCE BASED ON PARTICIPATION OF WOMEN IN DEVELOPMENT

By

Rahma Nurkholiz

Cluster analysis is one of the multivariate statistical methods used to group objects based on their similar characteristics without prior labeling. In this study, cluster analysis is used to group regencies-cities in West Java Province based on women's participation in development in 2021-2023 using the Ward and K-Means cluster methods. To overcome the problem of multicollinearity between variables, data transformation is carried out using Principal Component Analysis (PCA) in order that the resulting new variables are independent to each other. The principal component scores were used for clustering the regencies-cities data using the two cluster methods, and the results were then evaluated using the Davies-Bouldin Index (DBI) and Calinski-Harabasz Index (CHI) to determine the optimal number of clusters. The results show that both the Ward and K-Means methods produce the same number of optimal clusters, which are four clusters for data 2021, three clusters for data 2022 and 2023, and four clusters for the average of data 2021-2023.

Keyword: Cluster Analysis, Ward, K-Means, PCA, Davies-Bouldin Index, Calinski-Harabasz Index, Women's Participation.

CLUSTER ANALYSIS OF WARD AND K-MEANS METHODS TO GROUP REGENCY-CITY IN WEST JAVA PROVINCE BASED ON PARTICIPATION OF WOMEN IN DEVELOPMENT

By

RAHMA NURKHOLIZ

Thesis

Submitted as a Partial Fulfilment of the Requirement for the Degree of BACHELOR OF MATHEMATICS

at

Department of Mathematics
Faculty of Mathematics and Natural Sciences

FACULTY OF MATHEMATICS AND NATURAL SCIENCES
UNIVERSITY OF LAMPUNG
BANDAR LAMPUNG
2025

Title of Thesis	: **Cluster Analysis of Ward and K-Means Methods to Group Regency-City in West Java Province Based on Participation of Women in Development**
Name of Student	: Rahma Nurkholiz
ID Number of Student	: **2117031076**
Program of Study	: **Mathematics**
Faculty	: **Mathematics and Natural Science**

APPROVE BY

1. Supervisory Committe

Dr. Khoirin Nisa, S.Si., M.Si.
NIP 197407262000032001

Misgiyati, S.Pd., M.Si.
NIP 198509282023212032

2. Head of The Mathematics Department

Dr. Aang Nuryaman, S.Si., M.Si.
NIP 197403162005011001

RATIFIED BY

1. Examination Committe

Head : **Dr. Khoirin Nisa, S.Si., M.Si.** .....

Secretary : **Misgiyati, S.Pd., M.Si.** .....

Examiner
Non-Supervisor : **Prof. Drs. Mustofa, M.A., Ph.D.** .....

2. Dean of The Faculty of Mathematics and Natural Science

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Date of Passing Thesis Examination: **09 April 2025**

STATEMENT

The undersigned:

Name	:	**Rahma Nurkholiz**
ID Number of Student	:	**2117031076**
Major	:	**Mathematics**
Title of Thesis	:	**Cluster Analysis of Ward and K-Means Methods to Group Regency-City in West Java Province Based on Participation of Women in Development**

Hereby declare that this thesis is my own work. If, at any time in the future, it is proven that this thesis is plagiarized or written by someone else, I am willing to accept any sanctions In accordance with the applicable academic regulations.

Bandar Lampung, 12 April 2025

Author

Rahma Nurkholiz

BIOGRAPHY

The author was born in Bogor on June 21, 2003, as the third child of Mr. Sukarna and Mrs. Syaripah and the younger sister of Risma Andriani and Sahrul Anwar, older sister of Nur Nayla Rabiatul Adawiyah and Raisa Shadiqah Panca Sukarna.

The author has studied at Raudhatul Athfal (RA) Nurul Hidayah in 2008-2009, Madrasah Ibtidaiyah (MI) Nurul Hidayah in 2009-2015, Madrasah Tsanawiyah Negeri (MTsN) 2 Bogor in 2015-2018, and Sekolah Menengah Atas Negeri (SMAN) 1 Leuwiliang in 2018-2021.

In 2021 the author was enrolled as a student in the S1 Mathematics Study Program, Faculty of Mathematics and Natural Sciences, University of Lampung through the SBMPTN route. During being a student, the author was quite active in campus organizations, including being a member of the Social and Community Empowerment team (SPM) of BEM FMIPA for the 2023 period, had been a practicum assistant for Basic Statistics, Statistics and Data Science, and had been a mentor for the Basic Statistics course in the MAFIA HIMATIKA activity in 2023.

As a form of application of the knowledge that had been learned, from December 2023 to February 2024 the author carried out Practical Work (KP) at the Bogor City Statistics Agency assigned to the Statistics Dissemination Division. In August to September 2024 the author carried out the International Real Work Lecture (KKNI) in Huta Tinggi Village, Pangururan District, Samosir Regency, North Sumatra Province as a form of student service and carrying out the Tri Dharma of Higher Education.

INSPIRATIONAL QUOTE

"Allah does not burden a person but according to his ability"

(Qs. Al-Baqarah: 286)

"For indeed after hardship there is ease, indeed after hardship there is ease"

(Qs. Al-Insyirah: 5-6)

"Be good. Verily, Allah loves those who do good"

(Qs. Al-Baqarah: 195)

DEDICATION

By expressing praise and gratitude to Allah SWT who has given guidance and ease to complete my studies, I dedicate this little
work to:

My beloved father and mother who always educate, pray, sacrifice, and other things that I cannot express in words

My dearest siblings

Supervisors and examiners who are very meritorious and not tired of providing direction and input so that the author can complete the thesis

My best friends and friends, thank you for your togetherness, prayers and the encouragement you always give me.

University of Lampung

ACKNOWLEDGEMENT

Alhamdulillahi Robbil 'alamin, Praise and gratitude to Allah SWT for His grace and compassion, so that the author can complete this thesis. Sholawat and salam always remains poured out to the Prophet Muhammad SAW, the main guide and role model for all mankind. The thesis with the title "Cluster Analysis of Ward and K-Means Methods to Group Regency-City in West Java Province Based on Participation of Women in Development" is one of the requirements for obtaining a Bachelor of Mathematics degree at the University of Lampung.

In completing this thesis, many parties have helped the author in providing guidance, encouragement, and suggestions. So, with all sincerity and humility on this occasion the author would like to thank:

1. Dr. Khoirin Nisa, S.Si., M.Si. as the first supervisor who always provides guidance, advice, motivation, advice and input so that the author can complete this lecture and thesis
2. Misgiyati, S.Pd., M.Si. as the second supervisor who has provided input and advice in the completion of the thesis
3. Prof. Drs. Mustofa, MA., Ph.D. as the examiner who has provided criticism and suggestions to the author in the completion of the thesis
4. Siti Laelatul Chasanah, S.Pd., M.Si., as the academic supervisor
5. Dr. Aang Nuryaman, S.Si., M.Si. as Head of the Department of Mathematics, Faculty of Mathematics and Natural Sciences
6. Dr. Eng. Heri Satria, S.Si., M.Si. as Dean of the Faculty of Mathematics and Natural Sciences, University of Lampung

7. Lecturers, staff and employees of the Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung
8. My parents who never tired of giving prayers, support, affection, and sacrifices during the completion of this thesis. This achievement is as much theirs as it is mine
9. My older siblings Risma Andriani and Sahrul Anwar
10. My dearest younger siblings Nayla and Raisa
11. My dearest best friend forever Desi Widianti who has accompanied me in this city, always provided encouragement and support until now
12. My college best friends Adin, Atun, Dita, Deyra, Eva, Fathan, Mei, Maya, Nabila, Sherina, Tasya, and Vedisya who have always provided encouragement, support, advice, and counsel during this study period
13. My best friends from childhood Zalo, Ntang, and Cipa who always provide encouragement and support even though we are separated by distance
14. My best friend from junior high school Azahra, Aray, Haza, Farah, Kiki, Nesa, and Dini who always gave me support even though we are separated by distance
15. Mrs. Nisa's student friends
16. Member of SPM Team BEM FMIPA 2023
17. Friends of KKNI in 2024 Huta Tinggi Village, Pangururan District, Samosir Regency, North Sumatra
18. Friends of Mathematics Class of 2021.

The author also realizes that in this writing there are still many shortcomings. Therefore, the author expects suggestions and constructive criticism for further research to be better.

Bandar Lampung, 12 April 2025
Author,

Rahma Nurkholiz

TABLE OF CONTENTS

Page

TABLE LIST

Table Page

FIGURE LIST

Figure Page

I. INTRODUCTION

1.1 Background and Problems

Cluster analysis is a statistical method that identifies groups of samples based on similar characteristics. The purpose of cluster analysis is to reduce the number of objects by classifying objects into relatively homogeneous clusters. In cluster analysis, clustering is used Euclid distance as a proximity measure (Rencher, 2002). The closer the distance between one individual and another, the greater the level of similarity, so they will be grouped together in one group (Usman & Nurdin, 2013). Cluster analysis is divided into two methods, that is hierarchical and non-hierarchical. Hierarchical cluster analysis is a method used to group observations in a structured manner based on similarity of properties and the number of groups that can be formed is unknown, while non-hierarchical cluster analysis is a cluster technique that requires manual determination of the number of clusters (Rencher, 2002).

In hierarchical cluster analysis, there are two methods, that is agglomerative and divisive. The agglomerative method is further divided into Single Linkage, Complete Linkage, and Average Linkage, Ward Method, Centroid Method, and Median Method (Nugroho, 2008). While in non-hierarchical cluster analysis, one of the most popular methods is K-Means (Baroroh, 2012). In cluster analysis, there is assumptions that must be met, it is the assumption of non-multicollinearity. Multicollinearity is a linear relationship that exists between independent variables that can be seen from the Variance Inflation Factor (VIF) value. If the VIF value is more than 10, it can be concluded that there is multicollinearity (Draper & Smith, 1998) (Widarjono, 2010).

The problem of multicollinearity can be overcome by using Principal Component Analysis (PCA) which aims to reduce the data dimension by transforming the existing independent variables into K principal components. Principal Component Analysis (PCA) is an analytical technique used to transform the initial correlated variables into a new set of variables that are no longer correlated (Johnson & Wichern, 2007).

After clustering, the next step is to evaluate the performance of the clustering results to assess the strength and quality of the grouping of an object against the cluster it produces (Wira *et al.,* 2019). Indices that can be used are the Davies-Bouldin index, Calinski-Harabasz index, and other indices. The Davies-Bouldin index is used to perform the process of measuring a clustering result based on the cohesion and separation values. This index minimizes the average distance between each cluster and the most similar to it (Ansari *et al*., 2011). The Calinski-Harabasz index is the ratio of the sum of dispersion between clusters and dispersion within clusters for all clusters (dispersion is defined as the sum of squared distances) (Caliński & JA, 1974).

Research by Wang & Lu (2021), examines the application of panel data cluster analysis in economic and social research in China using R software where the clustering of 31 provinces is divided into each year period, from 2008-2018. Ramadhan *et al.* (2020), examined cluster analysis of multivariable panel data using Ward's method, focusing on Gross Enrollment Ratio (GER) data in West Java Province in 2015-2018 and concluded that through cluster analysis using Ward's method, regency/city in West Java Province can be grouped into several clusters during 2015 to 2018. Iklima & Pujiyanta (2023), examined the comparison between two clustering methods, that is K-Means and Ward, in grouping customers of a mall and concluded that the use of the Ward method is better than the use of the K-Means method for the process of grouping Mall customers. Fathia *et al.* (2016) examined cluster analysis with Ward and Single Linkage methods to group sub-districts in Semarang Regency based on village potential data.

Research by Ansari *et al.* (2011), conducted an evaluation process with a comparison of several indices, including Davies-Bouldin index to cluster user navigation sessions based on web access logs. Another research by Sikana & Wijayanto (2024) used the Calinski-Harabasz index in clustering the Human Development Index (HDI) in Indonesia in 2019.

According to Escobar (1995), development is a concept generally used to manage people who are often identified as problems, clients, and underdeveloped, more specifically women are one of them. The Indonesian government has published Presidential Instruction (Inpres) No. 9 of 2000 on Gender Mainstreaming in National Development, as a reference to maximize the potential of women in development. According to the West Java Central Bureau of Statistics (BPS), the involvement of women in parliament in West Java Province in 2023 is only 22.69%. In the Trading Development and Gender Equality forum held on the sidelines of the 2019 Asian Development Bank Annual Meeting in Nadi, Fiji, the Minister of National Development Planning/Head of Bappenas stated that women are important assets, have great potential, and are valuable investments for Indonesia. With their capabilities and abilities, women can make a significant contribution to the country's development.

Based on previous research and problems, no one has used Ward and K-Means cluster analysis to conduct research on clustering regency/city in West Java Province based on women's participation in development. Therefore, the author is motivated to conduct research on cluster analysis using the Ward and K-Means methods to group regency/city in West Java Province based on women’s participation in development on the 2021-2023 period, and use the Davies-Bouldin index and Calinski-Harabasz index in the process of evaluating cluster performance.

1.2 Research Objectives

This research aims to apply the Ward and K-Means methods in clustering regency/city in West Java Province using the Davies-Bouldin index and Calinski-Harabasz index as an evaluation of cluster performance based on women's participation in development in 2021-2023.

1.3 Research Benefits

The benefits obtained from this research are:

1. Contribute ideas to expand the knowledge of statistics, especially Ward's agglomerative hierarchical cluster analysis and K-Means non-hierarchical cluster analysis in grouping regency/city based on women's participation in development.
2. Provide input and encouragement for other researchers to further research Ward's agglomerative hierarchical cluster analysis and K-Means non-hierarchical cluster analysis to group the other areas.
3. Provide input and encouragement for the government and related institutions to allocate resources and develop more efficient policies to increase women's participation in development.

II. LITERATURE REVIEW

2.1 Multivariate Analysis

Multivariate analysis is the analysis of multiple variables in one or more relationships. Multivariate analysis is one type of statistical analysis used to analyze data consisting of many variables, both independent and dependent variables (Wijaya & Budiman, 2016). According to Johnson & Wichern (2007), Multivariate analysis includes the analysis of research data using many variables that are subject to simultaneous measurement.

Multivariate analysis is divided into two methods: dependency and interdependency methods. The dependency method has two types of variables: independent and dependent variables. Some of the dependency method analysis techniques are multiple regression, discriminant analysis, canonical correlation, and Manova. Meanwhile, the interdependency method only has one variable, the independent variable. Interdependency analysis techniques consist of factor analysis, cluster analysis, and multidimensional scaling (Wijaya & Budiman, 2016).

Data in multivariate analysis can be expressed in matrix form. The size of a matrix is described by stating the number of rows (horizontal line) and the number of columns (vertical line) contained in the matrix. If $\boldsymbol{X}$ is a matrix, it is used to denote the entries contained in row i and column j of $\boldsymbol{X}$, $i = 1, 2, ..., m$; $j = 1, 2, ..., n$. So, an $m \times n$ matrix can generally be written as follows (Johnson & Wichern, 2007):

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

where $\boldsymbol{X}$ contains data consisting of all observations on all variables.

2.2 Non-Multicollinearity Assumption

Multicollinearity is a linear relationship that exists between independent variables. Multicollinearity can be seen from the Variance Inflation Factor (VIF) value. The formula for calculating VIF is as follows:

$$VIF = \frac{1}{(1-{R_i}^2)} \tag{2.1}$$

where:

${R_i}^2$: coefficient of determination on variable *i*

i : 1,2,3,…, n.

If the VIF value is more than 10, it can be concluded that there is multicollinearity (Draper & Smith, 1998) (Widarjono, 2010).

2.3 Data Standardization

Data standardization or transformation is a process carried out to overcome large differences in unit values between variables in data analysis. This is important because significant differences in units can cause distance calculations to be invalid. Standardization aims to transform data into a common scale, allowing for more accurate and meaningful analysis. Transformation can be done with *z-score*, which

is the transformation of data in the standard normal form N (0,1) formulated as follows (Rencher, 2002):

$$z_i = \frac{x_i - \bar{x}}{s} \quad for\ i = 1,2,3, \dots, n \tag{2.2}$$

where:

x_i = original value of data *i*

$\bar{x}$ = average of all data

s = standard deviation of the data.

2.4 Euclidean Distance

In cluster analysis, clustering is used a measure that can explain the similarity or closeness between data to explain the simple group structure of complex data, that is the distance or similarity measure. The distance measure that is often used is the Euclidean distance measure (Johnson & Wichern, 2007). The smaller the distance of an individual and another individual, the greater the similarity of the individual, so that the individual will be included in the same group (Usman & Nurdin, 2013). Euclidean distance function between two vectors $\boldsymbol{x} = (x_1, x_2, \dots, x_p)'$ and $\boldsymbol{y} = (y_1, y_2, \dots, y_p)'$, are defined as follows (Rencher, 2002):

$$\begin{aligned} d(x,y) &= \sqrt{(x_1 - y_1) + (x_2 - y_2) + \cdots + \left(x_p - y_p\right)^2} \\ &= \sqrt{(\boldsymbol{x} - \boldsymbol{y})'(\boldsymbol{x} - \boldsymbol{y})} = \sqrt{\sum_{j=1}^{p} \left(x_j - y_j\right)^2} \end{aligned} \tag{2.3}$$

where:

$d(x,y)$: Euclidean distance between two vectors x and y

p : number of dimensions or length of the vector

x_j : The jth component or data of the vector x

y_j : The jth component or data of the vector y.

2.5 Cluster Analysis

Cluster analysis is a method in multivariate analysis that aims to find groupings in data, where objects or observations within each cluster are similar to each other, but different from objects in other clusters. This analysis seeks to identify clusters without any prior knowledge of the number or characteristics of the clusters. In contrast to classification analysis that uses predefined groups or populations, cluster analysis groups observations based on measured similarities, for example, using a measure of distance between observations or a comparison of variability between clusters (Rencher, 2002).

According to Nugroho (2008), cluster analysis is a statistical method that identifies groups of samples based on similar characteristics. Cluster analysis groups similar elements as research objects that have a high level of homogeneity between objects into different clusters with a high level of heterogeneity of objects between clusters. The purpose of cluster analysis is to reduce the number of objects by classifying objects (cases or elements) into relatively homogeneous clusters.

2.5.1 Hierarchical Cluster Analysis

The hierarchical cluster method is a method used to group observations in a structured manner based on similarity of properties and the number of groups that can be formed is unknown (Rencher, 2002). The formation of a hierarchical method has the nature of developing a hierarchy or a branching tree-like structure. Hierarchical methods can be through sequential merging (agglomerative) or sequential division (divisive). Agglomerative clustering is a clustering method that is carried out from observing objects that have similarities so that they combine into a small cluster and from the small clusters formed will be combined into one large cluster that contains all clusters.

The Agglomerative Method consists of:

1. Linkage Method

   This method is further divided into three methods that is, single linkage, complete linkage, and average linkage.
2. Ward Method
3. Centroid Method
4. Median Method

Divisive clustering is the opposite of the agglomerative method. This clustering method starts by assuming that there is only one cluster that contains all objects (Nugroho, 2008).

Ward's Method

According to Johnson & Wichern (2007), the Ward method is a hierarchical approach to clustering that aims to minimize the information lost when merging two groups. This method combines the two clusters that have the smallest cost to join according to the equation below (De Amorim, 2015).

$$Ward(S_i, S_j) = \frac{N_{S_i} \cdot N_{S_j}}{N_{S_i} + N_{S_j}} d(c_{S_i}, c_{S_j}) \quad (2.4)$$

Where:

N_{S_i}, N_{S_j} : cardinality of the cluster S_i and S_j or the amount of data in each cluster

c_{S_i}, c_{S_j} : centroid of the cluster S_i and S_j

$d(c_{S_i}, c_{S_j})$: Euclidean distance between centroids c_{S_i} and c_{S_j}.

2.5.2 Non-hierarchical Cluster Analysis

Non-hierarchical cluster analysis is a cluster analysis that requires manual determination of the number of clusters. This method is designed to group objects into K clusters. The number of clusters, K, is predetermined before the clustering procedure begins. One of the popular non-hierarchical cluster methods is K-Means (Baroroh, 2012).

K-Means's Method

The K-Means method is an unsupervised hierarchical method and belongs to the partition-based clustering method. This is because the data analyzed does not have cluster labels, so in the clustering process, there are no definite cluster members. According to MacQueen, the term K-Means describes that this algorithm marks each object grouped into a cluster that has the closest center (average) (Johnson & Wichern, 2007).

The K-Means method aims to minimize variation within the same cluster and maximize variation between clusters. However, there are two main challenges in this non-hierarchical approach. First, the number of clusters must be predetermined. Second, the selection of the initial cluster centers is not always certain. Furthermore, the clustering result is greatly influenced by how the cluster centers are selected. Many algorithms start the process by selecting the first *k* (k = number of clusters) cases as the initial cluster centers. Therefore, the clustering result is highly dependent on the observed data. Despite some drawbacks, this method can be done quickly and is particularly useful when the number of observations is large (Simamora, 2005).

The steps of k-means cluster analysis are as follows:

1. Specify *k* as the number of clusters formed
2. Determining the centroid (cluster center point)

The initial centroid determination is done randomly from the available objects for k clusters. After that, the centroid for the next i-th cluster is calculated using the formula:

$$C_{ij} = \frac{\sum_{q=1}^{m} x_j}{u} \quad (2.5)$$

where:

C_{ij} : centroid value of the i-th cluster on the j-th variable

x_j : data value on the jth variable

u : number of objects in the formed cluster

m : number of variables.

3. Calculate the distance of each object to the centroid of each cluster. To calculate the distance between objects and centroids using Euclidean distance (2.5).
4. Allocate each object to the closest centroid.
5. Iterate and then determine the new centroid position using the equation in (2).
6. Repeat the steps in (3) if the new centroid position is not the same (still changing).

2.6 Evaluation of Cluster Analysis Results

After the clustering results are obtained, the next step is to check how high the quality of the clustering is by looking at the ability of the cluster to distinguish existing data according to the variables or characteristics of the subject used for clustering (Gudono, 2014). In this research, the methods used are the Davies-Bouldin index and the Calinski-Harabasz index.

2.6.1 Davies-Bouldin Index (DBI)

The Davies-Bouldin index is used to measure a clustering result based on the cohesion and separation values discovered by David L. Davies and Donald W. Bouldin. This index minimizes the average distance between each cluster and the most similar to it (Ansari *et al.*, 2011). The following are the steps in calculating the Davies-Bouldin index:

1. Calculate the cohesion/homogeneity matrix using the SSW (Sum of Square Within cluster) formula:

$$SSW_i = \frac{1}{C_i}\sum_{x \epsilon C_i} d(x, c_i) \tag{2.6}$$

   where C_i is the number of points in the cluster C_i, $d(x, c_i)$ is the distance between point x and the centroid c_i, and c_i is the centroid of the cluster C_i.
2. Calculate separation/heterogeneity using the SSB (Sum of Square Between clusters) formula:

$$SSB_{i,j} = d(c_i, c_j) \tag{2.7}$$

   where c_i is the centroid of cluster C_i, c_j is the centroid of cluster C_j, and $d(c_i, c_j)$ is the distance between the centroids of the two clusters.
3. Calculate the ratio to find out how good the comparison value is between one cluster and another.

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \tag{2.8}$$

   Where SSW_i is the i-th cluster, SSW_j is the jth cluster, and $SSB_{i,j}$ is the separation of the i-th and j-th clusters.
4. Calculating the Davies-Bouldin index.

$$DBI = \frac{1}{K}\sum_{i=1}^{K} max_{i \neq j}\left(R_{i,j}\right) \quad (2.9)$$

Where K is the number of clusters, $R_{i,j}$ is the ratio between clusters i and j, and max is the largest ratio between clusters.

The Davies-Bouldin Index gives a lower value for better clustering, with an ideal value close to zero. The larger the Davies-Bouldin Index value, the worse the cluster quality (Hasan, 2024).

2.6.2 Calinski-Harabasz Index (CHI)

Calinski-Harabasz Index, also known as Variance Ratio Criterion, can be used for clustering model evaluation (Caliński & JA, 1974). For a data set X of size n that has been grouped into k clusters, the Calinski-Harabasz index is expressed in the following equation:

$$CHI = \frac{B_k/(k-1)}{W_k/(n-k)} \ \ or\ CHI = \frac{B_k}{W_k} \times \frac{n-k}{k-1} \quad (2.10)$$

with

$$B_k = \sum_{i=1}^{k} n_i \|c_i - C\|^2$$

$$W_k = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2$$

where B_k is inter-cluster variance (between cluster centroid and overall centroid), W_k is intra-cluster variance (between a data point and its cluster centroid), k is number of clusters, and n is the total amount of data.

The higher of the CH index value, the better the clustering result as it shows that the clusters are clearly different from each other with less variation within each cluster (Caliński & JA, 1974).

2.7 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an analytical technique used to transform initial correlated variables into a new set of variables that are no longer correlated. The main objective of PCA is to reduce the dimensionality of the data by transforming the independent variables into *k* principal components. PCA emerged as a solution in data collection situations involving many variables, resulting in fewer new variables but still able to explain data variation effectively (Johnson & Wichern, 2007).

Although it takes a number of *p* components to reproduce the total variability, it is often the case that most of this variability can be explained by a small number of *k* principal components. If so, these k components have almost the same information as the original *p* variables. Therefore, *k* principal components can replace the original *p* variables, and the original data consisting of *n* measurements on *p* variables can be reduced to data with *n* measurements on *k* principal components (Johnson & Wichern, 2007).

Suppose there are *p* variable observations, given a random vector $\boldsymbol{X}' = [X_1, X_2, \dots, X_P]$ which has a covariance matrix $\boldsymbol{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Consider the following linear combination:

$$
\begin{aligned}
Y_1 &= \boldsymbol{a_1'X} = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \cdots + a_{1p}X_p \\
Y_2 &= \boldsymbol{a_2'X} = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \cdots + a_{2p}X_p \\
Y_3 &= \boldsymbol{a_3'X} = a_{31}X_1 + a_{32}X_2 + a_{33}X_3 + \cdots + a_{3p}X_p \\
&\vdots \\
Y_p &= \boldsymbol{a_p'X} = a_{p1}X_1 + a_{p2}X_2 + a_{p3}X_3 + \cdots + a_{pp}X_p
\end{aligned}
$$

then

$$Var(Y_i) = \boldsymbol{a_i'\Sigma a_i} \quad ; \ i = 1,2,3,\dots,p \tag{2.11}$$

$$Cov(Y_i, Y_k) = \boldsymbol{a_i'\Sigma a_k} \quad ; \ i = 1,2,3,\dots,p \tag{2.12}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix or can be replaced by the correlation matrix ρ and $\boldsymbol{a_i}$ is the eigenvectors associated **with** λ_i. The principal components formed are uncorrelated linear combinations $Y_1, Y_2, \dots, Y_P$ whose variance in (2.17) is maximized. In general, the i-th principal component is the linear combination $\boldsymbol{a_i'X}$ that maximizes $Var(\boldsymbol{a_i'X})$ against the constraints $\boldsymbol{a_i'a_i} = 1$ and $Cov(\boldsymbol{a_i'X}, \boldsymbol{a_k'X}) = 0$ for k < I which means

- Principal component 1 is a linear combination $\boldsymbol{a_1'X}$ that maximizes $Var(\boldsymbol{a_1'X})$ against the constraint $\boldsymbol{a_1'a_1} = 1$.
- Principal component 2 is the linear combination $\boldsymbol{a_2'X}$ that maximizes $Var(\boldsymbol{a_2'X})$ against the constraints $\boldsymbol{a_2'a_2} = 1$ and $Cov(\boldsymbol{a_1'X}, \boldsymbol{a_2'X}) = 0$**.**
- Principal component 3 is the linear combination $\boldsymbol{a_3'X}$ that maximizes $Var(\boldsymbol{a_3'X})$ against the constraints $\boldsymbol{a_3'a_3} = 1$ and $Cov(\boldsymbol{a_1'X}, \boldsymbol{a_2'X}) = 0$**,** $Cov(\boldsymbol{a_1'X}, \boldsymbol{a_3'X}) = 0$, and $Cov(\boldsymbol{a_2'X}, \boldsymbol{a_3'X}) = 0$**.**
- The principal component *i* is a linear combination $\boldsymbol{a_i'X}$ that maximizes $Var(\boldsymbol{a_i'X})$ against the constraints $\boldsymbol{a_i'a_i} = 1$ and $Cov(\boldsymbol{a_i'X}, \boldsymbol{a_k'X}) = 0$ for k < 1.

Suppose the eigenvalues and eigenvectors are pairwise $(\lambda_1, \boldsymbol{a_1}), (\lambda_2, \boldsymbol{a_2}), \dots, (\lambda_p, \boldsymbol{a_p})$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then:

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

The part of the total population contributed by the *k*th principal component is the value:

$$\begin{pmatrix} Proportion\ of\ total\ population \\ variance\ due\ to\ the \\ kth\ principal\ component \end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1,2, \ldots, p.$$

If $Y_1 = \boldsymbol{a_1'X}$, $Y_1 = \boldsymbol{a_1'X}$,…., $Y_1 = \boldsymbol{a_1'X}$ is the principal component obtained from the covariance matrix $\boldsymbol{\Sigma}$, then to calculate the correlation between the i-th principal component and the kth X variable is:

$$\rho Y_i, X_k = \frac{a_{ik}\sqrt{\lambda_1}}{\sqrt{\sigma_{kk}}} \qquad k = 1,2, \ldots, p \tag{2.13}$$

where $(\lambda_1, \boldsymbol{a_1}), (\lambda_2, \boldsymbol{a_2}), \ldots, (\lambda_p, \boldsymbol{a_p})$ is the pair of eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ (Johnson & Wichern, 2007).

According to Johnson & Wichern (2007), the criteria for selecting *k* is that the main component is considered to explain the diversity of the original data well if the cumulative proportion of the original data diversity explained by the main component is at least 80%.

III. RESEARCH METHODOLOGY

3.1 Time and Place of Research

This research was conducted in the odd semester of the 2024/2025 academic year and took place at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Lampung.

3.2 Research Data

The data used in this research is secondary data, that is data on women's participation in development by Regency/City in West Java Province in 2021-2023, which is obtained from the official website of the West Java Central Statistics Agency (BPS), that is, https://jabar.bps.go.id/id. In this research, there are 27 samples (Regency/City) with five (5) variables used:

Table 1. Research Data Variables

Variables	Description
X_1	Women's involvement in parliament (%)
X_2	Women as professionals (%)
X_3	Women's income contribution (%)
X_4	Gender Empowerment Index
X_5	Gender Development Index

3.3 Research Methods

This research was conducted by literature study, that is textbooks, journals and internet access that support the research process. The steps taken are as follows:

1. Descriptive analysis of data for 2021.
2. Test the assumption of non-multicollinearity with the Variance Inflation Factor (VIF).
3. Solving data containing multicollinearity by using Principal Component Analysis (PCA).
4. Standardize data into Z-Score values.
5. Perform clustering with Ward's method using principal component scores.
6. Perform clustering with the K-Means method using the principal component score.
7. Calculate the Davies-Bouldin index for each method.
8. Calculate the Calinski-Harabasz index for each method.
9. Repeat step 1 (one) to step 9 (nine) for the data 2022 and 2023.
10. Evaluate the Davies-Bouldin index and Calinski-Harabasz index.
11. Analysis of results.

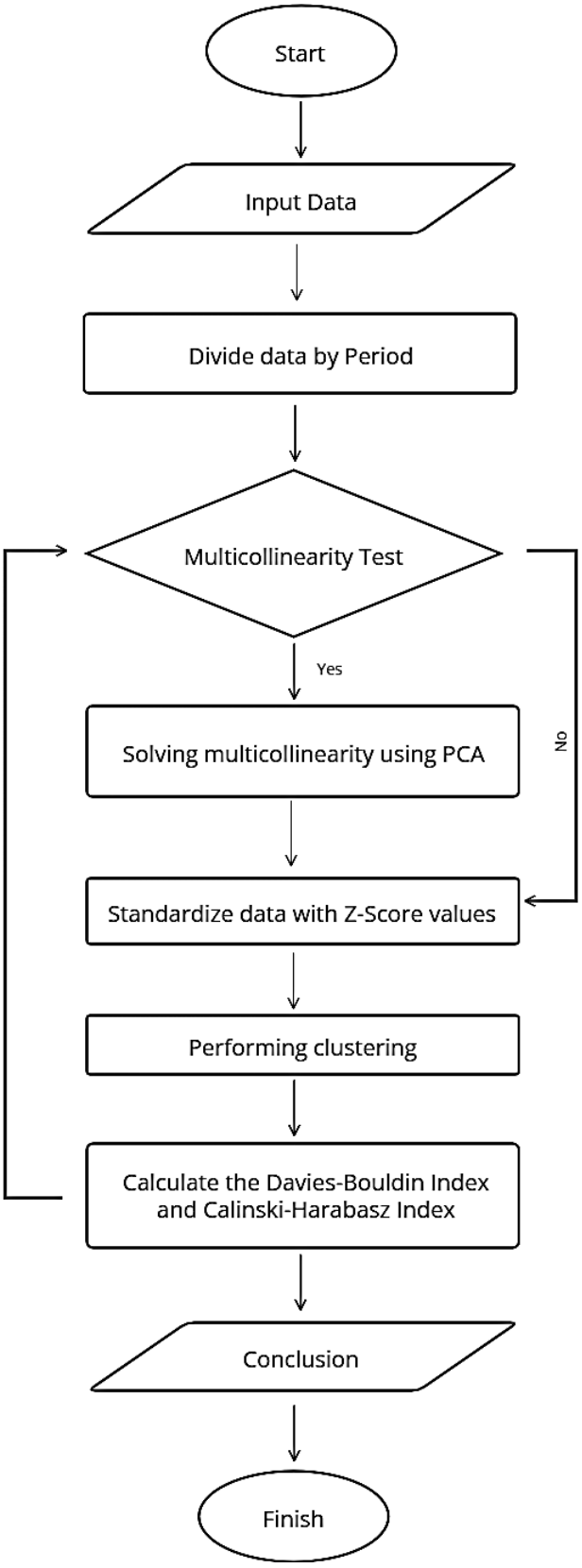

Figure 1. Research Flowchart

previous cluster. Cluster 3 consists of regencies/cities that have the highest level of women's participation than the other clusters.

3. During 2021-2023, using the average value of the three years, based on the DBI and CHI index values, the Ward and K-Means methods provide the same optimal clustering results, 4 clusters.

4. The clustering results show a change in the clustering pattern of regency/city from year to year. However, there are several regencies/cities that are consistently in the same cluster, specifically Sukabumi, Bogor, Cianjur, and Bekasi, which are in cluster 1. This means that these regencies have the lowest level of women's participation in development in West Java. Meanwhile, the regencies/cities with the highest level of women's participation in development in West Java are Sumedang and Bandung City.

5.2 Recommendation

In future research, clustering analysis of an area can be developed using other clustering methods. Further studies can be conducted to look at other factors that may affect the results of this clustering. Regions in cluster one need to get more attention in women's empowerment policies in West Java, for example through increasing women's involvement in parliament and access to professionals. By understanding this pattern, it is expected that the West Java regional government can make more specific policies in accordance with the characteristics of each cluster.

V. CONCLUSION

5.1 Conclusion

Based on the results and discussion described in Chapter IV, the following conclusions can be concluded:

1. In 2021, based on the Davies-Bouldin (DBI) and Calinski-Harabasz (CHI) index values, the Ward and K-Means methods provide the same optimal clustering results, 4 clusters. The results of clustering 27 regencies/cities in West Java based on women's participation in development show that cluster 1 consists of regencies/cities that have lower levels of women's participation than others. Cluster 2 consists of a regency/city that has higher levels of women's participation than the previous cluster. Cluster 3 consists of regencies/cities that have slightly higher levels of women's participation than other clusters. Meanwhile, cluster 4 consists of regency/cities that have the highest level of women's participation in development compared to other clusters.

2. In 2022 and 2023, based on the DBI and CHI values, the Ward and K-Means methods provide the same optimal cluster results, 3 clusters. The results of clustering 27 regencies/cities in West Java based on women's participation in development show that cluster 1 consists of regencies/cities that have lower levels of women's participation than other clusters. Cluster 2 consists of regencies/cities with higher levels of women's participation than the

previous cluster. Cluster 3 consists of regencies/cities that have the highest level of women's participation than the other clusters.

3. During 2021-2023, using the average value of the three years, based on the DBI and CHI index values, the Ward and K-Means methods provide the same optimal clustering results, 4 clusters.

4. The clustering results show a change in the clustering pattern of regency/city from year to year. However, there are several regencies/cities that are consistently in the same cluster, specifically Sukabumi, Bogor, Cianjur, and Bekasi, which are in cluster 1. This means that these regencies have the lowest level of women's participation in development in West Java. Meanwhile, the regencies/cities with the highest level of women's participation in development in West Java are Sumedang and Bandung City.

5.2 Recommendation

In future research, clustering analysis of an area can be developed using other clustering methods. Further studies can be conducted to look at other factors that may affect the results of this clustering. Regions in cluster one need to get more attention in women's empowerment policies in West Java, for example through increasing women's involvement in parliament and access to professionals. By understanding this pattern, it is expected that the West Java regional government can make more specific policies in accordance with the characteristics of each cluster.

REFERENCES

Anonim. 2024. *Optimalisasi Peran Perempuan dalam Pembangunan Kementerian Koordinator Bidang Pembangunan Manusia dan Kebudayaan*. Retrieved October 29, 2024, from https://www.kemenkopmk.go.id/optimalisasi-peran-perempuan-dalam-pembangunan

Ansari, Z., Azeem, M. F., Ahmed, W., & Babu, A. V. 2011. Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions. *World of Computer Science and Information Technology Journal (WCSIT)*. **1**(5): 217–226. https://arxiv.org/pdf/1507.03340

Baroroh, A. 2012. *Analisis Multivariat dan Time Series dengan SPSS 21*. PT. Elek Media Komputindo.

Caliński, T., & JA, H. 1974. A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods*. **3**: 1–27. https://doi.org/10.1080/03610927408827101

De Amorim, R. C. 2015. Feature Relevance in Ward's Hierarchical Clustering Using the L p Norm. *Journal of Classification*. **32**(1): 46–62. https://doi.org/10.1007/s00357-015-9167-1

Draper, N. R., & Smith, H. 1998. *Applied Regression Analysis* (Third). A Wiley-Interscience Publication.

Escobar, A. 1995. *Encountering Development: The Making and the Unmaking of the Third World*. Princeton University Press.

Fathia, A. N., Rahmawati, R., & Tarno, T. 2016. Analisis Klaster Kecamatan di Kabupaten Semarang Berdasarkan Potensi Desa Menggunakan Metode Ward dan Single Linkage. *Jurnal Gaussian*. **5**(4): Article 4. https://doi.org/10.14710/j.gauss.5.4.801-810

Gudono. 2014. *Analisis Data Multivariat*. BPFE-Yogyakarta.

Hasan, Y. 2024. Pengukuran Silhoutte Score dan Davies-Bouldin Index Pada Hasil Cluster K-Means dan DBSCAN. *Jurnal Informatika Dan Teknik Elektro Terapan*. **12**(3S1): Article 3S1. https://doi.org/10.23960/jitet.v12i3S1.5001

Iklima, T., & Pujiyanta, A. 2023. Perbandingan Metode K-Means Clustering Dan Metode Ward Dalam Mengelompokkan Pelangan Mall. *Jurnal FASILKOM*. **13**(3): 349–357. https://doi.org/10.37859/jf.v13i3.6040

Johnson, R. A., & Wichern, D. W. 2007. *Applied Multivariate Statistical Analysis* (Sixth Edition). Pearson International Edition.

Nugroho, S. 2008. *Statistika Multivariat Terapan* (Pertama). UNIB Press.

Ramadhan, R., Awalluddin, A., & Cahyandari, R. 2020. Multivariable Panel Data Cluster Analysis using Ward Method Gross Enrollment Ratio (GER) Data in West Java in the Year 2015-2018. *Proceeding International Conference on Science and Engineering*. **3:** 291–296. https://doi.org/10.14421/icse.v3.515

Rencher, A. C. 2002. *Methods of Multivariate Analysis* (Second). A John Wiley & Sons, Inc. Publication.

Sikana, A. M., & Wijayanto, A. W. 2024. Analisis Perbandingan Pengelompokan Indeks Pembangunan Manusia Indonesia Tahun 2019 dengan Metode Partitioning dan Hierarchical Clustering. *ResearchGate*. **14**(2). https://doi.org/10.24843/JIK.2021.v14.i02.p01

Simamora. 2005. *Analisis Multivariat Pemasaran* (Pertama). PT. Gramedia Pustaka Tama.

Usman, H., & Nurdin, S. 2013. *Aplikasi Teknik Multivariate untuk Riset Pemasaran*. PT. Raja Grafindo Persada.

Wang, W., & Lu, Y. 2021. Application of Clustering Analysis of Panel Data in Economic and Social Research Based on R Software. *Academic Journal of Business & Management*. **3**(10). https://doi.org/10.25236/AJBM.2021.031018

Widarjono, A. 2010. *Analisis Statistika Multivariat Terapan*. UPP STIM YKPN.

Wijaya, T., & Budiman, S. 2016. *Analisis Multivariat untuk Penelitian Manajemen*. Penerbit Pohon Cahaya.

Wira, B., Budianto, A. E., & Wiguna, A. S. 2019. Implementasi Metode K-Medoids Clustering untuk Mengetahui Pola Pemilihan Program Studi Mahasiswa Baru Tahun 2018 Di Universitas Kanjuruhan Malang. *RAINSTEK: Jurnal Terapan Sains Dan Teknologi*. **1**(3): Article 3. https://doi.org/10.21067/jtst.v1i3.3046