Bicluster Analysis of Cheng and Church's Algorithm to Identify Patterns of People's Welfare in Indonesia

Laradea Marifni[1*] , I Made Sumertajaya[2] , Utami Dyah Syafitri[3]

[1,2,3] *IPB University, Bogor, Indonesia*

*corr_author: laradeamarifni@apps.ipb.ac.id

Abstract – Biclustering is a method of grouping numerical data where rows and columns are grouped simultaneously. The Cheng and Church (CC) algorithm is one of the bi-clustering algorithms that try to find the maximum bi-cluster with a high similarity value, called MSR (Mean Square Residue). The association of rows and columns is called a bi-cluster if the MSR is lower than a predetermined threshold value (delta). Detection of people's welfare in Indonesia using Bi-Clustering is essential to get an overview of the characteristics of people's interest in each province in Indonesia. Bi-Clustering using the CC algorithm requires a threshold value (delta) determined by finding the MSR value of the actual data. The threshold value (delta) must be smaller than the MSR of the actual data. This study's threshold values are 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8. After evaluating the optimum delta by considering the MSR value and the bi-cluster formed, the optimum delta is obtained as 0.1, with the number of bi-cluster included as 4.

Keywords: Bi-clustering; CC algorithm; MSR; Cheng and Church

I. INTRODUCTION

The Covid-19 pandemic, which has entered Indonesia since early 2020, has brought down almost all aspects of the economy. Until the late 2020s, at least 3 to 4 million people lost their jobs. The potential for unemployment is predicted to reach 4-5 million people by 2021 [1]. Reflecting on these conditions, the presence of the state in various sectors of public life is very important. The form must protect the entire nation and bloodshed and promote the general welfare. Welfare has a dynamic, quantitative and relative meaning because its formulation will never be final and will continue to evolve in line with human needs. In general, welfare is the condition of all citizens who always have their material and spiritual needs met [2]. Therefore, the government must be able to guarantee public health, provide education for the community, provide jobs, provide facilities and infrastructure to support community activities, provide a clean environment, and create security for society equally for all people. [3] issues an annual publication on community welfare indicators. In this study, clustering was carried out in 34 provinces in Indonesia based on similarities in the characteristics of the people's welfare in 2020. After that, a mapping and profile analysis was carried out on the cluster results.

Cluster analysis is used for grouping different observations based on their similar characteristics [4]. Cluster analysis generally used is one-way clustering which assumes that things have similar features in all rows or columns so that objects in rows are grouped based on similarity in columns or variables in columns are grouped based on similarities in rows. Ref. [5] has carried out one-way clustering (classical clustering) separately by grouping objects in rows using a distance measure matrix and then grouping variables in columns using a correlation matrix. Clustering like this still has limitations for two-way data that want to know the relationship of a certain group of objects with a certain group of variables.

This two-way clustering method is not new that has introduced two-way clustering on matrix data [6]. Another study of two-way clustering was also implemented by [7] by clustering rows and columns together on matrix data, known as bicluster. Then it was reapplied by [8] on the gene expression data matrix. If clustering of the gene expression data matrix is performed using classical clustering, the gene or condition clustering results will be imprecise. Because the clustering of each gene is defined using all states, and the activity of all genes characterizes the clustering of conditions. So what is obtained is a global cluster, which causes many gene activation patterns only under certain experimental conditions, and a lot of information is ruled out. Bicluster analysis has advantages in terms of robustness, which is better than one-way clustering because it involves two object characteristics in rows and columns together [9]. In addition, bicluster analysis can identify subgroups of rows or columns that are interrelated and are not found in one-way clustering [10]. This approach's main idea is to partition the X(I, J) matrix by arranging I rows and J

columns into submatrices X(P, Q) with associated rows P and columns Q.

Initially, bicluster was applied to biological data to analyze gene expression microarray data. However, it has been widely used in other fields, such as social data, to identify patterns in people's welfare, health, poverty, education, food insecurity, etc. One of the basic, popular, and widely applied bicluster algorithms is Cheng and Church (CC). This algorithm aims to find bi-clusters with a Mean Square Residual (MSR) smaller than the specified threshold. Ref. [11] applied the bicluster method to group districts/cities in Indonesia using various indicators of Indonesia's readiness to face the industrial revolution 4.0. This study used the CC algorithm and produced five biclusters with the best parameter tuning based on MSR/V, namely δ=0.004123 (80% residue). Ref. [12] also applied bicluster analysis with the Cheng and Church (CC) algorithm approach and produced the best tuning parameter, namely δ=0.4 based on the smallest MSR/V (the average ratio of residue over volume) value (0.0021) which resulted in 2 biclusters. Ref. [13] also conducted a bicluster analysis study using the Cheng and Church algorithm to identify patterns of food insecurity in Indonesia. This study produced five biclusters with the best parameter tuning based on the smallest MSR/V (0.017), namely δ=0.1.

However, in biclustering analysis, there are no guidelines for choosing the right bicluster algorithm for certain data criteria because the algorithm in biclustering is not specific [14]. The algorithm's success in previous research in determining the optimal cluster can be a good criterion in deciding the algorithm to be used. In various studies, the CC algorithm is an algorithm that is often used and is a popular algorithm because it is the basic algorithm in the bicluster. The CC algorithm is interesting to apply to Indonesian social welfare data because many previous researchers have obtained optimal clustering using the CC algorithm on social data. Therefore, this study aims to apply the CC bicluster algorithm to public welfare data in Indonesia. In addition, this study also discusses the results of the CC algorithm in this case and produces spatial patterns of Indonesian people's welfare. The results of this study are expected to be useful for policymakers through the spatial pattern detection results obtained.

II. METHOD

A. Data

The data used is sourced from the official website of the Central Bureau of Statistics, namely https://www.bps.go.id. The data selected is based on the annual publication of people's welfare indicators in 2020, such as Population, Health and Nutrition, Education, Employment, Consumption Levels and Patterns, Housing and Environment, Poverty, and other social which are references in efforts to improve the quality of life by province [3]. The study began by preparing data for 36 variables and 34 provinces. The data is presented in Table I.

B. Research Stages

1) Preprocessing: There are three steps at this stage: 1) forming a data matrix from 34 provinces as rows and 36 variables as columns, 2) carrying out a standard normalization approach to the data matrix, and 3) exploring the data to see the initial characteristics of the data matrix which is scaled using a heat map.

2) Cheng and Church's Algorithm: Cheng and Church (CC) algorithm tries to find the maximum bi-cluster with a high similarity value, so CC algorithm is designed as a greedy algorithm. This equivalence value is called MSR (Mean Square Residue), and the association of rows and columns is called a bi-cluster if this value is lower than a predefined threshold value (delta). Bi-clustering using the CC algorithm is depicted through the flow chart in Fig. 1. In general, through Fig. 1, it can be seen that there are three phases in CC algorithm, that are the deletion, addition, and substitution phases [7], [15], [16]. The deletion and addition phases are performed iteratively until there are no more processes that can reduce the MSR value. Both phases aim to reduce the MSR so that a bi-cluster with an MSR value smaller than the predetermined delta threshold is obtained. Meanwhile, the substitution phase is done by closing the sub matrix (bi-cluster) with a disk number. This phase aims to prevent overlap between the resulting bi-clusters.

The description for each stage of the Cheng and Church (CC) algorithm flow chart in Fig. 1 is as follows:

- Determining the threshold (δ)
  1. Calculate the MSR value of the overall data, called MSR_A.
  2. Determine delta (δ) under the condition $\delta < MSR_A$
- Finding bi-cluster with CC algorithm (1)
  1. k=1
  2. The overall input data is a matrix called matrix A_{ijk} and the threshold values (δ)
  3. Calculate the MSR value of matrix A_{ijk}

TABLE I
VARIABLES OF INDONESIAN PEOPLE'S WELFARE

Dimensions	Variable
Population	Total Population (X1); Population Growth Rate (X2); Sex Ratio (X3); Population Density (X4); Percentage of population aged 0-14 (X5); Percentage of population aged 15-64 (X6); Percentage of population aged 65+ (X7); Percentage of population (X8)
Social	People who have been victims of crime (X9); Households receiving business loans (X10); Households receiving health care insurance (X11); Households with internet access (X12); Population traveling (X13)
Housing	Households by Ownership Status Owned house (X14); Households with Access to Decent Water (X15); Households with Electric Lighting Source (X16); Households with Owned Closet with Septic Tank (X17); Average Floor Area per Capita of Dwellings (X18)
Consumption	Protein Consumption per Capita per Day (X19)
Employment	Child Labor (Age 10-17 Years) (X20); TPAK (labor force participation rate) (X21); TPT (open unemployment rate) (X22)
Education	School Enrollment Rate (X23); Pure Enrollment Rate of High School (X24); Highest education completed (X25); Literacy rate of population aged 15-24 (X26)
Health	Percentage of Toddlers Who Have Received Measles Immunization (X27); Special Index for Stunting Management (X28); AKB (Infant Mortality Rate) (X29); AKABA (Under 5 Mortality Rate) (X30); AHH (Life Expectancy Rate) (X31)
Poverty	Poverty depth index (X32); Non-food Poverty Line (X33); Food Poverty Line (X34); Percentage of Poor Population (X35); Number of Poor Population (X36)

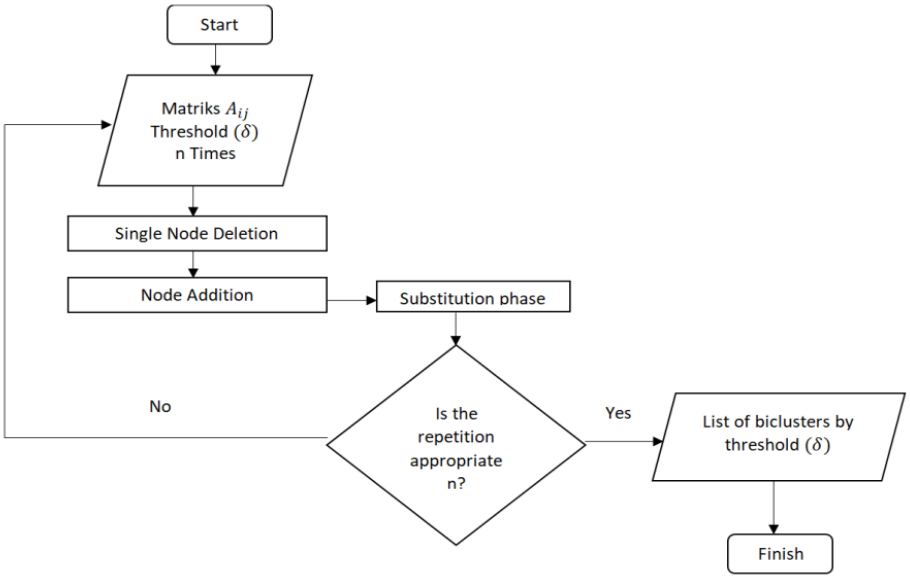

Fig. 1 Flowchart of CC (Cheng and Church) algorithm

$$MSR(\mathrm{A}_{ijk}) = \frac{1}{|I||J|}\sum_{i\in I, j\in J}\left(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}\right)^2 \quad (1)$$

a_{ij} : data row i column j
a_{iJ} : average in row i
a_{Ij} : average in column j
a_{IJ} : average across the bi-cluster

4. Is $MSR(\mathrm{A}_{ijk}) < \delta$?
   Yes, go to point 16
   No, go to point 5 (Single Node Deletion)

Single Node Deletion (5-10)

5. Calculate the row value $d(i)$ and column value $d(j)$ with the following (2) and (3).

$$d(i) = \frac{1}{|J|}\sum_{j\in J}\left(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}\right)^2$$

$$d(j) = \frac{1}{|I|}\sum_{i\in I}\left(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}\right)^2$$

6. Delete row $d(i)$ and column $d(j)$ as follows
   $MSR(\mathrm{A}_{ijk}) < d(i)$
   $MSR(\mathrm{A}_{ijk}) < d(j)$
7. K=K+1, go back to step 3
8. If no more rows or columns are deleted, then k=1 and a new matrix is formed, named matrix B_{ijk} .
9. Calculate the value of MSR (B_{ijk})
10. Is $MSR(\mathrm{B}_{ijk}) < \delta$?
    Yes, go to point 16
    No, go to point 11 (Node Addition)

Node Addition (11-15)

11. Calculate the row value $d(i \notin I)$ and column value $d(j \notin J)$ with the following (4) and (5)

$$d(i \notin I) = \frac{1}{|J|}\sum_{j \in J}\left(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}\right)^2 \quad (4)$$

$$d(j \notin J) = \frac{1}{|I|}\sum_{i \in I}\left(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}\right)^2 \quad (5)$$

12. Add row $(i \notin I)$ and column $d(j \notin J)$ as follows
    $MSR(\mathrm{B}_{ijk}) > d(i)$
    $MSR(\mathrm{B}_{ijk}) > d(j)$
13. k=k+1, go back to step 9
14. If there is nothing more to add to the iteration, then refer to the names I and J as I' and J'. The result of adding nodes to the algorithm is called the C matrix
15. The iterative deletion and addition phase stops $MSR(\mathrm{C}_{ij}) < \delta$.

Subtitution (16)

16. Bi-cluster C is formed, and replace the elements in A that have entered bi-cluster C with random values.
17. Repeating points 1 to 16 n times, that is, as many as n bi-cluster are to be found.

3) Evaluation to Find the Optimum Delta: The bi-cluster formed at each threshold type was evaluated using the mean value of the MSR. One kind of threshold is then selected to further analyze each bicluster's characteristics. From the specified/optimum CC threshold bi-clustering results, the pattern of people's welfare in Indonesia can then be depicted as a map. The performance of the bi-clustering algorithm can be evaluated using Mean Squared Residue (MSR).

Where a_{IJ} is the average across the bi-cluster, a_{Ij} is the average in column j, a_{iJ} is the average in row i, $I \times J$ is the bi-cluster dimension, which is the bi-cluster row size (I) which is the bi-cluster row size (J). $MSR_{(I,J)}$ represents the variation associated with the interaction between rows and columns in the bi-cluster [13], [14]. According to [15], the quality of a bi-cluster will be better as the residual value decreases and/or the volume (rows × columns) of the bi-cluster increases. The quality of a bicluster improves as the residual value (MSR) of the bi-cluster decreases and is obtained by calculating the average of the residuals $\left(\frac{1}{n}\sum_{i=1}^{n} MSR_i\right)$.

III. RESULTS AND DISCUSSION

A. Data Exploration

After the data is scaled, a heatmap is then carried out on the data by displaying data with different color representations. An initial overview of the scale data matrix is presented in the heatmap diagram in Fig. 2.

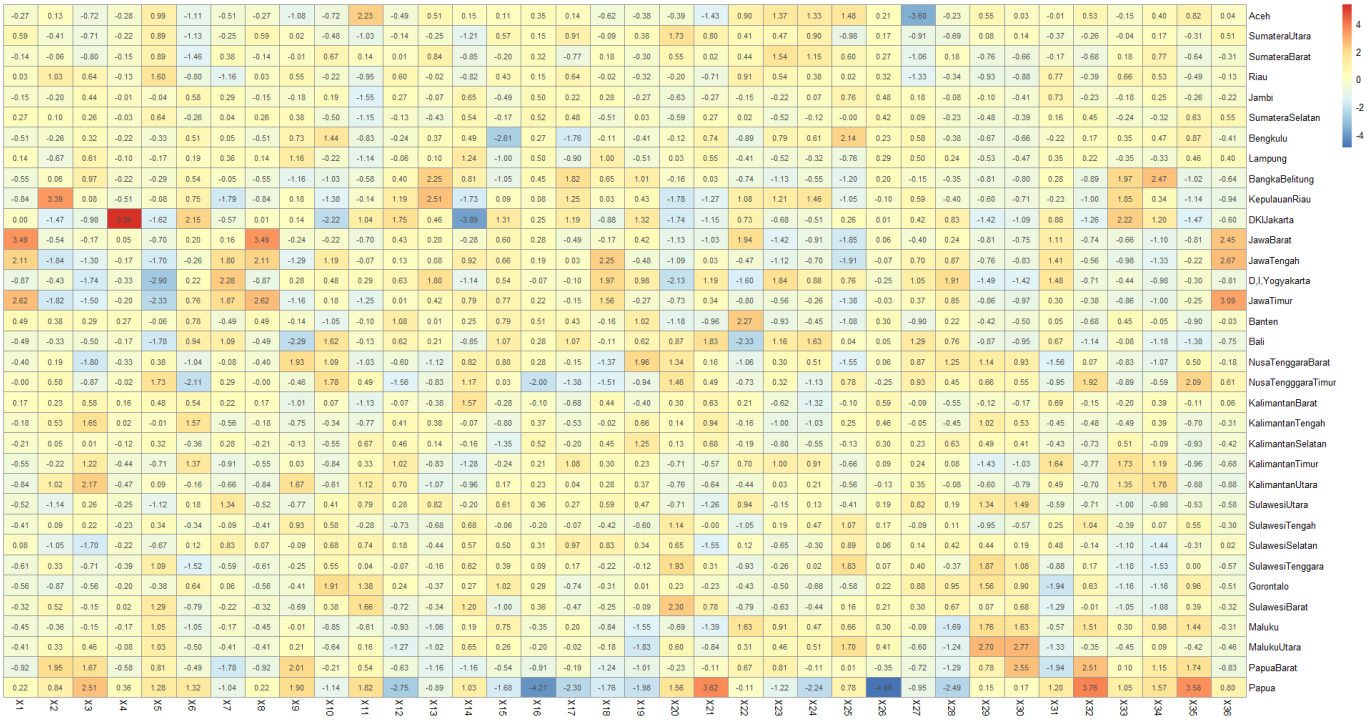

Fig. 2 Scaled Data Matrix Heatmap

The heatmap matrix in Fig. 2 shows that the level of welfare of Indonesian people based on 34 provinces in Indonesia in 2020 is divided into three color categories, namely provinces with high (red), medium (white) and low (blue) levels of people's welfare variables.

As an illustration, the province of DKI Jakarta has some very prominent changes, such as Population Density (X4) and Percentage of population aged 15-64 years (X6), meaning that the population density in DKI Jakarta and the population aged 15-64 years are higher than other provinces. This is certainly reasonable because DKI Jakarta is the capital city of Indonesia so that many people from other regions come to Jakarta in order to work, study and others. However, DKI Jakarta also has some low-value indicators such as Households by Ownership Status of Owned House (X14). It means that, in Jakarta, there are still many people who live in houses that do not belong to them such as rent, boarding houses, and so on. It is in contrast to the provinces of West Java, Central Java and East Java, which show more population (X1) and the number of poor people (X36) compared to other provinces. Meanwhile, Riau Islands Province shows a high value in the variable Population Growth Rate (X2).

Aceh Province has a high X11 variable, meaning that Aceh Province has more households insured for health services than other provinces. However, Aceh Province also has a low percentage of children under five who have received measles immunization (X27). Papua Province has several variables with high values, such as poverty depth index (X32), labor force participation rate (X21) and percentage of poor people (X35). However, there are also variables that have low values, such as the Literacy Rate of the 15-24 Year Old Population (X26), Households with an Electric Lighting Source (X16) and Households with Internet Access (X12). It means that the level of welfare in Papua province is still relatively low.

Overall, the heat map matrix above also illustrates that almost all Indonesian provinces have welfare levels that are already classified as medium to high, as seen from the dominance of white and red colors in the heat map. All of the variables used also have the potential to be found in a number of provinces in Indonesia, such as the variable X1, which is high in the provinces of West Java, Java and East Java; X27, which is low in the province of Aceh; and others. However, in terms of provinces, not all provinces in Indonesia have a high level of welfare.

In order to clarify the characteristics of each region or group of regions, further exploration will be conducted using the PCA Biplot presented in Fig. 3.

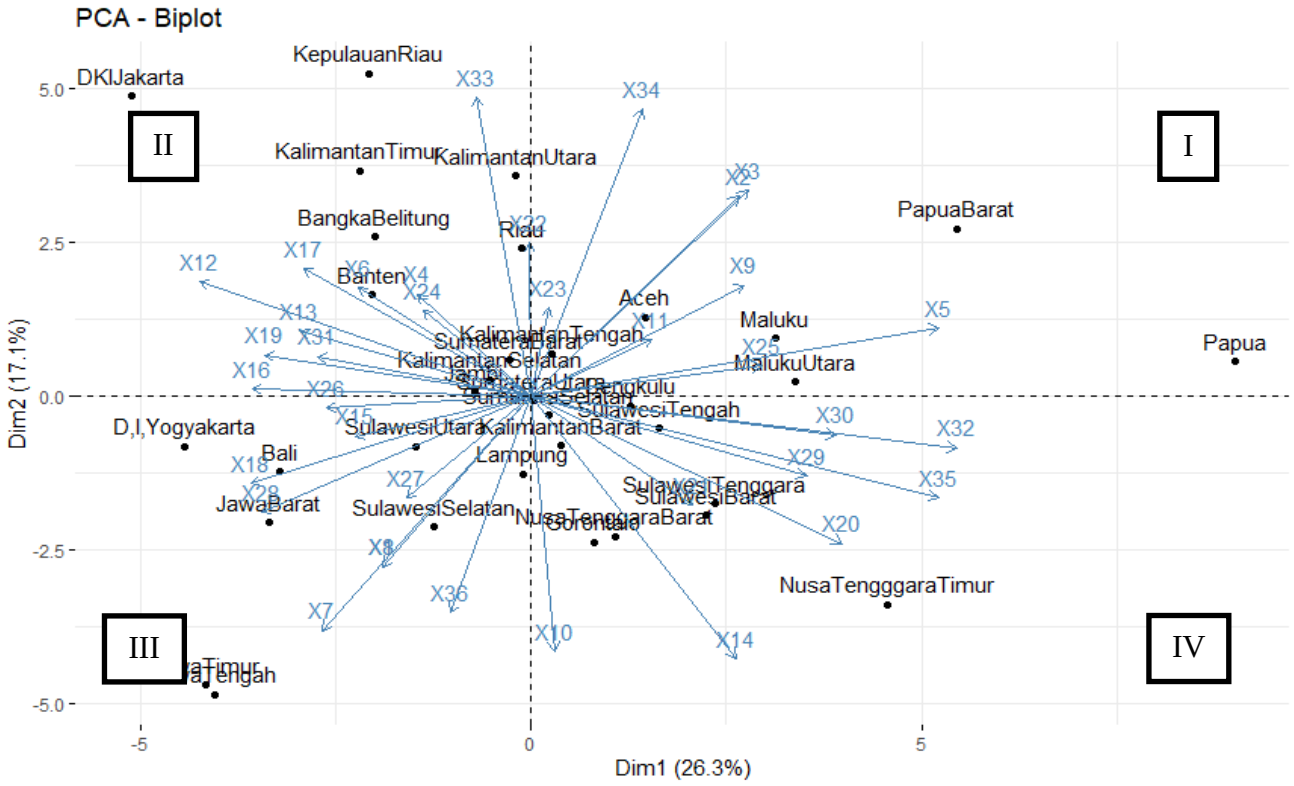

Fig. 3 Biplot Results of PCA Matrix Data Scale in Quadrant

The PCA biplot in Fig. 3 reduces the large-dimensional space of 36 variables to a two-dimensional space of dim1 and dim2. This results in less information that can be explained by the PCA biplot. The PCA biplot is only able to explain 43.4% of the diversity of the data matrix; the rest cannot be explained by this biplot. This shows that the diversity of the original variables cannot be explained well with only these two dimensions. Therefore, the PCA biplot is only used to provide an initial overview of the regional groupings and their characteristics.

Fig. 3 shows that of the 36 variables, the highest diversity is X14, X32, X33, X34 and X35. The variable X14 is Households with own ownership status, meaning that in Indonesia households with own ownership status still vary. Whereas X32, X33, X34 and X35 are variables from the Poverty Dimension, this also indicates that Poverty in Indonesia has a high level of diversity. Conversely, there are several variables that have a low level of diversity such as X15, X23, X24 and X25. The variable X15 is Households with access to decent water, which means that access to decent water is almost evenly distributed across provinces in Indonesia. Meanwhile, X23, X24 and X24 are variables from the Education dimension, meaning that education has also been evenly obtained in the provinces in Indonesia.

The provinces can be grouped into 4 quadrants; quadrant I consists of the provinces of West Papua, Papua, Maluku, North Maluku, Aceh and Central Kalimantan which are characterized by variables X2, X3, X5, X9, X11, X25, and X34. Quadrant II consists of DKI Jakarta, Riau Islands, East Kalimantan, North Kalimantan, Riau, Bangka Belitung, Banten, West Sumatra, South Kalimantan, and Jambi with variables characterized by X22, X33, X24, X4, X6, X17, X12, X13, X31, X19, X16 and X26. In Quadrant III, there are provinces of Yogyakarta, Bali, West Java, North Sulawesi, Lampung, East Java, Central Java and South Sulawesi which are characterized by variables X3, X15, X18, X27, X28 and X7. Quadrant IV consists of the provinces of NTT, NTB, West Sulawesi, Southeast Sulawesi, West Kalimantan, South Sumatra, Bengkulu, North Sumatra and Gorontalo with variables characterized by X10, X14, X20, X29, X30, X32 and X35. Each Quadrant has distinctive characteristics and different characterizing variables. One Quadrant is not explained by all variables, but only certain variables. Since the PCA Biplot is only able to explain 45.3% of the diversity of the data matrix, it is necessary to do bi-clustering to properly group regions or provinces.

B. Bi-cluster Analysis

Bi-cluster analysis has an approach by applying the algorithm on several parameters and tuning the parameters manually. The smallest average MSR/V value considering the number of clusters resulting from bi-cluster analysis is the best object clustering result. The CC algorithm will produce the largest possible bi-cluster with a low MSR value. MSR=0.098.

Table II shows that there is no linear relationship between the tuning parameter and the MSR result of the bi-cluster. The optimal bi-cluster pattern is the one with the smallest MSR value, which is at the tuning parameter δ equal to 0.1 with an MSR value of 0.098 by forming 4 bi-clusters with the following details.

TABLE II
PARAMETER TUNING RESULTS

Delta	Number of Bi-cluster	MSR
0.1	4	0,098226
0.2	4	0,182627
0.3	3	0,288207
0.4	3	0,491261
0.5	3	0,389297
0.6	3	0,581858
0.7	2	0,464434
0.8	2	0,749973

TABLE III
CHARACTERISTICS OF EACH BI-CLUSTER

Bi-cluster Identity	Bi-cluster Size	Province	Characteristics	
			Low	High
Bicluster 1	15 x 12	Aceh, North Sumatra, West Sumatra, Jambi, South Sumatra, Bengkulu, Lampung, Banten, West Kalimantan, Central Kalimantan, South Kalimantan, Central Sulawesi, Southeast Sulawesi, West Sulawesi, North Maluku	X1, X4, X7, X8, X12, X13, X18, X28, X33, X36	X16, X26
Bicluster 2	9 x10	Kepulauan Riau, West Java, Central Java, East Java, Bali, East Kalimantan, North Kalimantan, North Sulawesi, South Sulawesi	X4, X32, X35	X12, X15, X16, X19, X26, X27, X28
Bicluster 3	7 x 7	Riau, Bangka Belitung, DKI Jakarta, West Nusa Tenggara, Gorontalo, Maluku, West Papua	X1, X7, X8, X24, X36	X16, X26
Bicluster 4	3 x 10	Yogyakarta, East Nusa Tenggara, Papua	X1, X4, X8, X22, X29, X30	X2, X11, X14, X36

Table III shows that the number of clusters in bi-cluster 1 is 15 provinces and 12 variables, bi-cluster 2 has 9 provinces with 10 variables, bi-cluster 3 has 7 provinces with 7 variables, and bi-cluster 4 has 3 provinces and 10 variables. In the formation of bi-cluster, it is possible to have variables that overlap between bi-clusters and it is also possible that there are objects or provinces that cannot be categorized in the bi-cluster, but in bi-cluster with $\delta = 0.1$ all provinces are included in each of the existing bi-cluster.

Bi-cluster 1 has similar welfare characteristics such as Population Size, Population Density, Percentage of Population aged 65+, Households with internet and electricity access, Literacy Rate of Population aged 15-24, and Stunting and Poverty index. Bi-cluster 2 has similar welfare characteristics such as Population Density, Households with internet access, access to decent water and electricity, and poverty depth index. Bi-cluster 3 has similar welfare characteristics such as Population Size, Percentage of the Population aged 65+ years, Net enrollment rate at the high school level, Literacy Rate of the Population aged 15-24 years, and the number of poor people. Meanwhile, bi-cluster 4 has similar welfare characteristics such as Population Size, Population Growth Rate, Population Density, Households with Ownership status, Open Unemployment Rate, IMR, U5MR, and the number of poor people.

Bi-cluster 1 is characterized by several variables from the population, social, housing, education and poverty dimensions. The high value variables come from the housing and education dimensions, which are households with a source of electric lighting and literacy rate of the population aged 15-24 years. However, the provinces in bi-cluster 1 have a low health variable, which is the special index for handling stunting. Bi-cluster 2 is characterized by several variables from the population, social, housing, education, health and poverty dimensions. Provinces in bi-cluster 2 are characterized by households with high internet access, decent water and electricity. In the health dimension, there are also variables that have high

values such as the percentage of children under five who are immunized against measles and the special index for handling stunting. Meanwhile, the poverty dimension has a low poverty depth index variable and a low percentage of poor people. It means that the provinces in bi-cluster 2 have welfare characteristics that are high in the housing and health dimensions and low in the poverty dimension. Bi-cluster 3 is similar to bi-cluster 1 in that it is characterized by a high share of households with a source of electric lighting and a high literacy rate of the population aged 15-24 years; however, there is no variable from the health dimension that has a low value but rather the percentage of the population aged 65+ years in bi-cluster 3. Meanwhile, bi-cluster 4 is slightly different from the other three bi-clusters in the poverty dimension. In bi-cluster 4, the poverty variable has a high value, which means that provinces in bi-cluster 4 have a large number of poor people compared to provinces in other bi-clusters. However, in terms of the health dimension, there are AKABA (Under Five Mortality Rate) and AKB (Infant Mortality Rate) variables that are low in bi-cluster 4.

The distribution of bi-clusters can be depicted in the form of a map of Indonesia as shown in Fig. 4. Fig. 4 visually illustrates the bi-cluster into four colors that show the characteristics of the bi-cluster based on the welfare variables of the Indonesian people. The islands of Sumatra and Kalimantan offer dominance in Bi-cluster 1, except for Riau province, which is in Bi-cluster 3, and East Kalimantan, which is in Bi-cluster 2. It means that the islands of Sumatra and Kalimantan have similar welfare characteristics based on variables X1, X4, X7, X8, X12, X13, X16, X18, X26, X28, X33, X36. Meanwhile, Java's island shows dominance following bi-cluster 2, except for Yogyakarta and Banten, which are in bi-cluster 4 and 2, respectively. It means that apart from Yogyakarta and Banten, all provinces on the island of Java have similar welfare characteristics based on variables X4, 12, X15, X16, X19, X26, X27, X28, X32, X35. In contrast to Sulawesi, Sulawesi Island is more varied in that there are provinces in bi-cluster 1: Southeast Sulawesi, Central Sulawesi, and West Sulawesi. North Sulawesi and South Sulawesi provinces are in Bi-cluster 2, while there is only Gorontalo province from Sulawesi Island in Bi-cluster 3. The provinces in bi-cluster 3, such as Riau, Bangka Belitung, Jakarta, NTB, Gorontalo, Maluku, and West Papua, have similar welfare characteristics based on variables X1, X7, X8, X16, X24, X26, and X36. Meanwhile, those in bi-cluster 4, the provinces of Papua, Yogyakarta, and NTT, have similar welfare characteristics based on variables X1, X2, X4, X8, X11, X14, X22, X29, X30, X36.

In Fig. 5, the profile of the characteristics that characterize each bi-cluster appears to have a similar pattern. It shows that the Cheng and Church algorithm can cluster data on the welfare of Indonesian people based on the province well.

Fig. 4 Bi-cluster map of people's welfare in Indonesia

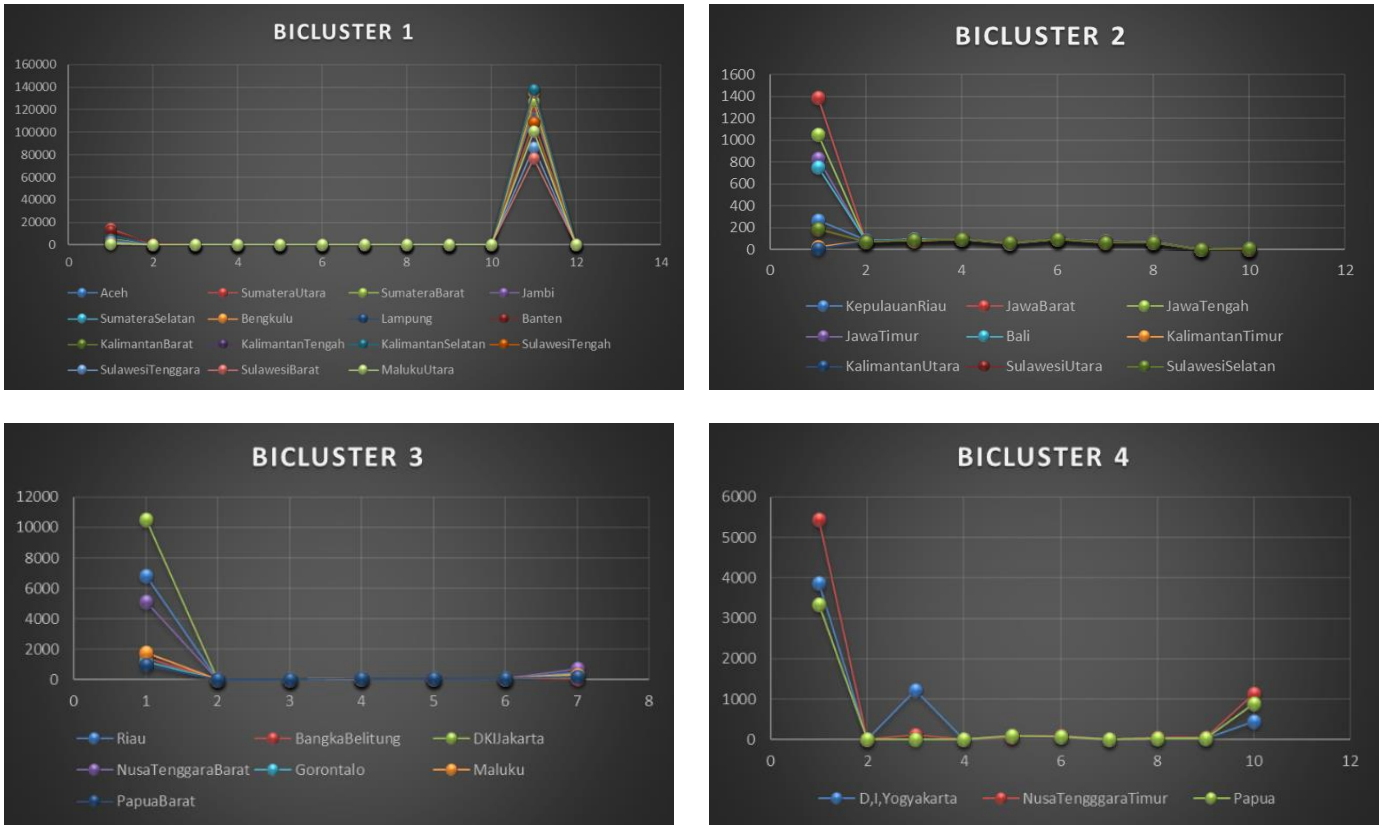

Fig. 5 Profile analysis of each bi-cluster

IV. CONCLUSION

Cheng and Church (CC) algorithm forms four bi-clusters with an optimum threshold value (delta) of 0.1 and an MSR value of 0.098. Bi-cluster 1 consists of 15 provinces with 12 variables, bi-cluster 2 consists of 9 provinces and ten variables, bi-cluster 3 consists of 7 provinces with 7 variables and bi-cluster 4 consists of 3 provinces with ten variables. This study uses a threshold value (delta) as a tuning parameter to obtain optimal bi-cluster results by considering the smallest MSR value. The selection of the threshold value is something that needs attention. Improper threshold values will result in less than optimal bi-clusters in clustering cases. The results of the study show that each province in the bi-cluster that is formed has different characteristics. The household variable with a source of electricity (X16) and the literacy rate for the population aged 14-24 (X26) always appear with high values in each bi-cluster, except for bi-cluster 4. It means that the province in bi-cluster 1, 2, and 3 already have a good source of electric lighting and good education. Whereas in bi-cluster 4, the variable Number of Poor Population (X36) looks high, and the IMR variable (X 29) and U5MR (X30) are low. It means that provinces in bi-cluster 4 need to get government attention, especially in the dimensions of poverty and health. For further research, we suggest investigating the profile of observations in provinces in bi-cluster 4. From the patterns found in each bi-cluster, it is expected that the government can pay attention to welfare in each region by paying attention to the dimensions appropriately. The Cheng and Church algorithm is an algorithm that produces constant bicluster types (for rows and columns), so further research is needed using other algorithms that make coherent bicluster types to understand better the structure of the dimensions that exist in the welfare of the Indonesian people.

REFERENCES

[1] Iskandar A Muhaimin. “Negara dan Politik Kesejahteraan: Reorientasi Arah Baru Pembangunan”. Jakarta: PT Gramedia Pustaka Utama, 2021.

[2] Roestam S. “Pembangunan Nasional untuk Kesejahteraan Rakyat”. Jakarta: Kantor Menteri Koordinator Bidang Kesejahteraan Rakyat Republik Indonesia, 1993.

[3] Badan Pusat Statistik. *Indikator Kesejahteraan Rakyat 2020*. Jakarta Pusat: Badan Pusat Statistik.2020

[4] Mattjik A, Sumertajaya IM. Sidik Peubah Ganda. Bogor: IPB Press.2011

[5] Tryon RC, Bailey DE. *Cluster* Analysis. New York (US): McGraw-Hill. 1970

[6] J. A. Hartigan, "Direct clustering of a data matrix," *J. Am. Stat. Assoc.*, vol. 67, no. 337, pp. 123–129, 1972, doi: 10.1080/01621459.1972.10481214.

[7] Mirkin B, "Mathematical Classification and *Cluster*ing" . Dordrecht (NL): Kluwer Academic Publishers, 1996.

[8] Y. Cheng and G. M. Church, "Biclustering of expression data.," *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 8, pp. 93–103, 2000.

[9] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, Spectral biclustering of microarray data: Coclustering genes and conditions, vol. 13, no. 4. 2003. doi: 10.1101/gr.648603.

[10] A. Prelić *et al.*, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006, doi: 10.1093/bioinformatics/btl060.

[11] Nurmawiya and R. Kurniawan, "Pengelompokan Wilayah Indonesia Dalam Menghadapi Revolusi Industri 4.0 Dengan Metode Biclustering," pp. 790–797, 2020

[12] Putri CA, Irfani R, Sartono B. Recognizing poverty pattern in Central Java using *Bicluster*ing Analysis. *Journal of Physics: Conference Series*. 1863(1).2021.

[13] Novidianto R, Irfani R. *Bicluster* CC Algoritm Analysis to Identify Patterns of Food Insecurity in Indonesia. *Jurnal Matematika, Statistika dan Komputasi*. 2021. 17(2):325-338

[14] B. Wang, Y. Miao, H. Zhao, J. Jin, and Y. Chen, "A biclustering-based method for market segmentation using customer pain points," Eng. Appl. Artif. Intell., vol. 47, pp. 101–109, 2016, doi: 10.1016/j.engappai.2015.06.005

[15] Tanay A, Sharan R, Shamir R. Biclustering Algorithms: A Survey. Handb Comput Mol Biol. 2024. May:709–726. doi:10.1201/9781420036275-40

[16] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, "Biclustering on expression data: A review," *J. Biomed. Inform.*, vol. 57, pp. 163–180, 2015, doi: 10.1016/j.jbi.2015.06.028

[17] N. Kavitha Sri and R. Porkodi, "An extensive survey on biclustering approaches and algorithms for gene expression data," *Int. J. Sci. Technol. Res.*, vol. 8, no. 9, pp. 2228–2236, 2019.

[18] H. Cho and I. S. Dhillon, "Coclustering of human cancer microarrays using minimum sum-squared residue coclustering," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 5, no. 3, pp. 385–400, 2008, doi: 10.1109/TCBB.2007.70268.

[19] A. Chakraborty and H. Maka, "Biclustering of gene expression data using genetic algorithm," *Proc. 2005 IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol. CIBCB '05*, vol. 2005, no. 2000, 2005, doi: 10.1109/cibcb.2005.1594893