

Comparative Evaluation of Cheng and Church and BCBimax Biclustering Algorithms on Sustainable Development Goals Data

Destriana Aulia Rifaldi^{1*}, Indahwati¹, Muhammad Nur Aidi¹

¹Department of Statistics and Data Science, School of Data Science,
Mathematics and Informatics, IPB University, Indonesia

13auliarifaldi@apps.ipb.ac.id

ABSTRACT

Article History:

Received : 21-06-2026

Revised : 09-07-2026

Accepted : 03-08-2026

Online : 22-10-2026

Keywords:

ASR;

BCBimax;

Biclustering;

Cheng and Church;

SDGs.



Sustainable Development Goals (SDGs) comprise multidimensional indicators that reflect various aspects of sustainable development and often exhibit heterogeneous patterns across countries. Although biclustering has been widely applied to analyze high-dimensional data, comparative studies evaluating the performance of the Cheng and Church (CC) and Binary Inclusion-Maximal Biclustering (BCBimax) algorithms on SDGs data remain limited. Therefore, this study compares the performance of the CC and BCBimax algorithms in identifying local patterns in the 2024 SDGs achievement data. The dataset consists of 167 countries and 17 SDG indicators obtained from the *Sustainable Development Report*. Data exploration was conducted using heatmaps and PCA biplots, followed by biclustering analysis. The CC algorithm was applied to standardized numerical data, while BCBimax was implemented on binarized data using several thresholding approaches. Algorithm performance was evaluated using the number of biclusters, Average Squared Residue (ASR), Jaccard Index, and row membership coverage. The results indicate that the optimal CC model was obtained at $\delta = 0.14$, producing 7 biclusters with complete coverage of all 167 countries and an ASR value of 0.2465. In contrast, the optimal BCBimax model generated 38 biclusters with a lower ASR value of 0.0450, indicating greater homogeneity, but covered only 112 countries. The Jaccard Index of 30.61% suggests low similarity between the bicluster structures produced by the two algorithms. Overall, CC provides broader coverage and better represents numerical data patterns, whereas BCBimax produces more homogeneous and specific biclusters.



Crossref

<https://doi.org/10.31764/jtam.v10i4.40389>



This is an open access article under the [CC-BY-SA](#) license

A. INTRODUCTION

Biclustering is a development of clustering that allows the simultaneous grouping of rows and columns on a data matrix, to be able to identify local patterns or submatrices that have homogeneous characteristics and more specific relationships than conventional clustering (Castanho et al., 2024). This method is also known as co-clustering or two-way clustering because it performs the grouping of objects and variables simultaneously in a data matrix. The goal is to identify the relationships between the object group and the variable group so that local patterns can be found that cannot be uncovered by one-dimensional clustering (Wang et al., 2024). Recent studies have highlighted the growing application of biclustering across various domains and emphasized its flexibility in identifying local structures in complex and high-dimensional datasets (Lin et al., 2022). In general, biclustering algorithms can be classified

into five main approaches, namely greedy, divide and conquer, thorough enumeration, two-dimensional iterative grouping, and parameter distribution-based approach (Izzati et al., 2024).

One of the commonly used algorithms is Bimax, which can find maximum binary biclusters (Chu et al., 2022). Bimax is also often used as a reference in various modern methods due to its high interpretability, especially in binary data such as genomic data (Neumann & Miettinen, 2020). Various developments such as parallel Bimax also improve computing efficiency on big data (Rubio et al., 2020). In addition, synthetic benchmark datasets have recently been developed to provide a standardized framework for evaluating and comparing biclustering algorithms under different data characteristics, noise levels, and bicluster structures (Castanho et al., 2023). In contrast to BiMax, the Cheng and Church (CC) algorithm works on numerical data by using a Mean Squared Residue (MSR) or H-score measure to measure the coherence of a bicluster. This algorithm looks for submatrices that have an MSR value below a certain threshold (δ), so it can identify local patterns that are continuous and relatively more tolerant of data variation than methods that work on binary data (Di Iorio et al., 2020). However, CC tends to be less stable on data with high overlap than methods such as Bimax (Padilha & Campello, 2017).

One of the widely used binary biclustering algorithms is Binary Inclusion-Maximal Biclustering (BCBimax), which identifies inclusion-maximal biclusters in binary matrices and has been widely applied to discover coherent local patterns in binary datasets (Castanho et al., 2024; Nicholls & Wallace, 2021). Unlike the Cheng and Church (CC) algorithm, which directly analyzes numerical data by optimizing the Mean Squared Residue (MSR) criterion, BCBimax requires a binarization or discretization process before biclustering analysis can be performed (Castanho et al., 2024). Consequently, BCBimax is more suitable for identifying the presence or absence of specific characteristics in binary data, whereas CC is more appropriate for detecting coherent patterns in continuous numerical datasets (Karim et al., 2019). These methodological differences indicate that BCBimax and CC employ distinct strategies for discovering bicluster structures, and their performance may vary depending on the characteristics of the analyzed data (Nicholls & Wallace, 2021).

In the context of the Sustainable Development Goals (SDGs), the data generally consist of multiple indicators representing various dimensions of sustainable development, with heterogeneous characteristics across countries that make pattern identification challenging (Wani, 2024). Conventional clustering methods are limited because they partition either observations or variables globally, making them less effective in identifying local relationships that exist only within subsets of countries and indicators (Castanho et al., 2024). In contrast, biclustering simultaneously groups rows and columns, enabling the identification of countries that exhibit similar behavior only for specific subsets of SDG indicators and providing a more informative representation of heterogeneous data structures (Castanho et al., 2024). This capability makes biclustering an appropriate approach for exploring complex local patterns in high-dimensional data, where global clustering methods may fail to reveal meaningful substructures (Nicholls & Wallace, 2021).

Although biclustering has been widely applied in various domains, previous studies have predominantly evaluated biclustering algorithms using gene expression, biomedical, or simulated datasets, while comparative studies on multidimensional Sustainable Development

Goals (SDGs) data remain limited. Moreover, existing studies generally focus on evaluating individual algorithms or specific data types, providing limited evidence regarding the relative performance of numerical- and binary-based biclustering approaches on heterogeneous SDGs data. Therefore, this study compares the performance of the Cheng and Church (CC) and Binary Inclusion-Maximal Biclustering (BCBimax) algorithms in identifying bicluster structures in the 2024 SDGs dataset. The comparison is conducted using the number of biclusters, Mean Squared Residue (MSR), Average Squared Residue (ASR), Jaccard Index, and row membership coverage as evaluation criteria. By providing a comprehensive comparison of these two representative biclustering algorithms on multidimensional SDGs data, this study contributes practical evidence for selecting an appropriate biclustering approach according to the characteristics of heterogeneous sustainable development indicators.

B. METHODS

1. Data

This study is a quantitative comparative study that evaluates the performance of two biclustering algorithms, namely Cheng and Church (CC) and Binary Inclusion-Maximal Biclustering (BCBimax), in identifying local patterns in the Sustainable Development Goals (SDGs) dataset. The data used in this study is secondary data obtained from *the Sustainable Development Report*, which is an annual report published by the Sustainable Development Solutions Network (SDSN) and contains information on the achievement of the Sustainable Development Goals or SDGs various countries in the world. The report presents the results of country performance measurement based on a set of indicators designed to evaluate progress in the implementation of the 17 sustainable development goals set by the United Nations (UN). The data can be accessed openly through the <https://dashboards.sdgindex.org/explorer> page.

The observation units in this study consisted of 167 countries representing various geographical regions in the world, including Asia, Europe, America, Africa, and Oceania. The diversity of social, economic, and environmental characteristics between countries leads to different levels of sustainable development achievement. The variables used in this study are 17 indicators of the main SDGs, each of which represents one sustainable development goal. The indicators cover various dimensions of development, such as poverty alleviation, food security, health, education, gender equality, access to clean water and sanitation, clean and affordable energy, economic growth, industry and innovation, inequality reduction, sustainable urban development, responsible consumption and production, action on climate change, preservation of marine and terrestrial ecosystems, strong peace and institutions, and global partnerships for sustainable development.

Each indicator has a value in the range of 0–100, where higher values indicate better achievement towards the sustainable development goals in question. This set of indicators is then compiled in the form of a data matrix measuring 167×17 , with rows representing countries and columns representing SDGs indicators. The data structure allows the application of the biclustering method to identify groups of countries that have similar characteristics in a particular group of indicators simultaneously. An explanation of each of the SDG's indicators used in this study is presented in Table 1.

Table 1. Secondary Data Research Variables

Variables (Columns)	Remarks
Goal Score 1 ,(X ₁)	Without Poverty
Goal Score 2 ,(X ₂)	No Hunger
Goal Score 3,(X ₃)	Healthy and Prosperous Life
Goal Score 4 ,(X ₄)	Quality Education
Goal Score 5 ,(X ₅)	Gender Equality
Goal Score 6 ,(X ₆)	Clean Water and Proper Sanitation
Goal Score 7 ,(X ₇)	Clean and affordable energy
Goal Score 8 ,(X ₈)	Decent Work and Growth
Goal Score 9,(X ₉)	Industry, Innovation, and Infrastructure
Goal Score 10 ,(X ₁₀)	Reduced Inequality
Goal Score 11,(X ₁₁)	Sustainable Cities and Settlements
Goal Score 12,(X ₁₂)	Responsible Consumption and Production
Goal Score 13 ,(X ₁₃)	Addressing Climate Change
Goal Score 14 ,(X ₁₄)	Ocean Ecosystems
Goal Score 15,(X ₁₅)	Terrestrial Ecosystems
Goal Score 16,(X ₁₆)	Peace, Justice, and Resilient Institutions
Goal Score 17 ,(X ₁₇)	Partnerships to Achieve Goals

2. Analysis Steps

The stages of BCBimax and CC biclustering analysis are as follows. A flowchart is also given as an illustration of the steps of secondary data analysis in Figure.

- Compile SDGs indicator data in 167-observation units as rows and 17 research variables as columns into the form of a measured matrix. 167×17
- Data preparation by checking data quality, missing values and the existence of outliers. The missing value is imputed using the MICE (*Multiple Imputation by Chained Equations*) method to ensure complete and valid data.
- Although all SDGs indicators are in the range of 0–100, the standardization process is still carried out using the Z-score method to equalize the distribution between variables and reduce the influence of differences in variability between indicators. Through this process, each indicator has a mean of zero and a standard deviation of one so that its contribution becomes more balanced in the biclustering process and is not dominated by indicators that have a greater diversity of values (Wongoutong, 2024). The standardization process follows the pattern in equation 1 below.

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (1)$$

- Data exploration is carried out through data visualization such as heatmaps, to obtain an overview of data characteristics, value distribution, variations between indicators, and relationship patterns that may be formed before biclustering analysis (Yoo et al., 2022).
- The standardized SDGs data were transformed into binary data prior to applying the BCBimax algorithm. Five thresholding strategies were evaluated, namely global mean, variable mean, global median, variable median, and system threshold. These

thresholding strategies were compared to determine the most appropriate binary representation of the SDGs indicators. The optimal thresholding approach was selected based on the quality of the resulting biclusters, considering the Average Squared Residue (ASR) and the number of biclusters obtained.

- f. The minimum number of rows (minr) and columns (minc) for the BCBimax algorithm were determined through a grid search. Candidate values ranged from 2–15 for rows and 2–8 for columns. The optimal thresholding approach was selected based on a comprehensive evaluation of the resulting biclusters, considering the Average Squared Residue (ASR), the number of biclusters, and the Jaccard Index.
- g. The Cheng and Church (CC) algorithm was applied to the standardized numerical data. The delta parameter was determined through a grid search over the range of 0.01 to the global Mean Squared Residue (MSR) of the standardized data with an increment of 0.01. The optimal delta value was selected based on the lowest Average Squared Residue (ASR) while maintaining representative bicluster coverage. Subsequently, the BCBimax algorithm was applied to the binarized dataset using the optimal thresholding strategy and the selected minimum row-column parameters. MSR is a measure used to evaluate the level of coherence of a bicluster through the deviation of the element values from the row average, column average, and overall submatrix average (Alpha et al., 2024). The quality of the biclustering models was evaluated using complementary performance measures, including Mean Squared Residue (MSR), Average Squared Residue (ASR), Jaccard Index, and row membership coverage. These measures were jointly used to evaluate bicluster coherence, similarity between algorithms, and the proportion of countries represented by the resulting biclusters. MSR calculation formula is written as follows:

$$MSR(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{IJ} + a_{IJ})^2 \quad (2)$$

The size of the bicluster is expressed as $|I| \times |J|$, by $|I|$ the number of rows included in the bicluster and $|J|$, indicating the number of columns that make up the bicluster. Furthermore, the quality of a bicluster based on the value of the MSR can be evaluated by calculating the average MSR divided by that volume, otherwise known as the average MSR per volume (Putri et al., 2021), which is defined through the equation below.

$$rataaan \text{ MSR per volume} = \frac{1}{b} \sum_{i=1}^b \frac{MSR_i}{volume_i} \quad (3)$$

By b symbolizing the number of biclusters generated by a given algorithm.

- h. Interpret the results by identifying the performance of the two algorithms and the meaning of bicluster groups in the context of sustainable development.

- i. The evaluation of the quality of the biclusters obtained was carried out using Average Squared Residue (ASR), which is the average Mean Squared Residue (MSR) value of all the biclusters formed. Smaller ASR values indicate a higher degree of coherence so that the quality of the resulting biclusters is better (Castanho et al., 2022).

$$A_v MSR = \frac{1}{k} \sum_{i=1}^k MSR_i(I, J) \quad (4)$$

The existence of overlap between biclusters can affect the assessment of the similarity of biclustering results. Therefore, the corrected Jaccard Index is used to obtain a more accurate measure of similarity on intersecting bicluster structures (Hess et al., 2021). As per equation 5 below.

$$jac_c(B_1, B_2) = \frac{jac(B_1, B_2)}{\max(jac(B_1, B_2); jac(B_1, B_2))} \quad (5)$$

Since the maximum value of the Jaccard index between non-overlapping sets of biclusters is 1, this correction aims to maintain the validity of the comparison when there is redundancy or wedges between biclusters in one set, so that the degree of similarity obtained is not excessively affected by the existence of overlap between groups (Ma et al., 2023). In general, the steps of the analysis are listed in Figure 1.

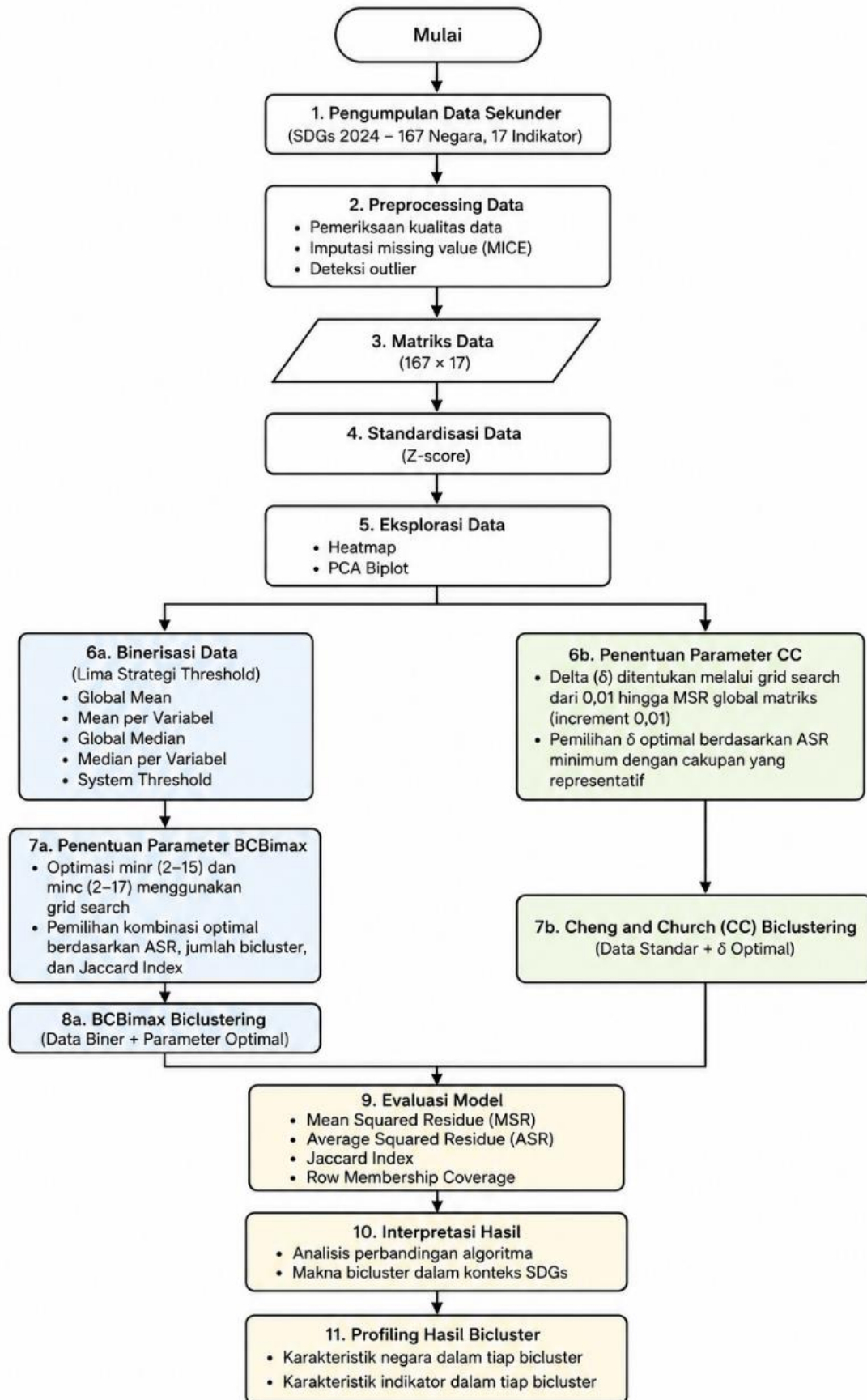


Figure 1. Research Flow

C. RESULT AND DISCUSSION

1. Data Exploration

The initial characteristics of each region based on the SDGs indicators of 167 countries are visualized through a heatmap formed from a standardized data matrix, as shown in Figure 2.

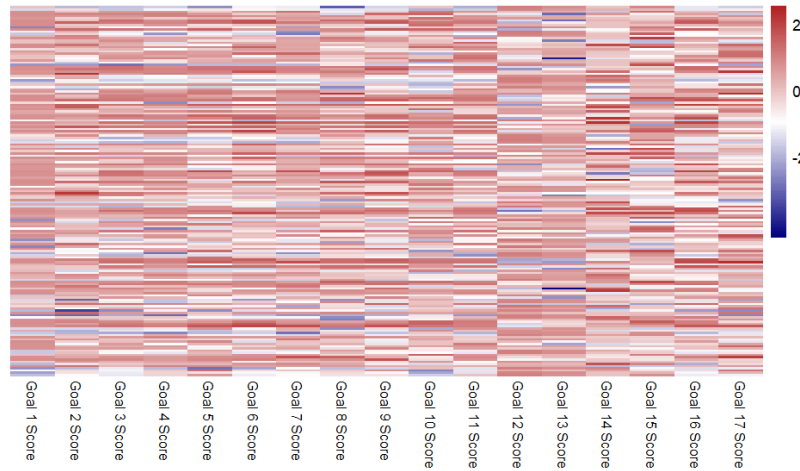


Figure 2. Heatmap scaling of SDGs

Figure 2 shows the heatmap of SDGs indicators after z-score standardization. The red color indicates above-average values, blue is below average, and white is close to average. The variation in achievement between countries can be seen quite large in Goal 3 (Healthy and Prosperous Lives), Goal 8 (Decent Work and Economic Growth), Goal 9 (Industry, Innovation, and Infrastructure), and Goal 10 (Reducing Disparities). Meanwhile, Goal 2 (No Hunger), Goal 6 (Clean Water and Decent Sanitation), Goal 7 (Clean and Affordable Energy), and Goal 15 (Terrestrial Ecosystems) tend to be more uniform between countries. Relatively similar color patterns in some parts indicate the presence of local structures in the data that have the potential to form biclusters, so biclustering analysis is needed to identify groups of countries and SDGs indicators that have similar characteristics. Furthermore, visualization is carried out with the PCA biplot listed in Figure 3.

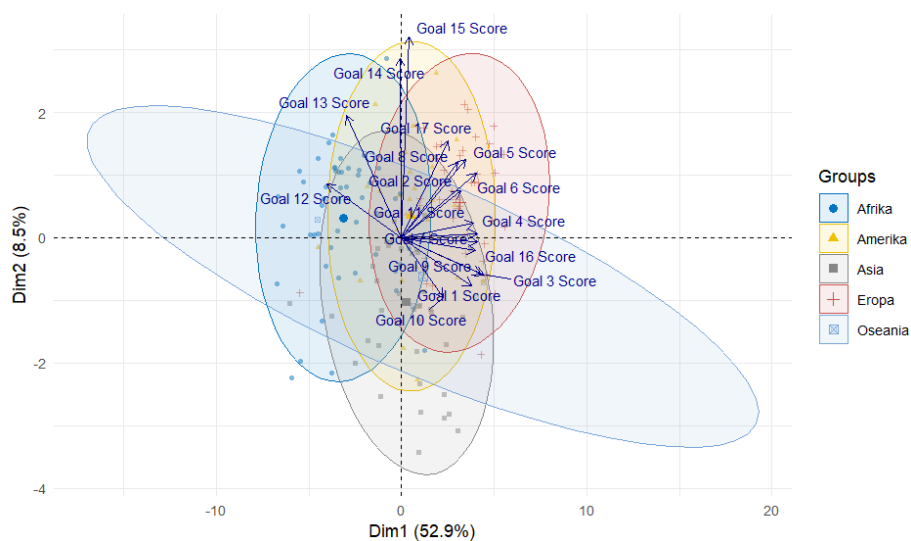


Figure 3. Biplot PCA SDGs Indicator

The results of projecting SDGs data into two-dimensional space show that Dim1 explains 52.9% of the data variation, while Dim2 is 8.5%, so the main variation between countries is mainly reflected in Dim1. With the grouping of continents, a clearer pattern can be seen, European and Oceanian countries tend to be on the right side of Dim1 which shows higher and relatively more equitable SDGs achievement, while African countries and parts of South Asia are on the left side which reflects development challenges, especially on basic indicators such as poverty and health. Dim2 provides additional variation in distinguishing countries with similar levels of Dim1, for example in high-income countries that still have different social and environmental characteristics. In general, this visualization shows a regional development gap, thus supporting the need for further analysis such as clustering or biclustering to identify more specific patterns.

2. Results of CC Biclustering Algorithm

Cheng and Church's (CC) analysis began with tuning the *delta* parameter in the range of 0.01–0.7174 based on a global MSR value of 0.7174. Evaluations using the number of biclusters, ASR, MSR per volume, and the Jaccard index showed that smaller delta values tended to result in more biclusters of better quality, while larger delta values resulted in fewer but less homogeneous biclusters. Therefore, the selection of deltas is carried out considering the balance between the number and quality of the biclusters, as shown in Table 2.

Table 2. Threshold Evaluation Results

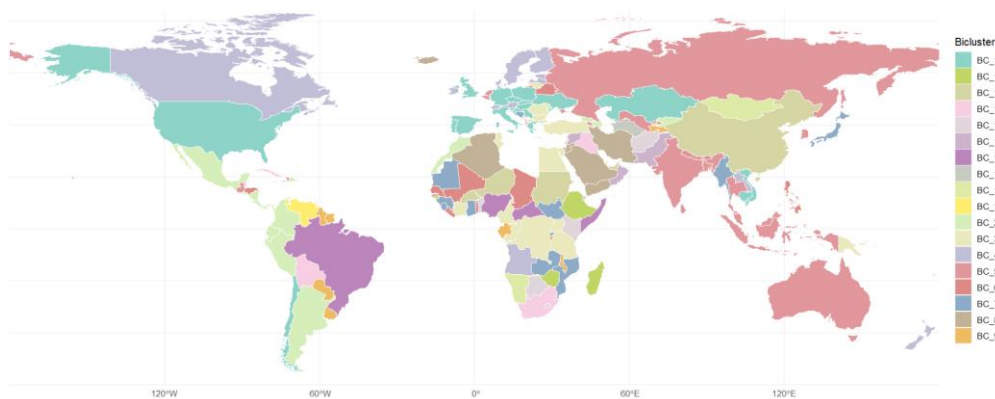
Delta Value	Quantity Bicluster	ASR	Track MSR per Volume	Table of Contents Jaccard
0,09	13	0,4278	0,0109	0,2999
0,14	7	0,2465	0,0069	1
0,17	6	0,5343	0,0074	0,4127
0,27	5	0,5101	0,0059	0,4307
0,50	2	0,5157	0,0004	0,5094

Based on Table 2, the delta parameter substantially affects the quality and number of biclusters produced by the Cheng and Church algorithm. Lower delta values impose stricter coherence requirements by restricting the allowable Mean Squared Residue (MSR), resulting in a larger number of highly homogeneous biclusters. Conversely, larger delta values permit greater variation within biclusters, leading to fewer but less coherent biclusters. This finding is consistent with the theoretical principle of the Cheng and Church algorithm, in which the delta parameter directly controls the maximum acceptable residue within a bicluster. Among the evaluated values, delta is 0.14 produced the best balance between bicluster quality and stability, as indicated by the lowest ASR value (0.2465), a low MSR per volume, and a Jaccard Index of 1. Therefore, delta 0.14 was selected as the optimal parameter for subsequent analysis, as shown in Table 3.

Table 3. Optimal Bicluster Characteristics of CC Algorithm

Dominant Characteristics	Number of Biclusters	Dominant SDGs Goals	Country Examples
High Reach	2	Goal 1 (No Poverty), Goal 3 (Healthy and Prosperous Lives), Goal 6 (Clean Water and Decent Sanitation), Goal 9 (Industry, Innovation, and Infrastructure), Goal 16 (Peace, Justice and Resilient Institutions)	United Kingdom, Georgia, Greece, Guatemala
Medium Reach	1	Most indicators have a Z-score value close to zero so that the achievement is relatively close to average	Denmark, Algeria, Estonia
Low Reach	2	Goal 1, Goal 3, Goal 6, Goal 7, Goal 9, Goal 10, and Goal 16 show a considerable negative Z-score value	Nigeria, Oman, Qatar, Saudi Arabia
Mixed Reach	2	Has a combination of indicators with high, medium, and low achievements simultaneously	Albania, Austria, Belgium

Based on Table 3, the seven biclusters generated by the CC algorithm can be classified into four dominant characteristics: high, medium, low, and mixed achievement. This classification indicates that the Cheng and Church algorithm effectively captures different levels of SDGs achievement while maintaining the continuous numerical relationships among indicators. Because the algorithm directly minimizes the Mean Squared Residue (MSR) on standardized numerical data, countries with similar overall achievement patterns can still be grouped together despite moderate differences in individual indicators. This characteristic explains why the CC algorithm successfully represented all 167 countries in the biclustering results. Then the visual visualization of the distribution of CC biclustering countries is shown in Figure 4.

**Figure 4.** Country Distribution of CC Algorithm Results

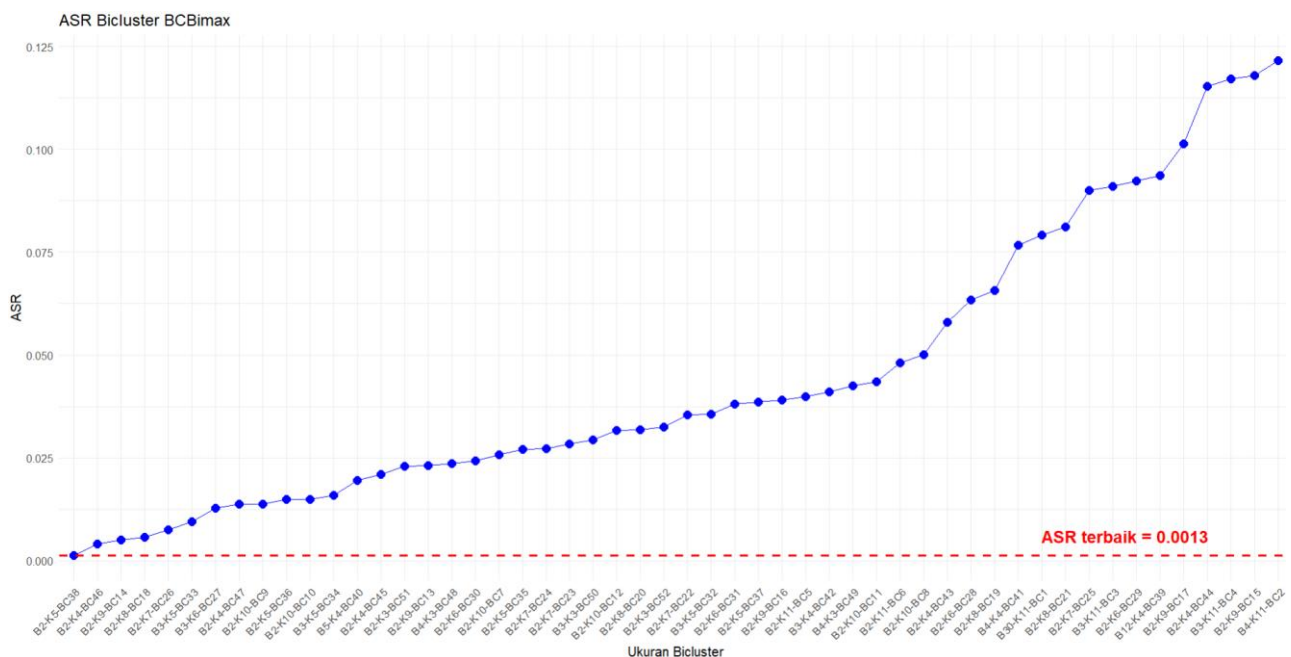
3. BCBimax Biclustering Algorithm Results

The BCBimax algorithm requires binary data, so SDGs data needs to be binary using a certain threshold value. The selection of thresholds is important because it affects the representation of data as well as the bicluster patterns that are formed. In this study, five threshold approaches were used, namely system, global mean, mean per change, global median, and median per change. A comparison of the results of the five approaches is presented in Table 4.

Table 4. BCBimax Algorithm Binary Scenario

Pendekatan	Threshold	Proportion of the value of "0"	Number of Biclusters
Threshold system	-0,973	0,190	26
Track global	0	0,443	0
Rataan peubah ^{*)}		0,443	52
Median global	0,184	0,500	52
Median peubah ^{*)}		0,503	55

Table 4 demonstrates that the choice of thresholding strategy substantially influences the binary data representation and consequently the biclustering results. The global median threshold produced the most representative binary matrix because it generated a balanced proportion of binary values while preserving local relationships among SDGs indicators. Compared with mean-based thresholds, the median is less sensitive to skewed distributions and extreme observations, making it more suitable for heterogeneous SDGs data. Consequently, the global median threshold produced biclusters with better structural quality and was selected for subsequent BCBimax analysis, as shown in Figure 5.

**Figure 5.** ASR Value of Row and Column Combination

Based on Figure 5, the combination of $\text{minr} = 2$ and $\text{minc} = 5$ resulted in the lowest ASR while simultaneously producing the highest number of biclusters. This parameter combination provides a balance between bicluster size and coherence. Smaller minimum dimensions enable the detection of localized development patterns that might otherwise be overlooked, whereas larger minimum dimensions tend to eliminate smaller but meaningful biclusters. Therefore, the selected parameter combination offers a more representative description of heterogeneous SDGs achievement patterns, as shown in Figure 6.

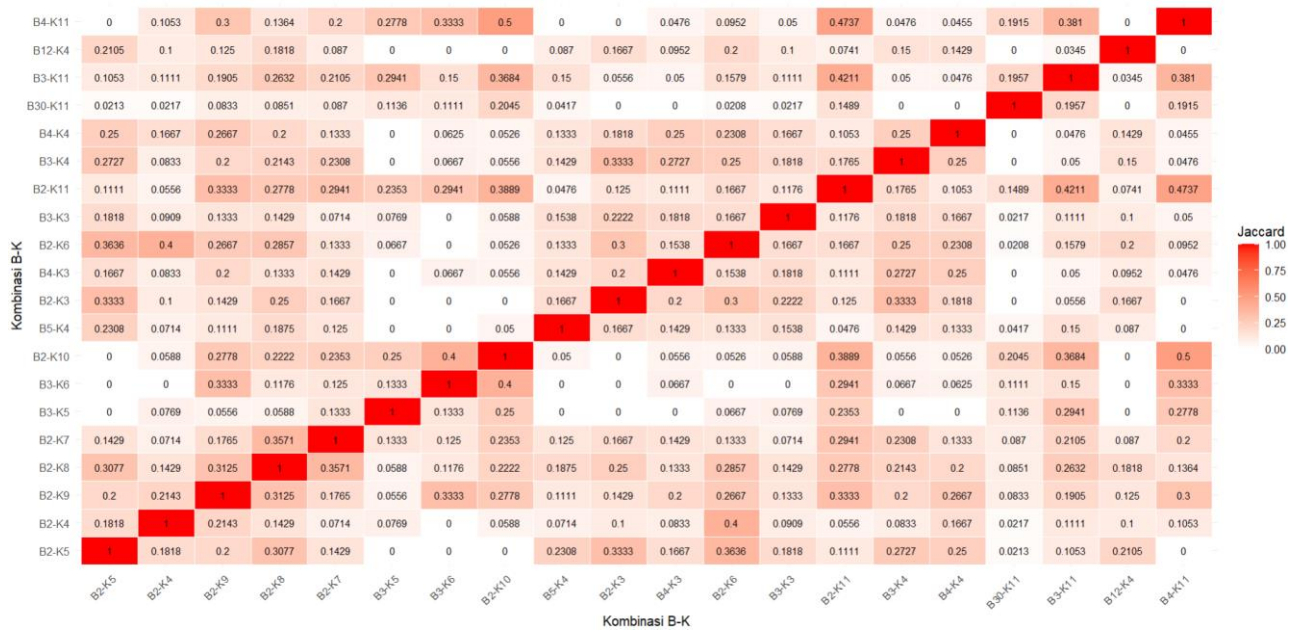


Figure 6 Index values of jaccard combinations of rows and columns

Based on Figure 6, most pairs of B-K parameter combinations show low to moderate Jaccard index values, which is indicated by the predominance of white to pink colors. The highest Jaccard value is seen on the main diagonal with a value close to 1, which indicates perfect similarity on the same parameter combination. Beyond the diagonal, only a few parameter combinations have relatively high Jaccard values, indicating the presence of a consistent bicluster structure. Conversely, the low Jaccard value in most combinations suggests that changes in parameters B and K have a considerable effect on biclustering results. Thus, these results indicate that the model is sensitive to parameter selection, so that a combination with higher Jaccard values can be considered as a more stable and representative parameter in forming a bicluster structure. Furthermore, these results are classified into high, medium, and low categories based on the concept of standard deviation classification, which is the interpretation of the position of values to the mean and one standard deviation after the data standardization process, as shown in Table 5.

Table 5. Optimal Bicluster Characteristics of the BCBimax Algorithm

Dominant Characteristics	Number of Biclusters	Dominant SDGs Goals	Country Examples
High Reach	8	Gender Equality (Goal 5), Decent Work and Economic Growth (Goal 8), Industry, Innovation and Infrastructure (Goal 9), Sustainable Cities and Settlements (Goal 11), Life on Land (Goal 15), Peace, Justice and Resilient Institutions (Goal 16), Partnerships to Achieve the Goals (Goal 17)	Austria, Australia, Japan, Singapore
Medium Reach	13	Most of the SDGs goals are in the medium category without the dominance of certain indicators	Malta, Tunisia, Peru, Costa Rica, Thailand
Mixed Reach	17	Some of the goals show high achievement,	Armenia,

Dominant Characteristics	Number of Biclusters	Dominant SDGs Goals	Country Examples
		notably Goal 14 (Underwater Life), Goal 15 (Life on Land), and Goal 17 (Partnership to Achieve Goals), while others are in the medium category	Vietnam, Indonesia, Barbados, dan Kenya

The results of biclustering in Table 5 show that the pattern of achievement of SDGs between countries can be grouped into three main characteristics. A total of 8 *biclusters* are dominated by countries with high achievements on various SDGs, 13 *biclusters* show relatively even moderate achievements, while the other 17 *biclusters* have mixed patterns with a combination of high and medium achievements in different development dimensions. These findings suggest that the characteristics of development between countries are heterogeneous and form specific local patterns. Then the visual visualization of the distribution of countries as a result of BCBimax biclustering is shown in Figure 7.

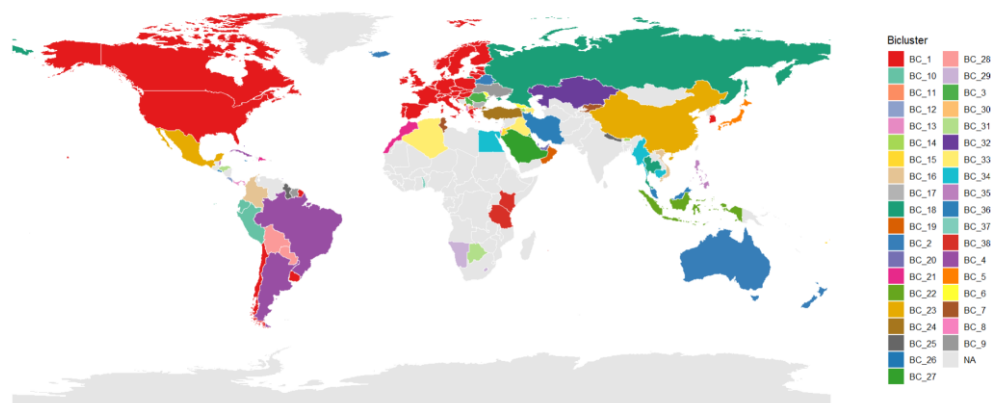


Figure 7. Country Distribution of BCBimax Algorithm

Based on Figure 7, countries tend to be concentrated in certain biclusters that show a similar pattern of achievement of SDGs indicators. On the other hand, several countries are members of different biclusters, reflecting the heterogeneity of development characteristics between countries. Countries such as Afghanistan, Pakistan, and Uganda do not fall into biclusters (non-BC) because they do not have a binary indicator pattern that is homogeneous enough to meet the criteria for bicluster formation. In general, these results show that the achievement of the SDGs shapes groups of countries with different development characteristics.

The findings of this study are generally consistent with previous studies on biclustering methods. The Cheng and Church (CC) algorithm demonstrated broader coverage by including all countries in the resulting biclusters, confirming its ability to identify coherent patterns in continuous numerical data. In contrast, BCBimax produced a larger number of biclusters with a lower Average Squared Residue (ASR), indicating greater homogeneity within each bicluster, although fewer countries were included in the overall biclustering results. These findings support previous studies reporting that binary biclustering methods tend to identify more specific and homogeneous local patterns, whereas numerical biclustering methods are more

suitable for capturing broader relationships in continuous datasets. Furthermore, the relatively low Jaccard Index indicates that the two algorithms identify different bicluster structures, suggesting that the choice of biclustering algorithm should be adapted to the characteristics of the analyzed data and the objectives of the study.

4. Performance Evaluation of Biclustering Algorithms

From a practical perspective, the biclustering results provide valuable information for mapping SDGs achievement across countries. The Cheng and Church algorithm is more appropriate for obtaining a comprehensive overview of global development patterns because it successfully represents all countries in the biclustering structure. In contrast, BCBimax is more suitable for identifying highly specific groups of countries that share similar strengths and weaknesses across selected SDG indicators. Such information can support policymakers in designing targeted development strategies, prioritizing international cooperation among countries with similar development profiles, and allocating development resources more efficiently. Therefore, biclustering offers a more informative alternative to conventional clustering for supporting evidence-based policy formulation in sustainable development.

The next stage is to evaluate the performance of the CC and BCBimax algorithms based on the quality of the biclustering results obtained. The evaluation was carried out using the number of biclusters formed, the Jaccard index to measure the similarity of the bicluster membership, the Average Squared Residue (ASR) value to assess the homogeneity of the biclusters, and the number of row memberships to see the coverage of the objects that were successfully grouped by each algorithm. Details of the size of the evaluation are described in Table 6.

Table 6. Results of CC and BCBimax Algorithm Evaluation

Algoritma	Jaccard Index	Number of Biclusters	ASR Values	Line Membership
CC	30,61%	7	0,2465	167
BCBimax		38	0,0450	112

Based on Table 6, the Jaccard Index of 30.61% indicates relatively low similarity between the bicluster structures generated by the CC and BCBimax algorithms, suggesting that each algorithm captures different local patterns within the SDGs dataset. This difference can be explained by the mathematical characteristics of the two algorithms. The Cheng and Church algorithm constructs biclusters by minimizing the Mean Squared Residue (MSR) on continuous numerical data. As long as the MSR remains below the specified delta threshold, countries with similar overall numerical patterns can still be included in the same bicluster. Consequently, CC successfully grouped all 167 countries, providing complete row coverage and a more comprehensive representation of the dataset. In contrast, BCBimax operates on binary data and identifies inclusion-maximal biclusters that require highly consistent binary patterns across both rows and columns. Countries that do not satisfy these strict binary conditions are excluded from biclusters, resulting in lower row coverage (112 countries) but substantially higher within-bicluster homogeneity, as reflected by the lower ASR value (0.0450). Therefore, the two algorithms exhibit complementary strengths: CC is more suitable for obtaining broad

representations of continuous numerical data, whereas BCBimax is more effective in identifying highly homogeneous and specific local structures.

D. CONCLUSION AND SUGGESTIONS

This study achieved its objective of evaluating and comparing the performance of the Cheng and Church (CC) and Binary Inclusion-Maximal Biclustering (BCBimax) algorithms in identifying local patterns within the 2024 Sustainable Development Goals (SDGs) dataset. The comparison demonstrates that the two algorithms exhibit complementary characteristics rather than indicating that one algorithm is universally superior. The CC algorithm generated seven biclusters and successfully covered all 167 countries, making it more suitable for representing broad numerical relationships and providing a comprehensive overview of SDGs achievement patterns. In contrast, BCBimax produced 38 biclusters with a lower Average Squared Residue (ASR), indicating higher internal homogeneity, although it covered fewer countries due to its stricter binary inclusion criteria. These findings contribute to the biclustering literature by providing an empirical comparison of numerical- and binary-based biclustering algorithms on multidimensional SDGs data, demonstrating that algorithm selection should be aligned with both the characteristics of the data and the analytical objectives. Furthermore, the identified bicluster structures reveal that SDGs achievements vary across groups of countries, highlighting the importance of local pattern analysis for understanding heterogeneous sustainable development performance.

Future research is recommended to extend this work by applying biclustering methods to SDGs datasets covering multiple years to investigate the temporal dynamics of sustainable development patterns across countries. In addition, comparisons involving other biclustering algorithms, such as Plaid, Spectral Biclustering, xMOTIFs, or FABIA, would provide a more comprehensive understanding of the strengths and limitations of different biclustering approaches. Future studies are also encouraged to incorporate additional evaluation metrics, such as row and column coverage, bicluster relevance, or stability measures, to obtain a more comprehensive assessment of bicluster quality. These developments are expected to enhance the robustness of biclustering analysis and broaden its application in supporting evidence-based decision-making for sustainable development.

ACKNOWLEDGEMENT

The author would like to thank the supervisor for the guidance, direction, and input provided during the research process. The author also expressed his gratitude to the Department of Statistics, Bogor Agricultural University, for the academic support and facilities that have been provided so that this research can be completed properly.

REFERENCES

- Alfó, M., Marino, M. F., & Martella, F. (2024). Biclustering multivariate discrete longitudinal data. *Statistics and Computing*, 34(1), 12. <https://doi.org/10.1007/s11222-023-10292-6>
- Castanho, E. N., Aidos, H., & Madeira, S. C. (2022). Biclustering fMRI time series: a comparative study. *BMC Bioinformatics*, 23(1), 223. <https://doi.org/10.1186/s12859-022-04733-8>
- Castanho, E. N., Lobo, J. P., Henriques, R., & Madeira, S. C. (2023). G-bic: Generating synthetic benchmarks for biclustering. *BMC Bioinformatics*, 24(1), 457. <https://doi.org/10.1186/s12859-022-04733-8>

- Castanho, E. N., Aidos, H., & Madeira, S. C. (2024). Biclustering data analysis: a comprehensive survey. In *Briefings in Bioinformatics*, 25(4). Bbae342. <https://doi.org/10.1093/bib/bbae342>
- Chu, H. M., Liu, J. X., Zhang, K., Zheng, C. H., Wang, J., & Kong, X. Z. (2022). A binary biclustering algorithm based on the adjacency difference matrix for gene expression data analysis. *BMC Bioinformatics*, 23(1), 436. <https://doi.org/10.1186/s12859-022-04842-4>
- 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. In *Bioinformatics*, 36(9), 2955-2957. <https://doi.org/10.1093/bioinformatics/btaa060>
- Hess, S., Pio, G., Hochstenbach, M., & Ceci, M. (2021). BROCCOLI: overlapping and outlier-robust biclustering through proximal stochastic gradient descent. *Data Mining and Knowledge Discovery*, 35(6), 2542–2576. <https://doi.org/10.1007/s10618-021-00787-z>
- Izzati, H., Indahwati, & Djuraidah, A. (2024). BCBimax Biclustering Algorithm with Mixed-Type Data. *JUITA: Jurnal Informatika*, 12(1), 131–139. <https://doi.org/https://doi.org/10.30595/juita.v12i1.21519>
- Karim, M. B., Kanaya, S., & Altaf-Ul-Amin, M. (2019). Implementation of BiClusO and its comparison with other biclustering algorithms. *Applied Network Science*, 4(1), 98. <https://doi.org/10.1007/s41109-019-0180-x>
- Ma, T., Guo, L., Yan, H., & Wang, L. (2023). Cobind: quantitative analysis of the genomic overlaps. *Bioinformatics Advances*, 3(1), vbad104. <https://doi.org/10.1093/bioadv/vbad104>
- Neumann, S., & Miettinen, P. (2020). Biclustering and boolean matrix factorization in data streams. *Proceedings of the VLDB Endowment*, 13(10), 1709–1722. <https://doi.org/10.14778/3401960.3401968>
- Nicholls, K., & Wallace, C. (2021). Comparison of sparse biclustering algorithms for gene expression datasets. *Briefings in Bioinformatics*, 22(6), bbab140. <https://doi.org/10.1093/bib/bbab140>
- Padilha, V. A., & Campello, R. J. G. B. (2017). A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics*, 18(1), 55. <https://doi.org/10.1186/s12859-017-1487-1>
- Pio, G., Ceci, M., & Malerba, D. (2021). Biclustering methods for biomedical data analysis: A survey based on methodological, biological and application perspectives. *Briefings in Bioinformatics*, 22(6), bbab421. <https://doi.org/10.1088/1742-6596/1863/1/012068>
- 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>
- Rubio, A. S., Viveros, A. M., Morales Luna, G. B., & López, M. P. (2020). Towards BiMAX: Binary inclusion-maximal parallel implementation for gene expression analysis. *Computacion y Sistemas*, 24(1), 255–267. <https://doi.org/10.13053/CyS-24-1-2979>
- Wang, H., Song, Y., Chen, W., Luo, Z., Li, C., & Li, T. (2024). A Survey of Co-Clustering. *ACM Transactions on Knowledge Discovery from Data*, 18(9), Article 223. <https://doi.org/10.1145/3681793>
- Wani, A. A. (2024). Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. *PeerJ Computer Science*, 10, 1–45. <https://doi.org/10.7717/PEERJ-CS.2286>
- Wongoutong, C. (2024). The impact of neglecting feature scaling in k-means clustering. *PLoS ONE*, 19(12), e0310839. <https://doi.org/10.1371/journal.pone.0310839>
- Yoo, K. H., Leung, C. K., & Nasridinov, A. (2022). Big Data Analysis and Visualization: Challenges and Solutions. In *Applied Sciences (Switzerland)*, 12(16), 8248. <https://doi.org/10.3390/app12168248>