JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 7, No. 3, July 2023, pp. 643-661

Biclustering Performance Evaluation of Cheng and Church Algorithm and Iterative Signature Algorithm

I Made Sumertajaya[1*], Wiwik Andriyani Lestari Ningsih[2], Asep Saefuddin[3], Embay Rohaeti[4]
[1,3]Department of Statistics, IPB University, Indonesia
[2]BPS-Statistics, Indonesia
[4]Departement of Mathematics, FMIPA, Pakuan University, Indonesia
imsjaya@apps.ipb.ac.id[1], wiwik.andriyani@apps.ipb.ac.id[2], asaefuddin@apps.ipb.ac.id[3], embay.rohaeti@unpak.ac.id[4]

Article History:
Received : 13-04-2023
Revised : 08-06-2023
Accepted : 14-06-2023
Online : 18-07-2023

Keywords:
Biclustering;
Cheng and Church algorithm;
Inter bicluster;
Intra bicluster;
Iterative signature algorithm.

ABSTRACT

Biclustering has been widely applied in recent years. Various algorithms have been developed to perform biclustering applied to various cases. However, only a few studies have evaluated the performance of bicluster algorithms. Therefore, this study evaluates the performance of biclustering algorithms, namely the Cheng and Church algorithm (CC algorithm) and the Iterative Signature Algorithm (ISA). Evaluation of the performance of the biclustering algorithm is carried out in the form of a comparative study of biclustering results in terms of membership, characteristics, distribution of biclustering results, and performance evaluation. The performance evaluation uses two evaluation functions: the intra-bicluster and the inter-bicluster. The results show that, from an intra-bicluster evaluation perspective, the optimal bicluster group of the CC algorithm produces bicluster quality which tends to be better than the ISA. The biclustering results between the two algorithms in inter-bicluster evaluation produce a deficient level of similarity (20-31 percent). This is indicated by the differences in the results of regional membership and the characteristics of the identifying variables. The biclustering results of the CC algorithm tend to be homogeneous and have local characteristics. Meanwhile, the results of biclustering ISA tend to be heterogeneous and have global characteristics. In addition, the results of biclustering ISA are also robust.

https://doi.org/10.31764/jtam.v7i3.14778

A. INTRODUCTION

Biclustering is an analytical tool for grouping data simultaneously from two directions and is a clustering development. Unlike the case with clustering, which only clusters data from one direction, the row or column side separately, biclustering does clustering from the row and column sides simultaneously (Castanho et al., 2022; Divina et al., 2019; Patowary et al., 2020). Clustering data from the row side using clustering will produce a group of rows (objects) that must contain all columns (variables) and vice versa (Flores et al., 2019; Kamranrad et al., 2021). However, grouping using biclustering can produce a group of rows containing only a few columns (submatrix) (Brizuela et al., 2013). Based on the developed algorithm, biclustering is like the two-way classification approach.

Biclustering, often applied in biological data analysis (Ahmed et al., 2014; Chen et al., 2022; Xie et al., 2018), has recently become famous for its application in other fields (Cotelo et al., 2020; Henriques et al., 2015; Wei et al., 2019). Putri et al. (2021) applied biclustering to identify

poverty patterns using Cheng and Church's algorithm (CC algorithm). In addition, Kaban et al. (2019) applied biclustering using the CC algorithm to identify cases of social vulnerability.

The biclustering application uses the latest CC algorithm by Ningsih et al. (2022a) to identify the economic and Covid-19 pandemic vulnerability cases. Biclustering using the CC algorithm is quite popular in its application because it has several advantages. One of its advantages is avoiding overlapping between the resulting groups (biclusters) (Di Iorio et al., 2020; Pontes et al., 2015). Another application of biclustering to identify patterns of economic and Covid-19 pandemic vulnerability was also carried out by Ningsih et al. (2022b) using a different algorithm, namely the Iterative Signature Algorithm (ISA). Biclustering using ISA also has several advantages, one of which is the use of two thresholds which are said to be the most potent tools in clearly distinguishing structures at different levels (Khalili et al., 2019; Zhang et al., 2021).

The results of biclustering in identifying economic and Covid-19 pandemic vulnerability patterns using the CC and ISA algorithms in studies by Ningsih et al. (2022a) and Ningsih et al. (2022b) yield the same general conclusions. Biclustering using the CC and ISA algorithms both yielded the result that most regions in Indonesia tend to have low economic and Covid-19 pandemic vulnerability in their respective spatial pattern characteristic variables (Ningsih et al., 2022a, 2022b). However, it gives different results when focusing on the Java Island region. Biclustering using the CC algorithm shows that most regions on Java Island tend to have low economic and Covid-19 pandemic vulnerability in the spatial pattern characteristic variables (Ningsih et al., 2022a). However, biclustering using ISA gives the opposite results. Most areas on Java Island tend to have high economic and Covid-19 pandemic vulnerability in the spatial pattern characteristic variables (Ningsih et al., 2022b).

In more detail from the studies by Ningsih et al. (2022a) and Ningsih et al. (2022b), there are particularly significant differences. This difference is interesting to study in evaluating the performance of the biclustering results between the two algorithms. Therefore, this research will evaluate the performance of biclustering algorithms, namely the Cheng and Church algorithm (CC algorithm) and the Iterative Signature Algorithm (ISA). The performance evaluation results expect to provide information related to the characteristics of the biclustering results produced by each algorithm, especially in the case of economic and Covid-19 pandemic vulnerability in Indonesia.

B. METHODS

1. Data Sources

This study uses secondary data from the BPS-Statistics; Ministry of Health; Ministry of Villages, Development of Disadvantaged Regions and Transmigration; and the Ministry of Environment and Forestry. The units of observation in this study are regions (34 provinces) in Indonesia with variable data based on 2020, as presented in Table 1. Due to the limited data obtained, the X3 variable is based on 2018, and the X16 variable is based on 2019. The indicators that make up the Economic Vulnerability Index (EVI) and the Pandemic Vulnerability Index (PVI), according to the United Nations (UN) and National Institute of Environmental Health Sciences (NIEHS), are the basis for determining the variables used in this study (United Nations, 2011; National Institute of Environmental Health Sciences, 2020), as shown in Table 1.

Table 1. Research Variables Based on Indicators on EVI and PVI

Indicator	Variable (notation)
A. EVI	
1. Size	Population percentage (X1)
2. Location	Remoteness and underdeveloped areas (X2) $^{-1}$
3. Environment	Percentage of population in coastal areas (X 3) **
4. Economic Structure	Export concentration (X4)
	Contribution of agricultural, forestry and fishery products (Category A) in Gross Regional Domestic Product/GRDP (X5)
5. Shocks to trading conditions	Instability of exports of goods and services (X6) $^{-1}$
6. Natural shock	Instability of agricultural production results (X7)
	Ratio of natural disaster victims per 1,000 population (X8)
B. PVI	
1. Infectious case	Infectious case of Covid-19 (X9)
2. Spread of disease	Covid-19 death rate (X10)
3. Mobility	Estimated percentage of daytime population (X11)
	Average traffic volume (X12)
4. Residential density	Average number of household members (X13)
5. Testing	Covid-19 testing (X14)
6. Social distancing	Social distancing score (X15)
7. Air pollution	Air quality index (X16) $^{-1*}$
8. Age distribution	Percentage of population aged 65 years and over (X17)
9. Comorbidities which include premature death, smoking, diabetes, and obesity	Morbidity rate (X18)
	Percentage of smokers in the adult population (X19)
	Percentage of population without insurance (X20)
10. Health disparities	Percentage of poor people (X21)
	Open unemployment rate (X22)
11. Hospital bed	Ratio of population to availability of hospital beds (X23)

*Data for 2019, **Data for 2018, $^{-1}$Inverse Value

The use of inverse values is intended to align the variable condition with the concept of vulnerability in other variables. A brief explanation for some of the variables in Table 1 is as follows:

a. Remoteness and underdeveloped areas use the inverse value approach of the Developing Village Index.
b. The population in the coastal area is estimated using the proportion of the number of villages located by the sea multiplied by the total population.
c. Export concentration uses the percentage of category A commodity exports to total exports.
d. Instability of goods and services exports uses the inverse value approach of the ratio of the contribution of total exports in 2020 divided by the previous year.
e. Instability of agricultural production results using the contribution ratio approach to Category A in 2020 divided by the previous year.
f. The percentage of the daytime population is estimated by multiplying the total population by the proportion of passenger cars and motorcycles.
g. The social distance score uses the ratio of domestic tourists (number of residents travelling other than for work or school) per resident.

2. Procedure of Analysis

a. Data Exploration

To explore a data matrix measuring 34 regions × 23 variables that have been standardized using the standard normal. The standardized data matrix is called the scaling data matrix. Exploration was carried out using a heatmap to obtain an overview of the data related to the initial characteristics of each region according to the constituent variables of the EVI and PVI indicators.

b. Biclustering Algoritme CC dan ISA

Biclustering is a helpful methodology for finding hidden local coherent patterns in a data matrix by classifying patterns simultaneously in both directions, the rows and columns of the data matrix (Alzahrani et al., 2017; Henriques & Madeira, 2014; Huang et al., 2020). According to Ferraro et al. (2021) biclustering consists of simultaneously partitioning a set of rows and columns into classes or biclusters. Given a matrix $A_{I\times J} = (U, V)$ with a set of rows U consisting of I rows ($|U|$) and a set of columns V consisting of J columns ($|V|$). A bicluster is a submatrix $B_{n\times m} = (U', V')$ with a row subset U' consisting of n rows sample and a column subset V' consisting of m column sample. Then a_{ij} is the value in the matrix A corresponding to the-ith row and the-jth column (Siswantining et al., 2021) Noise in a bicluster is the residual calculated from the difference between the element value of a_{ij} with the predicted value. The predicted value (($\hat{a}_{ij}$) is calculated from the corresponding row and column averages and their bicluster averages (Pang, 2022). The existence of these residues makes the element values of a_{ij} follow equation (1) and the bicluster residues are denoted by e_{ij} like equation (2). Furthermore, the average value of the-ith row in the bicluster (row average) is denoted by a_{iJ} and follows equation (3), the average value of the-jth column in the bicluster (column average) is denoted by a_{Ij} and follows equation (4), and the average value of all elements in a bicluster (bicluster average) is denoted by a_{IJ} and follows equation (5) (Ramkumar et al., 2022).

$$a_{ij} = e_{ij} + a_{iJ} + a_{Ij} - a_{IJ} \quad (1)$$

$$e_{ij} = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ} \quad (2)$$

$$a_{iJ} = \frac{1}{m}\sum_{j=1}^{m} a_{ij} \quad (3)$$

$$a_{Ij} = \frac{1}{n}\sum_{i=1}^{n} a_{ij} \quad (4)$$

$$a_{IJ} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij} \quad (5)$$

The Cheng and Church algorithm (CC algorithm) is a biclustering algorithm that looks for biclusters with constant values, rows or columns (Ardaneswari et al., 2017). This algorithm searches for biclusters simultaneously by considering row and column coherence for a submatrix that is the residual score average (Oghabian et al., 2014). Pontes et al. (2015) stated that Cheng and Church were the first to apply biclustering to gene expression data by adopting a sequential covering strategy to return a list of n biclusters from an expression data matrix. Bicluster quality was measured by the mean squared residue (MSR) size. The measurement aims to evaluate the coherence of genes (rows) and conditions (columns) of the bicluster using the gene expression values (objects) and conditions (variables) in it. Given a data matrix A and a threshold $\delta > 0$,

the goal of the CC algorithm is to find δ-bicluster, i.e., row subsets and column subsets with a score not greater than δ (Di Iorio et al., 2020). The score is the coherence score in the form of a residual score average and the algorithm makes the smallest MSR as the goal. The MSR of a matrix is denoted by $H_{(I,J)}$ and defined by equation (6) with e_{ij} defined as equation (2)(Ramkumar et al., 2022).

$$H_{(I,J)} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(e_{ij}\right)^2 \tag{6}$$

In addition, the row-squared residuals average of a matrix ($d(i)$) and the column-squared residuals average of a matrix ($d(j)$) are defined by equations (7) and (8), respectively.

$$d(i) = \frac{1}{m}\sum_{j=1}^{m}\left(e_{ij}\right)^2 \tag{7}$$

$$d(j) = \frac{1}{n}\sum_{i=1}^{n}\left(e_{ij}\right)^2 \tag{8}$$

The following is the Cheng and Church Algorithm Chart (Modified from Pontes et al., 2015), as shown in Figure 1.

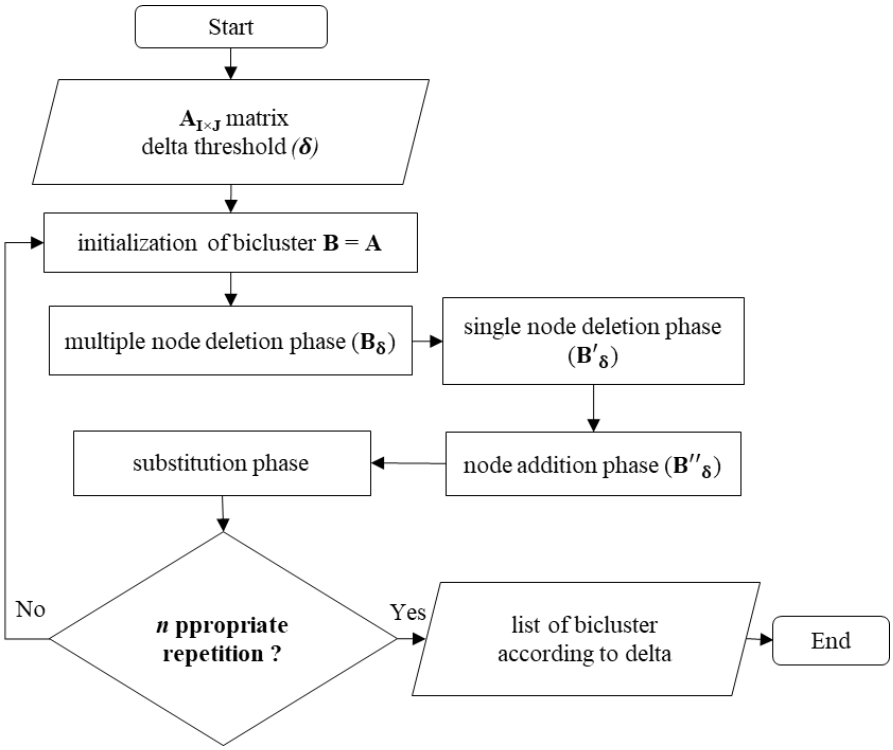


Figure 1. Cheng and Church Algorithm Flowchart (Modified from Pontes et al. (2015)

According to Pontes et al. (2015), the CC algorithm generally works by taking input as a matrix denoted by $\boldsymbol{A}$ and a threshold δ used to reject non-δ biclusters. Consequently, a list of δ-biclusters is returned as output. The following are the stages of the CC algorithm biclustering CC and are illustrated through a flowchart in Figure 1 (Pontes et al., 2015).

1) Bicluster initialization sets the initial matrix of the input ($\boldsymbol{A}$) data and the delta threshold (δ).

2) The multiple node deletion phase, i.e., deleting rows and columns based on the residuals average of row-squared ($d(i)$) and column-squared ($d(j)$) that greater than 1,5 × the squared residual average of the entire matrix ($\alpha\ H_{(I,J)}$), as long as it satisfies the mean square residue (MSR) condition > δ.
3) The single node deletion phase, i.e., deleting rows or columns based on conditions $d(i)$ or $d(j)$, is the maximum, as long as it satisfies the MSR condition > δ.
4) The node addition phase is the addition of rows and columns based on conditions of $d(i) \leq H_{(I,J)}$ and $d(j) \leq H_{(I,J)}$ as long as it satisfies the MSR conditions of adding nodes results $\leq H_{(I,J)}$.
5) The substitution phase, i.e., replacing bicluster resulting matrix elements with a random number to prevent overlapping between biclusters.
6) Repeat steps 1 to 5 as many as *n* times, that is, as many as *n* biclusters want to find.

Iterative Signature Algorithm (ISA) is a biclustering algorithm with input in the form of a matrix while the resulting output is in the form of a bicluster set and is defined as transcription modules (TM) (Balamurugan et al., 2015). A TM contains a subset of rows and columns that depend on a pair of thresholds, and it is the row and column thresholds that determine the degree of similarity of the TM (Pontes et al., 2015). Pontes et al. (2015) classified ISA into non-metric-based linear algebra groups. The algorithm does not use a specific evaluation size in the process of bicluster search. However, it uses vector spaces and linear mapping between these spaces to describe and find the most correlated submatrix (TM) (Pontes et al., 2015). It is known that a matrix $A_{|U|\times|V|} = (U, V)$, is a matrix with the number of rows denoted by $|U|$ and the number of columns denoted by $|V|$. The row score on the ISA is the row average of the column sample ($a^C_{uV'}$) that meets the conditions in equation (9), while the column score is the column average of the row sample ($a^G_{U'v}$) that meets the conditions in equation (10)(Ningsih et al., 2022b).

$$a^C_{uV'} > t_C\sigma_C \tag{9}$$
$$\frac{1}{|V'|}\sum_{v=1}^{|V'|} a^C_{uv} > \frac{t_C}{\sqrt{|V'|}}$$
$$a^G_{U'v} > t_G\sigma_G \tag{10}$$
$$\frac{1}{|U'|}\sum_{u=1}^{|U'|} a^G_{uv} > \frac{t_G}{\sqrt{|U'|}}$$

with a^C_{uv} is a matrix element of a normalized column matrix (A^C), and a^G_{uv} is a matrix element of a normalized row matrix (A^G). The following are the biclustering stages of ISA and the illustration is in Figure 2 (Ningsih et al., 2022b).

1) Sets the row and column thresholds (t_C, t_G), the seed value, and the number of seeds (n).
2) Creates a normalized row matrix (A^G) and a normalized column matrix (A^C) from matrix $A_{|U|\times|V|}$ $[UV]$.
3) Randomly select multiple column vectors (column samples) $[UV']$.
4) Computes each row average of a column sample ($a^C_{uV'}$) using A^C.

5) Selects a sample of rows that meet the conditions $a^{C}_{uV'} > t_C\sigma_C$ and its average value becomes the "row score" $[U'V']$.
6) Computes each column average of a row sample $(a^{G}_{U'_v})$ using A^G.
7) Selects a sample of columns that satisfies the condition $a^{G}_{U'_v} > t_G\sigma_G$ and its average value becomes the "column score" $[U'V'']$.
8) Repeat stages 4 to 7 as many as the number of seeds (n) when convergent conditions are unmet.
9) When the convergent condition is met, i.e., $\frac{|V'\setminus V''|}{|V'\cup V''|} < \varepsilon$, rows and columns (bicluster) are selected.
10) Repeat stages 3 to 9 as many as the number of biclusters that may be formed, as shown in Figure 2.

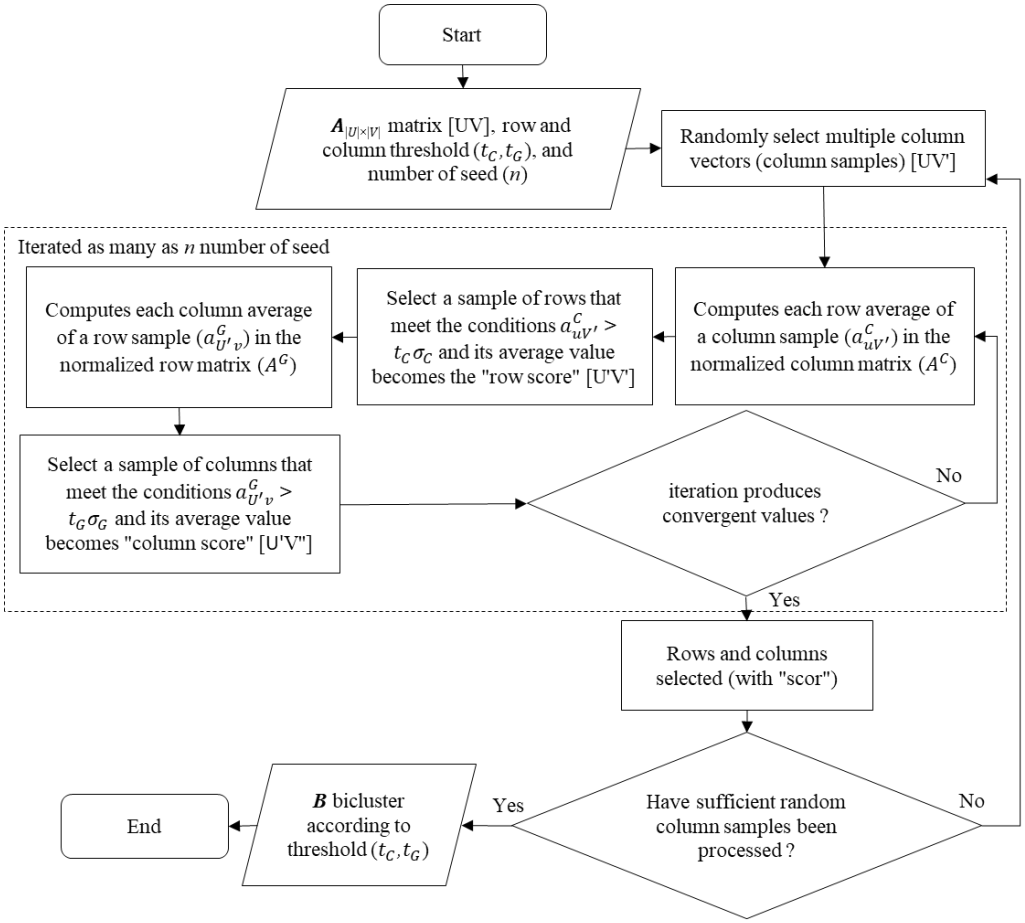


Figure 2. Iterative Signature Algorithm Flowchart (Modified from Ningsih et al., 2022b)

c. Performance Evaluation of Biclustering Algorithm

According to Kavitha Sri & Porkodi (2019), the biclustering algorithm's performance evaluation uses two categories of evaluation functions: the intra-bicluster evaluation function and the inter-bicluster evaluation function. The intra-bicluster evaluation function is a function that measures the quality of a bicluster using the level of coherence in a bicluster (Ben Saber & Elloumi, 2014). The size of the intra-bicluster evaluation

function used in this study is the mean squared residue (MSR) and is defined by equation (11) (Kavitha Sri & Porkodi, 2019).

$$MSR_{(I,J)} = \frac{\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}(a_{ij}-a_{iJ}-a_{Ij}+a_{IJ})^2}{|I|\times|J|} = \frac{\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}(e_{ij})^2}{|I|\times|J|} \quad (11)$$

with a_{ij} is a bicluster element in the-ith row and the-jth column, a_{IJ} is the average across all biclusters, a_{Ij} is the average in the-jth column, a_{iJ} is the average in the-ith row, $|I| \times |J|$ is the bicluster dimension (volume), i.e., the bicluster row size ($|I|$) multiplied by the bicluster column size ($|J|$). The value of $MSR_{(I,J)}$ represents the variation (diversity) associated with the bicluster interaction between rows and columns (Kavitha Sri & Porkodi, 2019; Saber & Elloumi, 2015). According to (Putri et al., 2021) , the bicluster quality will be better as the residual value decreases and/or the volume of the bicluster increases. The quality of the bicluster group based on MSR can then be measured by calculating the average of MSR divided by the volume (the MSR average per volume) and defined by equation (12) (Putri et al., 2021),

$$MSR\ average\ per\ volume = \frac{1}{b}\sum_{i=1}^{b}\frac{MSR_i}{Volume_i} \quad (12)$$

where b is the number of biclusters generated by a particular algorithm. Apart from using the MSR value, the Akaike information criterion (AIC) value of a bicluster can also be calculated using the formula in equation (13) (Brewer et al., 2016),

$$AIC_{(B)} = v\ ln\left(MSR_{(I,J)}\right) + 2k \quad (13)$$

where k is the number of parameters adjusted independently to obtain a bicluster ($B_{|I|\times|J|}$), i.e., $k = |I| + |J| + 1$ and v is the volume or dimension of a bicluster, i.e., $v = |I| \times |J|$, and e_{ij} is the residue of a bicluster following the formula in equation (2). The AIC value measures the goodness of the biclustering results.

Meanwhile, the inter-bicluster evaluation function is a function that measures the quality of the bicluster group by assessing the accuracy of an algorithm to obtain actual biclusters in a data matrix (Ben Saber & Elloumi, 2014; Henriques & Madeira, 2018). The size of the inter-bicluster evaluation function used is the Liu and Wang index which is defined by equation (14)(Saber & Elloumi, 2015),

$$I_{Liu\&Wang}(M_{opt}, M) = \frac{1}{K_{opt}}\sum_{i=1}^{K_{opt}} max\left(\frac{|G_i\cap G_j|+|C_i\cap C_j|}{|G_i\cup G_j|+|C_i\cup C_j|}\right) \quad (14)$$

with M_{opt} is the bicluster group that has the smallest average value of MSR per volume, and M is the other bicluster group. K_{opt} is the number of biclusters in M_{opt}, $|G_i \cap G_j|$ is the number of rows (G) in M_{opt} which intersects with rows in M, and $|C_i \cap C_j|$ is the number of columns (C) in M_{opt} which intersects columns in M. $|G_i \cup G_j|$ is the number of combined rows of M_{opt} and M, and z$|C_i \cup C_j|$ is the number of combined columns of M_{opt}

and M. Liu and Wang's index compares two solutions (biclustering results) by considering the rows and columns of a bicluster (Kavitha Sri & Porkodi, 2019). The Liu and Wang index values indicate how well an optimal bicluster group (M_{opt}) will have similarities with other bicluster group (M). When $M_{opt} = M$, the Liu and Wang index values are 1.00 (Saber & Elloumi, 2015).

The flowchart of the biclustering algorithm's performance evaluation process in this study is illustrated in Figure 3. In general, through Figure 3, each algorithm's first evaluation is carried out separately to obtain biclustering results at the optimal threshold. The biclustering results at the optimal threshold are then evaluated for their performance separately using the inter-bicluster evaluation function. In addition, the evaluation also uses the intra-bicluster and inter-bicluster evaluation functions simultaneously. This evaluation is a comparability of the biclustering results. The results are compared in terms of membership, characteristics, and distribution of the biclustering results, as shown in Figure 3.

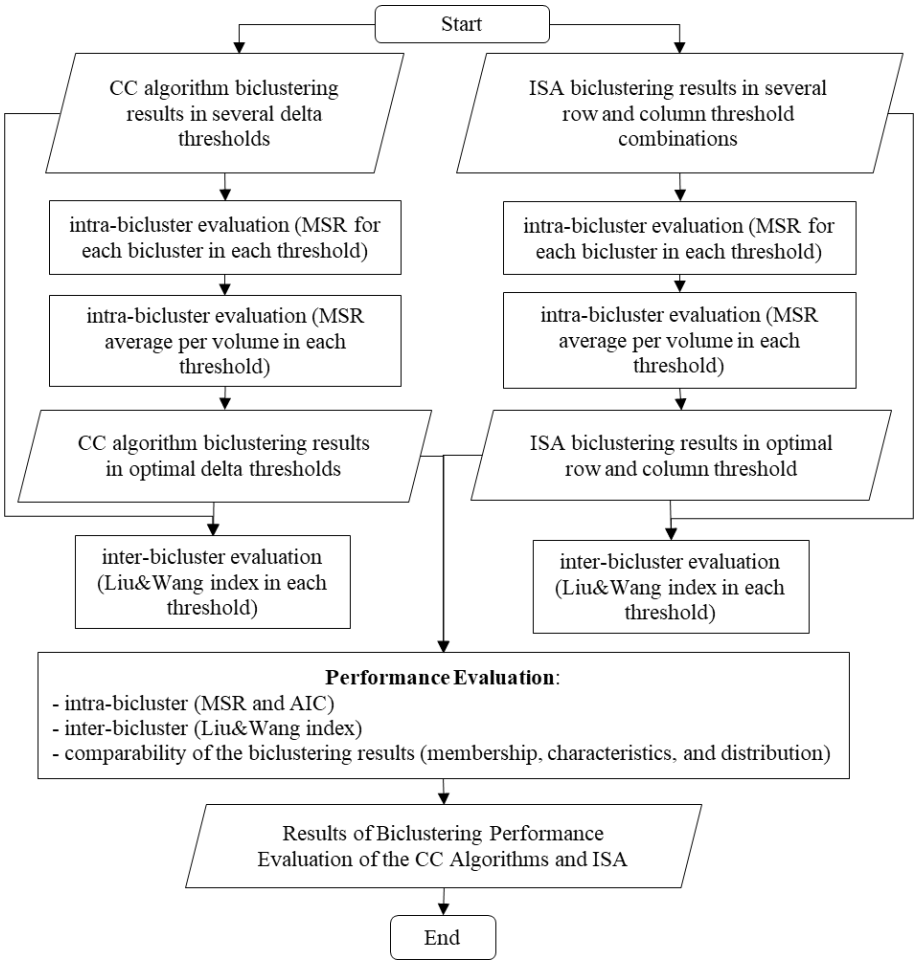


Figure 3. Flowchart of the Biclustering Algorithm's Performance Evaluation Process

C. RESULT AND DISCUSSION

1. Data Exploration

The description of the data related to the initial characteristics of each region according to the constituent variables of the EVI and PVI indicators is illustrated through the scaling data matrix heatmap in Figure 4. The heatmap describes several extreme values in the EVI (X1 to X8) and PVI (X9 to X23) indicator variables in a particular province. Some values are highly positive (dark orange), and some are highly negative (light yellow) (Guo et al., 2020).

The indication of provinces with values that tend to be highly positive is that the province tends to be vulnerable. Conversely, the province tends to be invulnerable. The example is on the PVI X9 indicator variable: the Covid-19 infectious case. DKI Jakarta Province has a highly positive value (dark orange). It indicates that DKI Jakarta Province tends to be highly vulnerable to the Covid-19 pandemic, especially regarding the indicator of Covid-19 infectious cases, as shown in Figure 4.

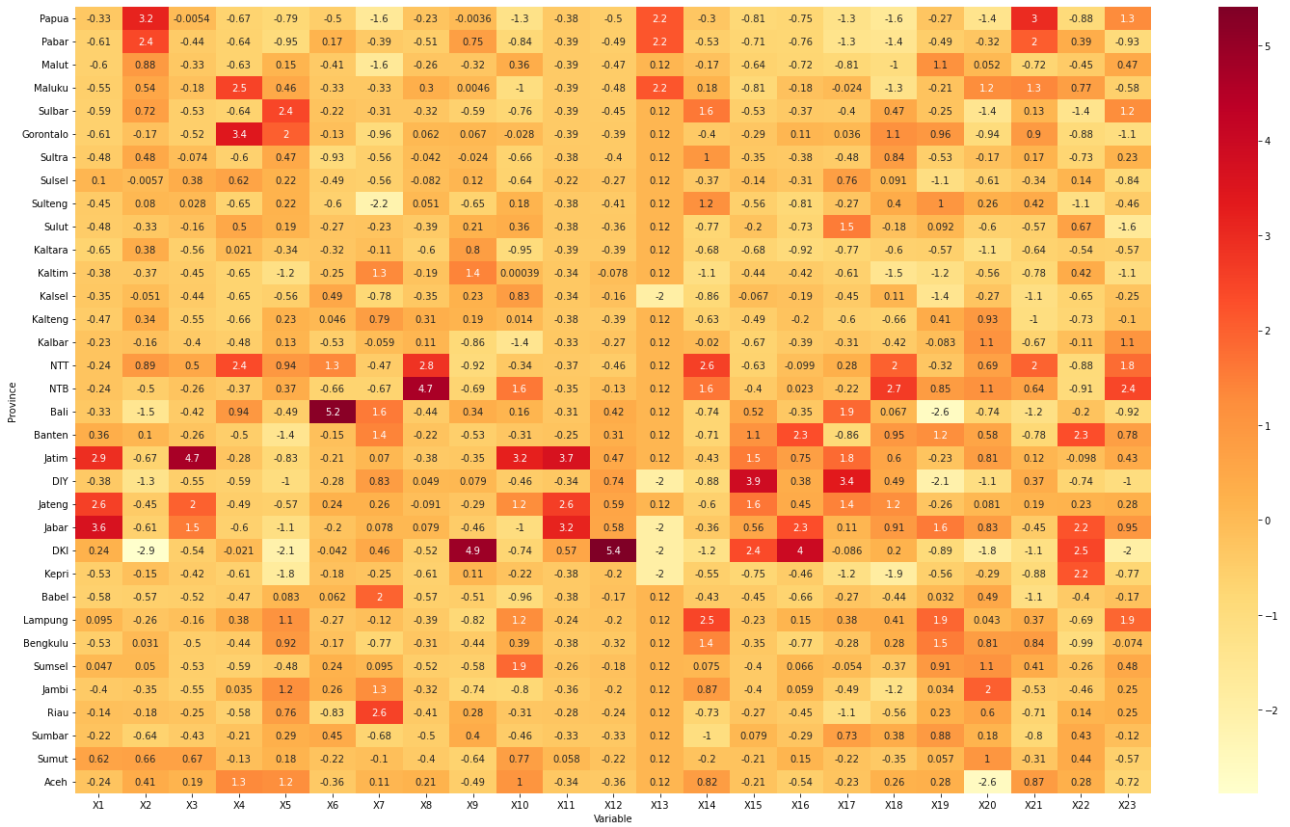


Figure 4. Heatmap of Scaling Data Matrices

Another example is the EVI X2 indicator variable: remoteness and underdeveloped areas. DKI Jakarta Province has a highly negative value (light yellow). It indicates that DKI Jakarta Province tends to have a low economic vulnerability, especially in remoteness and underdeveloped areas.

2. Biclustering Algorithms CC and ISA

Ningsih et al. (2022a) research on "biclustering applications in Indonesian economic and pandemic vulnerability" shows that using the CC algorithm produces optimal bicluster groups at the 0.01 delta threshold. Figure 5 shows six optimal bicluster groups, as shown in Figure 5.

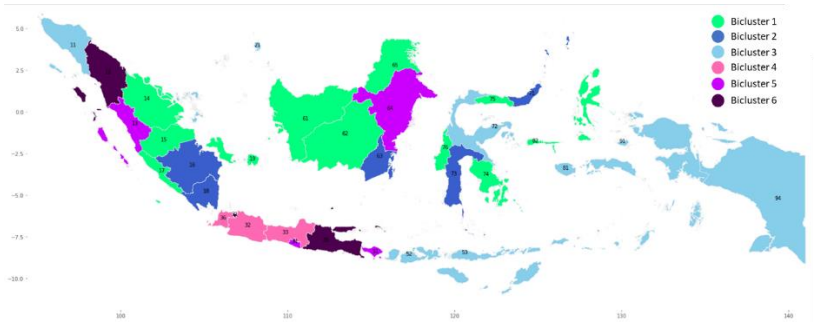


Figure 5. Map of Biclustering Results of Optimal Threshold CC Algorithm according to the Type of Spatial Pattern

The optimal bicluster group of the CC algorithm concludes that areas that dominate in Indonesia are the first type of spatial pattern with the most invulnerable characteristics. It indicates that most regions in Indonesia tend to have low economic and Covid-19 pandemic vulnerability in the first spatial pattern characteristic variable (Bicluster 1). Meanwhile, Ningsih et al. (2022b) research regarding "pattern detection of economic and pandemic vulnerability index in Indonesia using bi-cluster analysis" shows that biclustering using ISA produces optimal bicluster groups at the -1.0 row and -1.0 column threshold. The number of biclusters formed from the optimal bicluster group is three. However, due to the overlap between the three biclusters, five different spatial patterns are formed, as shown in Figure 6. The ISA optimal bicluster group concludes that areas that dominate in Indonesia are the fifth type of spatial pattern with invulnerable characteristics. It indicates that most regions in Indonesia tend to have low economic and Covid-19 pandemic vulnerability on the fifth spatial pattern characteristic variable (Overlap Bicluster 1, 2, and 3), as shown in Figure 6.

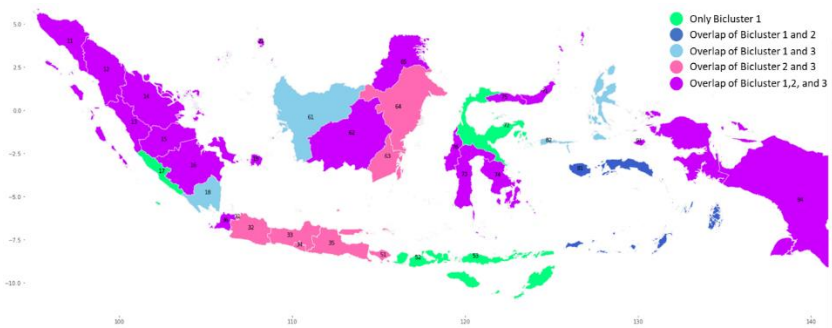


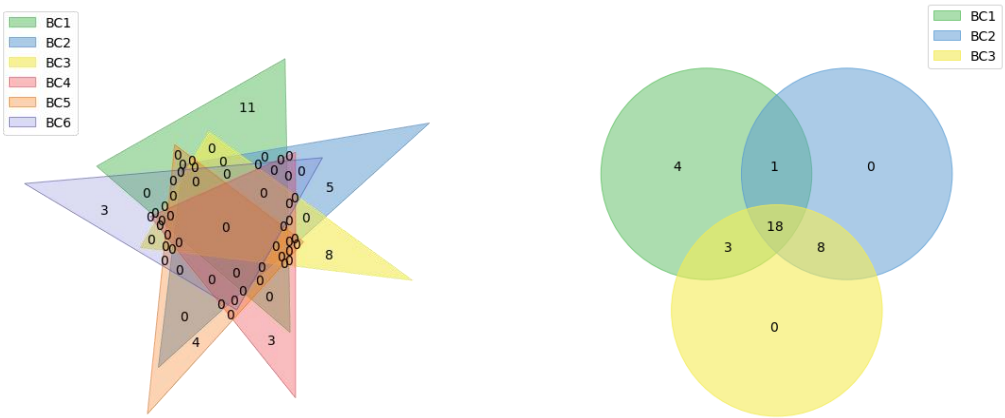
Figure 6. Map of Biclustering Results of Optimal Threshold ISA according to the Type of Spatial Pattern

3. Biclustering Algorithm Performance Evaluation

This study evaluated the biclustering application using the CC algorithms and ISA from research results by Ningsih et al. (2022a) and Ningsih et al. (2022b). The conducted performance evaluation is a comparability study of biclustering results for each optimal

threshold. The optimal threshold in the CC algorithm is at 0.01 delta (Ningsih et al., 2022a) and ISA at the -1.0 row and -1.0 column threshold (Ningsih et al., 2022b). The results of comparability study carried out in this study included the comparability of the objects' membership (region) and the formed variables, the characteristics of the mean and median of the same identifying variables, the values distribution of the region identifying variables, as well as the results of intra-bicluster and inter-bicluster evaluation.

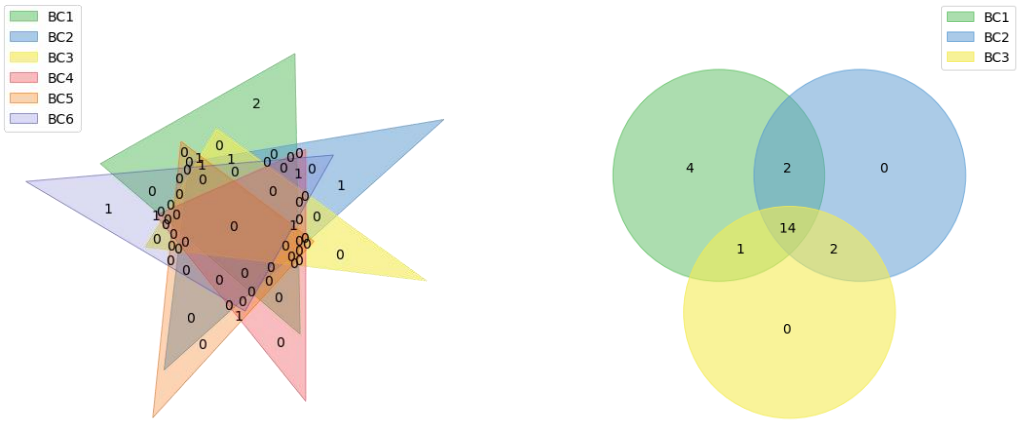
Figure 7 compares regional membership between the CC and ISA algorithms. This figure shows that both algorithms can group all objects (regions) into each formed bicluster (BC). However, the ISA produces overlapping regional memberships, forming five types of spatial patterns or bicluster groups. Obtaining the five spatial pattern types comes from the overlap of the three BC combinations resulting from biclustering ISA, as shown in Figure 7.



(a) (b)

Figure 7. Area Membership Results of the CC Algorithms (A) and ISA (B)

Figure 8 compares the results of the variable's membership between the CC and ISA algorithms. The figure shows that the membership of the CC algorithm variables tends to be less, i.e., only eleven identifying variables, in contrast to the membership of the ISA variables, which are as many as 23. It shows that the identifying variables that describe the characteristics of the biclustering results of the CC algorithm tend to be few. Therefore, the biclustering results of the CC algorithm are local characters. However, the results of biclustering ISA tend to be global characters. It is because ISA can transform all research variables into identifier variables for each formed bicluster, as shown in Figure 8.



(a) (b)

Figure 8. Variable Membership Results of the CC Algorithms (A) and ISA (B)

Table 2 presents a comparison of the characteristics of the mean and median values of the same identifying variables between the two results of the biclustering algorithm. Table 2 shows six identifying variables that are the same between the biclustering results of the CC algorithms and ISA. The six variables consist of three EVI indicator variables (X1, X3, and X8) and three PVI indicator variables (X11, X12, and X15). These variables tend to have the characteristics of an average value classified as an invulnerable value. The mean and median values of these identifying variables in total are classified as having an invulnerable value characteristic. However, there are differences between the mean and median values of each identifying variable in each bicluster, as shown in Table 2.

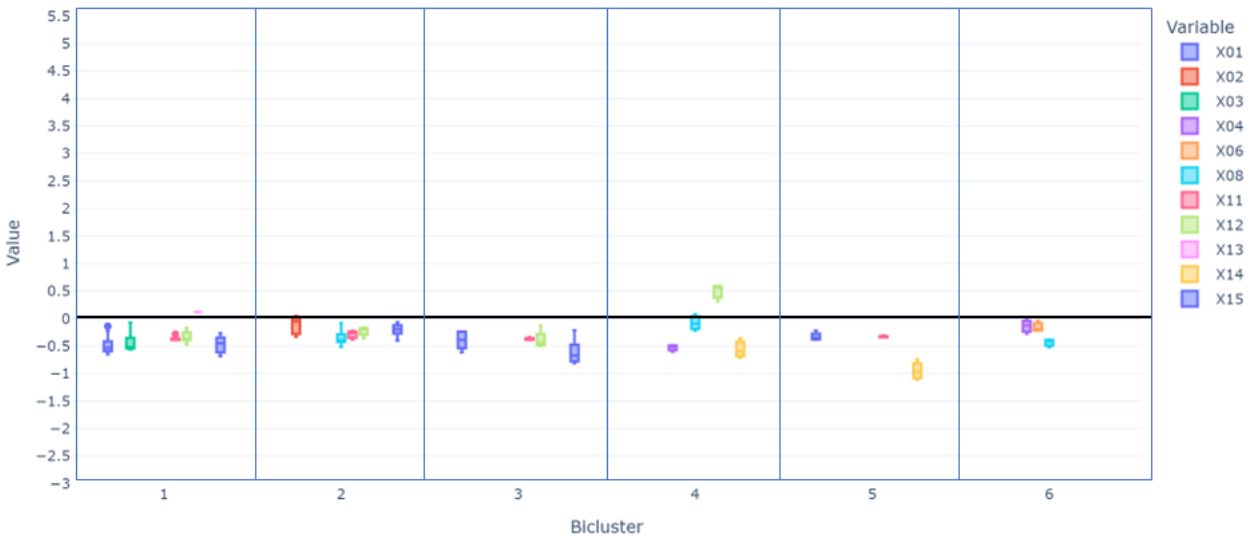
Table 2. Characteristics of the Mean and Median Values of the Same Identifying Variables between Biclustering Results of CC Algorithm and ISA according to Bicluster

Identifying variable	CC Algorithm			ISA		
	Bicluster	Mean	Median	Bicluster	Mean	Median
X1	1	-0.480^{T}	-0.534^{T}	1	-0.305^{T}	-0.421^{T}
	3	-0.397^{T}	-0.387^{T}	2	0.081^{R}	-0.348^{T}
	5	-0.326^{T}	-0.351^{T}	3	0.069^{R}	-0.328^{T}
	total	-1.203^{T}	-1.272^{T}	total	-0.155^{T}	-1.096^{T}
X3	1	-0.435^{T}	-0.524^{T}	1	-0.225^{T}	-0.295^{T}
				2	0.041^{R}	-0.425^{T}
				3	0.014^{R}	-0.420^{T}
	total	-0.435^{T}	-0.524^{T}	total	-0.170^{T}	-1.140^{T}
X8	2	-0.347^{T}	-0.393^{T}	1	0.071^{R}	-0.289^{T}
	4	-0.078^{T}	-0.091^{T}	2	-0.248^{T}	-0.322^{T}
	6	-0.435^{T}	-0.404^{T}	3	-0.261^{T}	-0.322^{T}
	total	-0.859^{T}	-0.888^{T}	total	-0.438^{T}	-0.932^{T}
X11	1	-0.368^{T}	-0.382^{T}	1	-0.333^{T}	-0.376^{T}
	2	-0.289^{T}	-0.258^{T}	2	0.090^{R}	-0.343^{T}
	3	-0.372^{T}	-0.379^{T}	3	0.064^{R}	-0.341^{T}
	5	-0.330^{T}	-0.337^{T}			
	total	-1.360^{T}	-1.355^{T}	total	-0.178^{T}	-1.059^{T}
X12	1	-0.335^{T}	-0.390^{T}	1	-0.306^{T}	-0.346^{T}
	2	-0.235^{T}	-0.200^{T}	2	0.084^{R}	-0.219^{T}

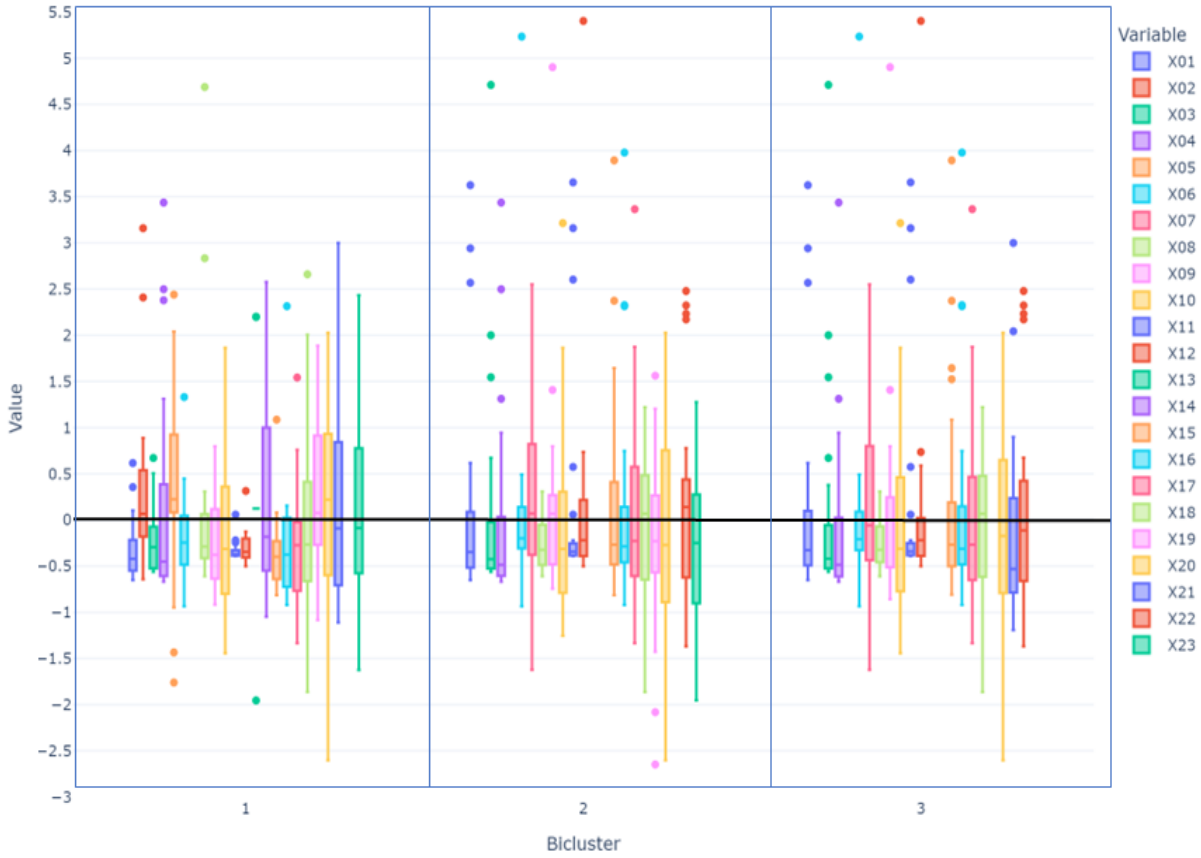
Identifying variable	CC Algorithm			ISA		
	Bicluster	Mean	Median	Bicluster	Mean	Median
	3	-0.378[T]	-0.436[T]	3	0.062[R]	-0.219[T]
	4	0.494[R]	0.580[T]			
	total	-0.454[T]	-0.445[T]	total	-0.160[T]	-0.785[T]
X15	1	-0.464[T]	-0.450[T]	1	-0.385[T]	-0.399[T]
	2	-0.207[T]	-0.196[T]	2	0.128[R]	-0.268[T]
	3	-0.610[T]	-0.672[T]	3	-0.095[T]	-0.268[T]
	total	-1.282[T]	-1.318[T]	total	-0.162[T]	-0.934[T]

[T]: having an invulnerable value. [R]: having a vulnerable value

It shows from Table 2 that the median value of each characterizing variable in each bicluster of the CC and ISA algorithms has the same invulnerable value characteristics but differs from the average value. Only one variable is classified as a vulnerable characteristic of the CC algorithm, i.e., variable X12 in Bicluster 4. Meanwhile, the average value of the other variable has an invulnerable value characteristic. Meanwhile, ten identifying variable values of ISA are classified as having vulnerable value characteristics. These variables are X1 (Bicluster 2 and 3), X3 (Bicluster 2 and 3), X8 (Bicluster 1), X11 (Bicluster 2 and 3), X12 (Bicluster 2 and 3), and X15 (Bicluster 2). It indicates that the results of biclustering ISA have outlier values (outliers) due to a significant difference between the characteristics of the mean and median values. The outlier values are more clearly seen through the values distribution of the regional identifying variables according to the CC algorithm and ISA biclustering results, which are depicted in Figure 9.



(a)



(b)

Figure 9. The Distribution of the Regional Identifying Variables according to the Biclustering Results of the CC Algorithm (A) and ISA (B)

Figure 9(a) shows that almost all regions in each identifying variable of the CC algorithm biclustering results have values below zero (low values). 89.74% of the 156 observation points of the CC algorithm results produce low values (invulnerable characteristics). It indicates that almost all regions of Indonesia tend to be a low vulnerability in the identifying variable of the CC algorithm result. Meanwhile, it can be seen from Figure 9(b) that there are several areas with outlier values, most of which are above zero in each identifying variable of the ISA biclustering results. However, when examined more closely, it was found that most areas in each identifying variable of the ISA biclustering results were at low values, i.e., around 63.28% of the 1,525 observation points. It indicates that most regions in Indonesia also tend to have a low vulnerability in the identifying variable of the ISA result.

Figure 9 describes that the CC algorithm and ISA results are equally dominant in areas with low identifying variables (low vulnerability) values. However, the values of the biclustering results of the CC algorithm tend to be homogeneous, as indicated by the relatively small mean values of the identifier variables, which range from 0.011 to 0.023. In addition, the results of the CC algorithm tend to be sensitive to outliers because there are almost no outlier values in the biclustering results of the CC algorithm. Meanwhile, the results of biclustering ISA tend to be scattered or heterogeneous. It is indicated by the relatively high mean value of the identifying variables' variance compared to the CC algorithm, which ranges from 0.634 to 1.045.

The ISA results are not sensitive to outliers because of many outlier values in its biclustering result, so the ISA tends to be robust.

Based on the intra-bicluster evaluation measure, the CC algorithm produces a smaller value of MSR average per volume (0.00041) compared to ISA (0.00141). It indicates that the bicluster quality of the CC algorithm results tends to be better when compared to the ISA results. The MSR average per volume value aligns with the MSR and AIC values for each bicluster result of the CC algorithm, and ISA presented in . From this table, the MSR and AIC values for each CC algorithm bicluster tend to be smaller when compared to the ISA bicluster.

Table 3. Size of the Intra-Bicluster Evaluation of the Biclustering Results of the CC Algorithm and ISA according to Bicluster

Bicluster	CC Algorithm		ISA	
	MSR	AIC	MSR	AIC
1	0.0086	-277.8925	0.5279	-252.8345
2	0.0087	-96.5408	0.8187	-5.1938
3	0.0096	-122.5568	0.7774	-30.1297
4	0.0068	-43.8744	-	-
5	0.0057	-46.0489	-	-
6	0.0056	-32.6406	-	-

Based on the inter-bicluster evaluation measure using the Liu and Wang index values, the biclustering results between the CC and ISA algorithms show a deficient similarity level, around 20 to 31 percent. When the biclustering results of the CC algorithm are assumed to be the optimal bicluster group, the Liu and Wang index values are 0.20. Meanwhile, when the ISA biclustering results are assumed to be the optimal bicluster group, the Liu and Wang index values are 0.31. It indicates that the biclustering results between the CC and ISA algorithms differ, resulting in biclusters with different memberships and characteristics. The 69 to 80 percent difference is supported by an explanation regarding the distribution of regional identifying variable values in Figure 9 and the results comparison of the area and variable membership in Figure 7 and Figure 8.

D. CONCLUSION AND SUGGESTIONS

The evaluation result of the CC algorithm and ISA performance in the form of its optimal threshold biclustering comparative study shows that the bicluster quality of the CC algorithm tended to be better. The indication is that the MSR average per volume of the CC algorithm is lower than the ISA. In addition, the two biclustering results show a deficient level of similarity (20-31 percent) supported by the differences in their membership and characteristics. The biclustering results of the CC algorithm tend to be homogeneous with a small number of identifying variables (local characters) and dominated by areas with low values (low vulnerability). Meanwhile, the results of biclustering ISA tend to be heterogeneous, with the number of identifying variables covering all research variables (global character) and dominated by areas with low vulnerability. Besides, the ISA result tends to be robust (not sensitive to outliers) because its biclustering results have many outlier values.

REFERENCES

Ahmed, H. A., Mahanta, P., Bhattacharyya, D. K., & Kalita, J. K. (2014). Shifting-and-Scaling Correlation Based Biclustering Algorithm. *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, *11*(6), 1–14. https://doi.org/10.1109/TCBB.2014.2323054

Alzahrani, M., Kuwahara, H., Wang, W., & Gao, X. (2017). Gracob: A novel graph-based constant-column biclustering method for mining growth phenotype data. *Bioinformatics*, *33*(16), 2523–2531. https://doi.org/10.1093/bioinformatics/btx199

Ardaneswari, G., Bustamam, A., & Siswantining, T. (2017). Implementation of parallel k-means algorithm for two-phase method biclustering in Carcinoma tumor gene expression data. *AIP Conference Proceedings*, *1825*, 020004. https://doi.org/10.1063/1.4978973

Balamurugan, R., Natarajan, A. M., & Premalatha, K. (2015). Stellar-mass black hole optimization for biclustering microarray gene expression data. *Applied Artificial Intelligence*, *29*(4), 353–381. https://doi.org/10.1080/08839514.2015.1016391

Ben Saber, H., & Elloumi, M. (2014). A Comparative Study of Clustering and Biclustering of Microarray Data. *International Journal of Computer Science and Information Technology*, *6*(6), 93–111. https://doi.org/10.5121/ijcsit.2014.6607

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, *7*(6), 679–692. https://doi.org/10.1111/2041-210X.12541

Brizuela, C. A., Luna-Taylor, J. E., Martinez-Perez, I., Guillen, H. A., Rodriguez, D. O., & Beltran-Verdugo, A. (2013). Improving an evolutionary multi-objective algorithm for the biclustering of gene expression data. *2013 IEEE Congress on Evolutionary Computation, CEC 2013*, 221–228. https://doi.org/10.1109/CEC.2013.6557574

Castanho, E. N., Aidos, H., & Madeira, S. C. (2022). Biclustering fMRI time series: a comparative study. *BMC Bioinformatics*, *23*(1), 1–30. https://doi.org/10.1186/s12859-022-04733-8

Chen, S., Zhang, L., Lu, L., Meng, J., & Liu, H. (2022). FBCwPlaid: A Functional Biclustering Analysis of Epi-Transcriptome Profiling Data Via a Weighted Plaid Model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(3), 1640–1650. https://doi.org/10.1109/TCBB.2021.3049366

Cotelo, J. M., Ortega, F. J., Troyano, J. A., Enríquez, F., & Cruz, F. L. (2020). Known by who we follow: A biclustering application to community detection. *IEEE Access*, *8*, 192218–192228. https://doi.org/10.1109/ACCESS.2020.3032015

Di Iorio, J., Chiaromonte, F., Cremona, M. A., & Cremona, M. A. (2020). On the bias of H-scores for comparing biclusters, and how to correct it. *Bioinformatics*, *36*(9), 2955–2957. https://doi.org/10.1093/bioinformatics/btaa060

Divina, F., Vela, F. A. G., & Torres, M. G. (2019). Biclustering of smart building electric energy consumption data. *Applied Sciences (Switzerland)*, *9*(2), 222. https://doi.org/10.3390/app9020222

Ferraro, M. B., Giordani, P., & Vichi, M. (2021). A class of two-mode clustering algorithms in a fuzzy setting. *Econometrics and Statistics*, *18*, 63–78. https://doi.org/10.1016/j.ecosta.2020.03.006

Flores, A., Tito, H., & Silva, C. (2019). Local Average of Nearest Neighbors: Univariate Time Series Imputation. *International Journal of Advanced Computer Science and Applications*, *10*(8), 45–50. https://doi.org/10.14569/ijacsa.2019.0100807

Guo, H., Zhang, W., Ni, C., Cai, Z., Chen, S., & Huang, X. (2020). Heat map visualization for electrocardiogram data analysis. *BMC Cardiovascular Disorders*, *20*(1), 1–8. https://doi.org/10.1186/s12872-020-01560-8

Henriques, R., Antunes, C., & Madeira, S. C. (2015). A structured view on pattern mining-based biclustering. *Pattern Recognition*, *48*(12), 3941–3958. https://doi.org/10.1016/j.patcog.2015.06.018

Henriques, R., & Madeira, S. C. (2014). BicSPAM: Flexible biclustering using sequential patterns. *BMC Bioinformatics*, *15*(1), 1–20. https://doi.org/10.1186/1471-2105-15-130

Henriques, R., & Madeira, S. C. (2018). BSig: evaluating the statistical significance of biclustering solutions. *Data Mining and Knowledge Discovery*, *32*(1), 124–161. https://doi.org/10.1007/s10618-017-0521-2

Huang, Q., Chen, Y., Liu, L., Tao, D., & Li, X. (2020). On Combining Biclustering Mining and AdaBoost for Breast Tumor Classification. *IEEE Transactions on Knowledge and Data Engineering*, *32*(4), 728–738. https://doi.org/10.1109/TKDE.2019.2891622

Kaban, P. A., Kurniawan, R., Caraka, R. E., Pardamean, B., Yuniarto, B., & Sukim. (2019). Biclustering method to capture the spatial pattern and to identify the causes of social vulnerability in Indonesia: A new recommendation for disaster mitigation policy. *Procedia Computer Science*, *157*, 31–37. https://doi.org/10.1016/j.procs.2019.08.138

Kamranrad, R., Soltanzadeh, S., & Mardan, E. (2021). A Combined Data Mining Based-Bi Clustering and Order Preserved Sub-Matrices Algorithm for Set Covering Problem. *Journal of Quality Engineering and Production Optimization*, *6*(2), 1–16. https://doi.org/10.22070/JQEPO.2021.5330.1144

Kavitha Sri, N., & Porkodi, R. (2019). An extensive survey on biclustering approaches and algorithms for gene expression data. *International Journal of Scientific and Technology Research*, *8*(9), 2228–2236. https://ipb.link/ijstr-2277-8616

Khalili, B., Tomasoni, M., Mattei, M., Mallol Parera, R., Sonmez, R., Krefl, D., Rueedi, R., & Bergmann, S. (2019). Automated Analysis of Large-Scale NMR Data Generates Metabolomic Signatures and Links Them to Candidate Metabolites. *Journal of Proteome Research*, *18*(9), 3360–3368. https://doi.org/10.1021/acs.jproteome.9b00295

Nations, U. (2011). *EVI Indicators*. ipb.link/un-evi

Ningsih, W. A. L., Sumertajaya, I. M., & Saefuddin, A. (2022a). Biclustering Application In Indonesian Economic And Pandemic Vulnerability. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *16*(4), 1453–1464. https://doi.org/10.30598/barekengvol16iss4pp1453-1464

Ningsih, W. A. L., Sumertajaya, I. M., & Saefuddin, A. (2022b). Pattern Detection of Economic and Pandemic Vulnerability Index in Indonesia Using Bi-Cluster Analysis. *JUITA : Jurnal Informatika*, *10*(2), 273. https://doi.org/10.30595/juita.v10i2.14940

Oghabian, A., Kilpinen, S., Hautaniemi, S., & Czeizler, E. (2014). Biclustering methods: Biological relevance and application in gene expression analysis. *PLoS ONE*, *9*(3), e90801. https://doi.org/10.1371/journal.pone.0090801

Pang, C. (2022). Construction and Analysis of Macroeconomic Forecasting Model Based on Biclustering Algorithm. *Journal of Mathematics*, *2022*, 7768949. https://doi.org/10.1155/2022/7768949

Patowary, P., Sarmah, R., & Bhattacharyya, D. K. (2020). Developing an effective biclustering technique using an enhanced proximity measure. In *Network Modeling Analysis in Health Informatics and Bioinformatics* (Vol. 9, Issue 1, p. 6). https://doi.org/10.1007/s13721-019-0211-7

Pontes, B., Giráldez, R., & Aguilar-Ruiz, J. S. (2015). Biclustering on expression data: A review. *Journal of Biomedical Informatics*, *57*, 163–180. https://doi.org/10.1016/j.jbi.2015.06.028

Putri, C. A., Irfani, R., & Sartono, B. (2021). Recognizing poverty pattern in Central Java using Biclustering Analysis. *Journal of Physics: Conference Series*, *1863*(1), 012068. https://doi.org/10.1088/1742-6596/1863/1/012068

Ramkumar, M., Basker, N., Pradeep, D., Prajapati, R., Yuvaraj, N., Arshath Raja, R., Suresh, C., Vignesh, R., Barakkath Nisha, U., Srihari, K., & Alene, A. (2022). Healthcare Biclustering-Based Prediction on Gene Expression Dataset. *BioMed Research International*, *2022*(Special Issue), 1–7. https://doi.org/10.1155/2022/2263194

Saber, H. Ben, & Elloumi, M. (2015). A New Survey on Biclustering of MicroArray Data. *International Journal for Computational Biology*, *4*(1), 21–37. https://doi.org/10.5121/csit.2014.41314

Sciences, N. I. of E. H. (2020). *Details for PVI Maps*. ipb.link/niehs

Siswantining, T., Bustamam, A., Puspa, S. D., Rustam, Z., & Zubedi, F. (2021). Biclustering of diabetic nephropathy and diabetic retinopathy microarray data using a similarity-based biclustering algorithm. *International Journal of Bioinformatics Research and Applications*, *17*(4), 343–362. https://doi.org/10.1504/ijbra.2021.10041400

Wei, W. J., Shi, B., Guan, X., Ma, J. Y., Wang, Y. C., & Liu, J. (2019). Mapping theme trends and knowledge structures for human neural stem cells: a quantitative and co-word biclustering analysis for the 2013-2018 period. *Neural Regeneration Research*, *14*(10), 1823–1832. https://doi.org/10.4103/1673-5374.257535

Xie, J., Ma, A., Fennell, A., Ma, Q., & Zhao, J. (2018). It is time to apply biclustering: A comprehensive review of biclustering applications in biological and biomedical data. *Briefings in Bioinformatics*,

20(4), 1449–1464. https://doi.org/10.1093/bib/bby014

Zhang, L., Chen, S., Ma, J., Liu, Z., & Liu, H. (2021). REW-ISA V2: A Biclustering Method Fusing Homologous Information for Analyzing and Mining Epi-Transcriptome Data. *Frontiers in Genetics*, *12*(5), 1–10. https://doi.org/10.3389/fgene.2021.654820