InPrime: Indonesian Journal of Pure and Applied Mathematics
Vol. 8, No. 1 (2026), pp. 56 – 70, doi: 10.15408/inprime.v8i1.50887
p-ISSN 2686-5335, e-ISSN: 2716-2478

The Influence of the Binary Process Threshold on the BCBimax Algorithm in Finding Optimal BC

Femmy Diwidian[1,2], I Made Sumertajaya[2]*, Indahwati[2], Hari Wijayanto[2], and Bagus Sartono[2]
[1]Mathematics Education Program, Faculty of Education and Teaching Science,
UIN Syarif Hidayatullah Jakarta, Tangerang Selatan, Banten, Indonesia
[2]Statistics and Data Science Study Program, Bogor Agricultural University, Bogor, West Java, Indonesia
Email:*imsjaya@apps.ipb.ac.id

Abstract

Selecting an appropriate threshold is crucial in the binarization process because it directly affects the quality of biclusters generated by the BCBimax algorithm. However, studies comparing thresholding strategies for optimal bicluster formation are still limited. Therefore, this study proposes an ASR-based evaluation framework to analyze the effect of several thresholding methods on biclustering performance. This study uses 2024 fruit production data from 34 provinces and three special regions in Indonesia, with 25 variables representing fruit commodities. Data preprocessing includes scaling and threshold determination using global and variable-based statistical approaches. The novelty of this research lies in the comparative analysis of threshold selection strategies in BCBimax biclustering using Average Spearman's Rho (ASR) to evaluate bicluster quality. The results show that the variable-wise median produces patterns similar to the global median, while the variable-wise mean yields the same ASR value as the global mean (0.56909). Although the system-generated threshold yields the smallest ASR (0.09533), it produces only one bicluster. These findings highlight the significant influence of threshold selection on the quality and interpretability of biclustering.
Keywords: BcBimax algorithm; Binarization; Threshold.

Abstrak

Pemilihan nilai ambang batas yang tepat sangat penting dalam proses binarisasi karena secara langsung mempengaruhi kualitas bicluster yang dihasilkan oleh algoritma BCBimax. Namun, studi yang membandingkan strategi penentuan ambang batas untuk pembentukan bicluster optimal masih terbatas. Oleh karena itu, penelitian ini mengusulkan kerangka evaluasi berbasis ASR untuk menganalisis pengaruh beberapa metode penentuan ambang batas terhadap kinerja biclustering. Penelitian ini menggunakan data produksi buah tahun 2024 dari 34 provinsi dan tiga daerah khusus di Indonesia, dengan 25 variabel yang mewakili komoditas buah. Praproses data meliputi penskalaan dan penentuan ambang batas menggunakan pendekatan statistik global dan berbasis variabel. Kebaruan penelitian ini terletak pada analisis komparatif strategi pemilihan ambang batas dalam biclustering BCBimax menggunakan Average Spearman's Rho (ASR) untuk mengevaluasi kualitas bicluster. Hasil menunjukkan bahwa median per variabel menghasilkan pola yang mirip dengan median global, sedangkan rata-rata per variabel menghasilkan nilai ASR yang sama dengan rata-rata global (0,56909). Meskipun ambang batas yang dihasilkan sistem menghasilkan nilai ASR terkecil (0,09533), ambang batas tersebut hanya membentuk satu bicluster. Temuan ini menyoroti pengaruh signifikan pemilihan ambang batas terhadap kualitas dan interpretasi biclustering.
Kata Kunci: *Algoritma BcBimax; Binerisasi; Ambang batas.*

2020MSC: 62H30

*) Corresponding author
Submitted April 22nd, 2026, Revised May 15th, 2026,
Accepted for publication May 24th, 2026, Published Online May 31st, 2026

1. INTRODUCTION

Statistical analysis is important for organizing and understanding complex, unstructured data. Clustering is a valuable analysis for finding hidden patterns, simplifying analysis, and providing deeper insight into the structure of unstructured data. Cluster analysis is a method for grouping objects based on the similarity of their characteristics [1]. However, classical cluster analysis has limitations in classifying objects based on rows or columns, so the diversity of information is often not appropriately grouped [2]. To overcome these limitations, the biclustering technique was developed. This technique enables two-way clustering, grouping rows and columns with similar characteristics to produce more in-depth, useful information [3]. Lai et al. [4] explained that biclustering was developed to overcome the weaknesses of clustering analysis in grouping data matrices.

Clustering methods in two-way analysis (biclustering) have been widely developed. Madeira and Oliveira [5] stated that there are five main categories of biclustering algorithms or techniques for finding biclusters. Divide and Conquer is a method that divides the initial matrix into a set of subproblems of smaller size. The BCBimax algorithm finds biclusters using a divide-and-conquer approach. Prelić et al. [6] The BCBimax algorithm finds biclusters by identifying a matrix that contains all 1s in the binary data matrix [7]

The BCBimax algorithm is a biclustering method designed to detect submatrices with uniform binary values in binary data; therefore, the binarization process must be performed before running the algorithm. Binarization is a widely used process in several studies because it simplifies data processing. Lejeune et al. [8] explained that the process of transforming data into binary data was first proposed by Boros, Hammer, Ibaraki, and Kogan in 1997, which is currently known as binarization. Rodriguez-Baena et al. [9] also explained that transforming data into a set of binary data can be one way to concisely explain several relationships between a group of objects and possible properties. Mayoraz and Moreira [10] Also stated that transforming data into binary data will retain the most important information, especially in classification problems. One important aspect of binarization is determining the threshold value.

The threshold value determines the point at which the data will be classified into one of two binary categories. Choosing the correct threshold value is very important because it can affect the quality and accuracy of the binarization results. Considering the threshold in the binarization process is known as the cutpoint. The selection of the cutpoint or threshold, in some cases, is adjusted to several interests of the research being conducted. However, the median can be considered in several cases where the cutpoint or threshold value is unclear [9]. According to Anthony and Ratsaby [11], choosing the wrong threshold will affect the results, such as losing important information in the data. Several studies related to images also mention that choosing a threshold that does not match the characteristics of the image can cause segmentation errors, such as over-segmentation or under-segmentation [12][13][14]. Therefore, by considering the ASR Evaluation, this study will discuss several important things related to the threshold and research data on the binarization process in selecting the optimal Bicluster. The agricultural sector is an important contributor to Indonesia's economic growth. One of the agricultural subsectors, namely horticulture, still has potential to be encouraged to improve farmers' welfare, the regional and national economies, and even increase the country's foreign exchange through exports.

Based on BPS data, the production of Indonesian plants and fruits continues to increase. In 2021, it reached 25.96 million tons, an increase of 5.4% from 2020's production of 24.63 million tons. The commodities with the most significant production volume are bananas (8.74 million tons/33.67%),

pineapples (2.89 million tons/11.13%), mangoes (2.84 million tons/10.94%), tangerines (2.4 million tons/9.24%), and durians (1.35 million tons/5.2%). Two-way clustering is expected to be one way to manage areas with plant and fruit producers in Indonesia. This article is expected to be one of the considerations in mapping regional potential based on similarities in climate, soil, and topography, thus ensuring optimal plant growth. The BCBimax algorithm, which is relatively fast and precise at finding biclusters, is expected to group the potential of optimally produced fruit. In addition, this article will discuss the influence of threshold selection on finding optimal biclusters, considering the number of BCs, overlapping membership, and bicluster evaluation.

2. METHOD

2.1. Data

This study uses secondary data from the official BPS website (https://www.bps.go.id). The data used is fruit and vegetable production data in 2023, released in 2024. The observation unit in this study covers 34 provinces in Indonesia, with the addition of special areas in Papua, namely West Papua, Southwest Papua, Central Papua, and Papua Pegunungan. The variables used in this study consist of 25 variables that represent the amount of production for each type of fruit and vegetable in Quintal (Kw) units. The elements that make up the data matrix in this study are presented in Table 1.

Table 1. Elements of Data Matrix in Fruit and Vegetable Production

Provinces (Rows)	Variables (Coluamns)
34 Provinces and 3 Special Regions of Papua	Avocado (X1), Starfruit (X2), Duku (X3), Durian (X4), Guava (X5), Guava (X6), Jengkol (X7), Siamese orange (X8), Large orange (X9), Mango (X10), Mangosteen (X11), Melinjo (X12), Jackfruit (X13), Pineapple (X14), Papaya(X15), Petai (X16), Banana (X17), Rambutan (X18), Salak (X19), Sapodilla (X20), Soursop (X21), Breadfruit (X22), Dragon Fruit (X23), Lemon (X24), Longan (X25)

2.2. Research Stages

The stages of research carried out can be written in several core steps, including:

1. Data pre-processing
   The data used in this study is the 2023 Fruit and Vegetable Production data released in 2024. The production of fruit and vegetable crops in each province varies. Scaling is an important step in data pre-processing that helps ensure the model works more efficiently and accurately by minimizing the influence of non-uniform data scales. Data pre-processing is done by scaling the fruit and vegetable production data matrix. After scaling, the data will be visualized as a heatmap.
2. The Binarization Process uses several Thresholds.
   The following process changes the data matrix into a binary data matrix. The threshold in the binarization process in this article is important. Several thresholds will be considered in determining the effect of threshold selection in finding the optimal BC. Some of the thresholds used are as follows:
   a) The median of each variable

The median is one of the threshold considerations that is often used. [15][16][9] explained that the median can be considered when no specific provisions exist in choosing a threshold. Wulandari et al. [7] mentioned the median of each variable as the best threshold for finding Optimal BC. The different production potentials of fruits and vegetables cause the median of each variable to be considered in the binarization process in forming a binary data matrix. In calculating the median value, the following equation can be used:

$$\hat{y}_p = median_p\{y_{11}, y_{21}, y_{31}, \dots, y_{mn}\}, \quad (1)$$

where $\hat{y}_p$ indicates the median value of the missing data to be calculated, y_{mn} indicates the value of each element of the data matrix of the m-th column and the n-th row.

b) Median of All Data
After the initial scaling of the data, the median of all the data ($Median_{all}$) can be considered. Data scaling, which makes the data scale uniform, is the next consideration. This can also provide additional new information on the influence of scaling on optimal BC clustering results. The median threshold for all data can be written as follows. Suppose we get a mxn data matrix; m indicates the number of observations, and n indicates the number of variables used. Calculate the number of elements $N = m \times n$; if the position of N is odd, then

$$Median_{all} = x_{\left(\frac{N+1}{2}\right)}. \quad (2)$$

c) Mean of Each Variable
Mean becomes the next consideration; for example, a mxn data matrix exists. The Mean of each variable can be written as follows:

$$\bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij}, \text{ for } i = 1,2,\cdots,m, j = 1,2,\cdots,n. \quad (3)$$

d) Mean of All Data
The next consideration is the Mean of all the data; the Mean of all the data can be additional information by paying attention to the influence of outliers in the data after scaling.

$$\bar{x} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}, \quad (4)$$

where N is the total number of data elements.

e) System Threshold
The system threshold is one of the thresholds used when there is no reference to determine the threshold value in data. Suppose there is a data matrix (x). The system threshold is calculated using the following formula:

$$System\ Threshold = \frac{\min(x)+\max(x)}{2}. \quad (5)$$

3. Running the BCBimax algorithm on each binarization threshold
The BCBimax algorithm is applied to each binary matrix obtained from the thresholding process, considering combinations of minimum row and column sizes (from 2 to 10). Define the binarization as:

$$b_{ij} = \begin{cases} 1, & x_{ij} \geq T, \\ 0, & x_{ij} < T. \end{cases}$$

Let B denote a binary matrix with row set R and column set C. The algorithm proceeds by recursively partitioning the matrix as follows:

First, the column set C is divided into two subsets, C_U and C_V, such that the initial rows in C_U contain elements equal to 1. Then, the row set $R$is partitioned into three subsets:

- $G_U = \{i \in R \mid B_{ij} = 1 \text{ for } j \in C_U\}$,
- $G_V = \{i \in R \mid B_{ij} = 1 \text{ for } j \in C_V\}$,
- $G_W = G_U \cap G_V$.

Next, two submatrices are constructed:

$$U = (G_U \cup G_W, C_U), V = (G_W \cup G_V, C_U \cup C_V).$$

The same partitioning procedure is recursively applied to submatrices $U$and Vuntil submatrices consisting entirely of ones are obtained. Each such submatrix is recorded as a bicluster, provided that it satisfies the minimum row and column constraints. To avoid redundancy, identified biclusters may be removed from the search space before continuing the process. The algorithm terminates when no further valid submatrices can be generated. Pseudocode for BCBimax Algorithm is:

Algorithm 1. Threshold-Based BCBimax Biclustering

```
Input:
  X : Real-valued data matrix (m × n)
  T : Threshold value
  minRow : Minimum number of rows
  minCol : Minimum number of columns
  Output:
  ℬ : Set of biclusters
---------------------------------------------------
Step 1: Binarization
---------------------------------------------------
1.1. Construct binary matrix B = (b_ij) as:
     For each i = 1,...,m and j = 1,...,n:
        If x_ij ≥ T then
           b_ij ← 1
        Else
           b_ij ← 0
---------------------------------------------------
Step 2: BCBimax Procedure
---------------------------------------------------
2.1. Initialize ℬ ← ∅
2.2. Define recursive function BCBimax(B):
  a. If number of rows < minRow OR number of columns < minCol:
        Return ∅
  b. If all elements in B are equal to 1:
        Add B to ℬ
        Return
  c. Select a column set C and partition into C_U and C_V
  d. Partition row set R into:
```

G_U = {i | row i has 1s in C_U}
G_V = {i | row i has 1s in C_V}
G_W = G_U ∩ G_V
e. Construct submatrices:
U = (G_U ∪ G_W, C_U)
V = (G_W ∪ G_V, C_U ∪ C_V)
f. Recursively apply:
BCBimax(U)
BCBimax(V)
2.3. Run BCBimax(B)

Step 3: Output

3.1. Return $\mathcal{B}$

4. Evaluation of biclusters obtained from each threshold.
After running the BCBimax algorithm on all combinations of row and column thresholds performed. The following process is carried out by evaluating the BC results obtained. [17][3]]explain that one simple hypothesis for evaluating biclusters is to calculate the bicluster variance with the following equation:

$$MSR_{(I,J)} = \frac{1}{|I|\times|J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2, \quad (5)$$

$$\text{with } a_{iJ} = \frac{1}{m}\sum_{j=1}^{m} a_{ij} \;;\; a_{Ij} = \frac{1}{n}\sum_{i=1}^{n} a_{ij} \;;\; a_{IJ} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}. \quad (6)$$

BC evaluation is obtained by calculating the ASR value

$$ASR = \frac{1}{n}\left(\sum_{i=1}^{n} E_{MSR_i}(I,J)\right), \quad (7)$$

where a_i describes the (i,j) element of the Bicluster A_{IJ}, a_{ij} represents the overall average of the elements in the Bicluster. $|I|$ represent the total number of rows, and $|J|$ represents the total number of columns.

5. Optimal Bicluster selection from each threshold
After calculating the ASR value at each threshold, the process will be evaluated by considering several things, such as profiling, overlapping bicluster membership, ASR value, and the number of biclusters formed.

3. RESULT

3.1. Data Exploration

The first stage of this study was scaling. Scaling was carried out to determine the value of fruit and vegetable production for each province and special region of Papua, which has a uniform value scale. After scaling, the next step is to visualize the results as a heatmap, as shown in Figure 1. This figure shows that almost all variables have the potential for vegetable and fruit production in East Java, Central Java, and West Java. The regions of North Sumatra and Lampung are major producers of vegetables and fruits. Several other regions produce vegetables and fruit that are typically small to medium. Based on the heatmap, selecting the mean threshold is a concern. This is because the East

Java region has fruit production that tends to be higher than in several other regions. Data matrices containing outliers can have fewer elements, so bicluster clustering will also be less. The BCBimax algorithm enables clustering of areas with predominantly high production levels. At the same time, some areas that tend to be medium to low cannot be clustered. Therefore, the resulting clustering is less informative.

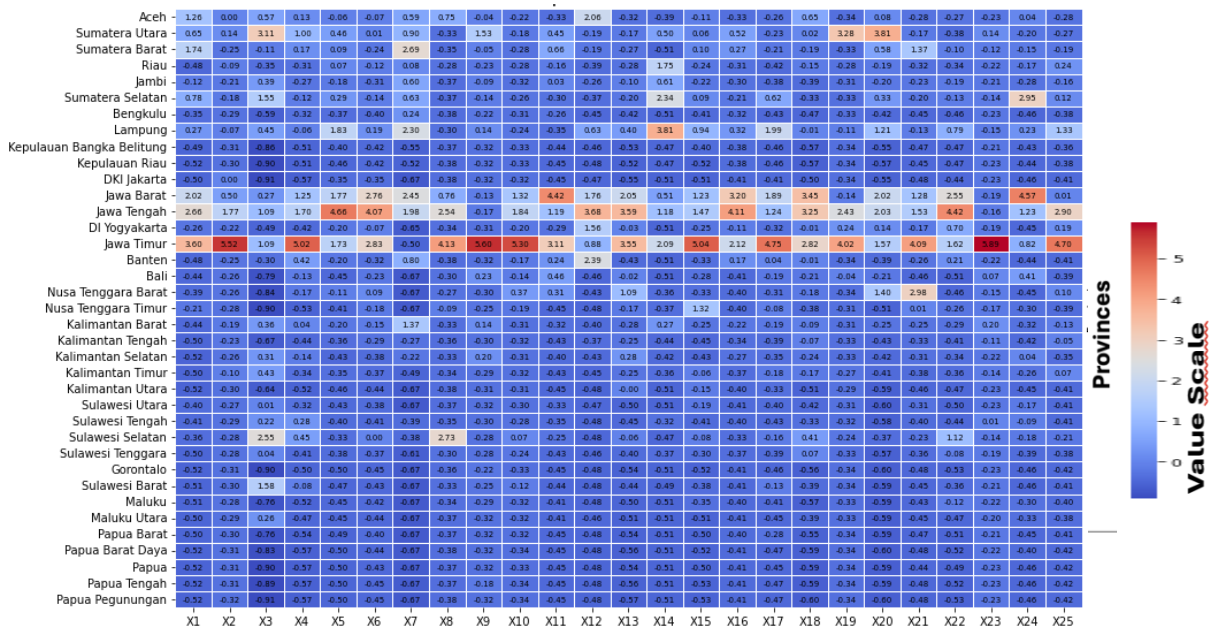

Figure 1. Fruit Data Matrix Scaling Heatmap

3.2. Optimal Bicluster of each selected threshold

After scaling, the following process is carried out by transforming the binary data matrix using each selected threshold. Binarization is an important preprocessing step before running the BCBimax algorithm for biclustering. Several thresholds will be used to see the effect of group results using the BCBimax algorithm. The BCBimax algorithm groups the data matrix values of 1 from the binarization process. Several thresholds are tried in the binarization process to find the optimal Bicluster. After transforming the binary data matrix, the BCBimax algorithm will be run to find several biclusters by considering the combination of row and column thresholds. The biclusters resulting from groups of the binary data matrix from each threshold are presented in the following heatmap:

3.2.1. Binarization Using the Median Threshold of Each Variable

The heatmap of the scaling data matrix in Figure 1 shows that several regions, such as the eastern part of Indonesia, have low vegetable and fruit production across all variables. So, the median of each variable will be used in this study. The median of each variable is chosen, recognizing that regions have different vegetable and fruit production potentials, so further analysis will be carried out accordingly. After binarizing each variable with the median threshold, the BCBimax algorithm captures several biclusters across all row and column thresholds. The resulting row and column clustering results are presented in Figure 2.

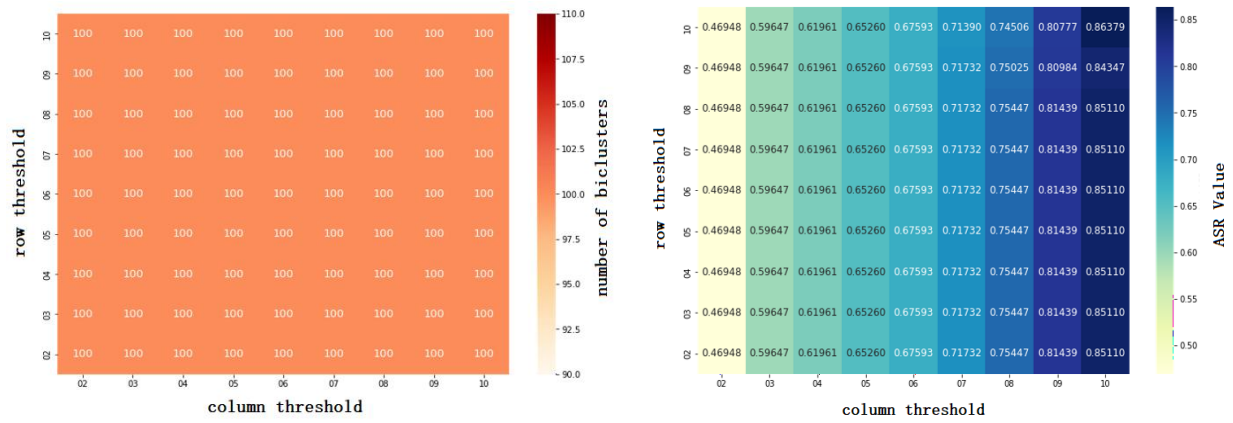

Figure 2. Number of biclusters (left) and ASR Value (right) from the Median Threshold of Each Variable

3.2.2. Binarization Using Global Data Median Threshold

The heatmap in Figure 2 shows that each variable's median yields several too-large biclusters (100 BC) for each threshold combination tested. This will undoubtedly yield inaccurate analysis results because many bicluster memberships will overlap. The presence of too many large biclusters will result in many areas with the same vegetable and fruit production in each Bicluster. This certainly does not yield good results because the concept of clustering to identify distinct biclusters within each Bicluster is not implemented. Further research uses the binarization process on the global data median matrix. The results of bicluster clustering using the BCBimax algorithm on the global data matrix are shown in Figure 3.

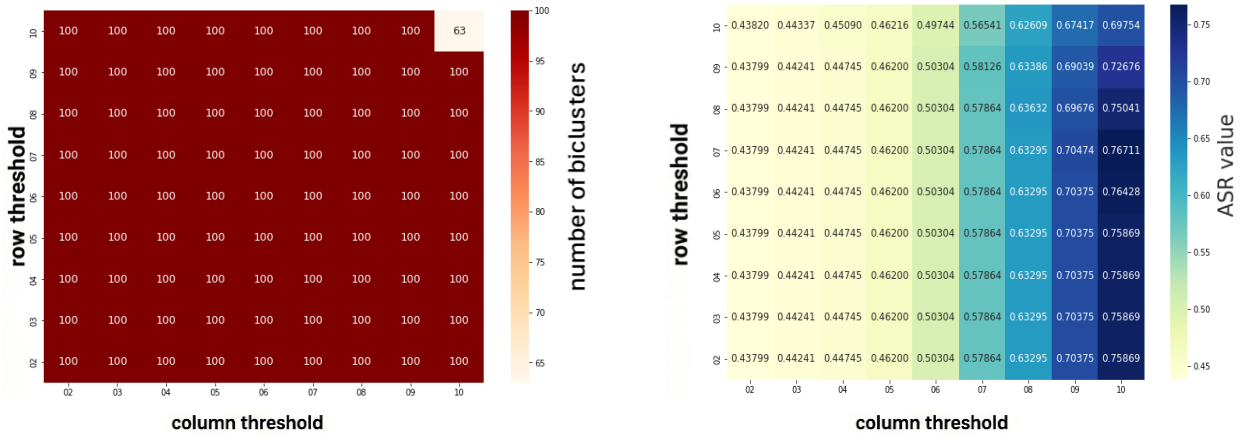

Figure 3. Number of biclusters (left) and ASR Value (right) Global Median Threshold

The heatmap in figure 3 shows that the BCBimax algorithm captures almost the same number of biclusters using the median of each variable. However, at the global median threshold, the BCBimax algorithm produces 63 biclusters at the threshold of row 2 and column 10. The number of biclusters is still relatively large. The results of this study also show that binarizing the median of each data set produces a more substantial ASR value than the global median, with the number of BCs produced

remaining almost the same, namely 100 BC. Using the median in grouping fruit and vegetable production is less appropriate. This is because the number of biclusters is too large, reaching 100. In addition, using the median with data characteristics that tend to have small production will cause all regions to be grouped in the same Bicluster. In Figure 1, all regions with low production of each vegetable and fruit (variable) are depicted in the blue heatmap. Almost all vegetable and fruit production heatmaps are dominated by the colour blue. So, in this case, the Mean can be further considered in grouping.

3.2.3. Binarization using the Mean Threshold of Each Variable

The next threshold is to use the mean of each variable after scaling and seeing the results of the median groups. The next consideration is to know the group's results in the threshold binarization process using the Mean of each variable. The Mean was chosen because the groups will be tighter, given the outliers in the fruit data matrix. Outliers are a significant concern in clustering. Outliers can bias the group's results because data containing them can render binarization ineffective. The production of vegetables and fruits in some areas is either too high or too low, which will cause areas near the threshold to be classified as 0. So that the number of objects and variables classified as one will be smaller. This will only cluster areas with high potential.

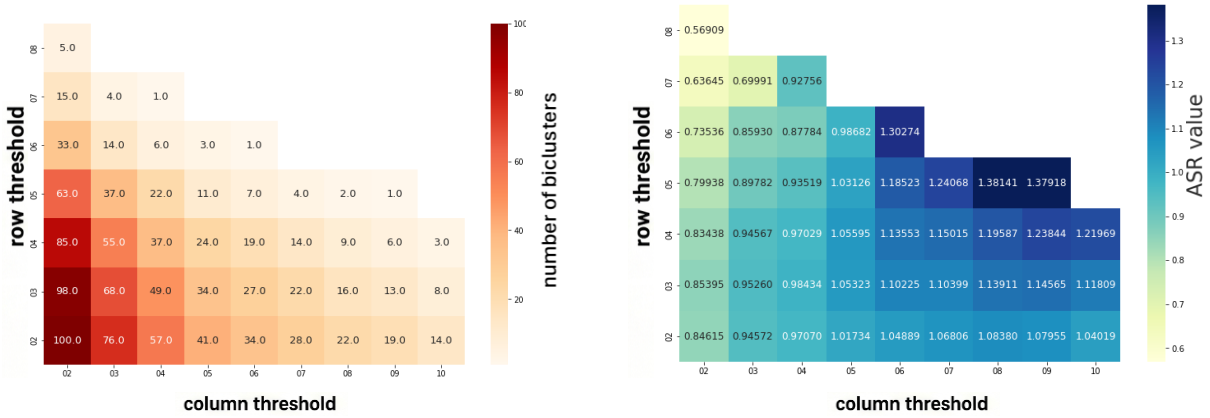

Figure 4. Number of biclusters (a) and ASR Value (b) Mean Threshold of Variables

However, grouping by the median can also be considered if the purpose of the grouping is only to identify production results that tend to be high. The results of BCBimax biclustering using the mean threshold of each variable will be presented in Figure 4. The heatmap in Figure 4 shows that the BCBimax algorithm produces a range of BCBimax values at the median threshold for each variable. The resulting ASR value is more significant than using the median. However, if further analysis is conducted, the number of biclusters produced by Mean will not be too large. The BCBimax algorithm produces five biclusters at row 8 and column 2 threshold as the optimal bicluster result with an ASR value of 0.56909. The number of biclusters is classified as informative because it is not too large to produce BC. The group's results will be stored to evaluate further and analyze the resulting bicluster membership.

3.2.4. Binarization using the mean threshold of all data

The next threshold is to use the Mean of all data from the data matrix. The Heatmap Matrix in Figure 5 shows that the BCBimax algorithm captures almost the same number of biclusters using the Mean of each variable. As with the Mean of each variable, the BCBimax algorithm also finds 5 BCs in the combination of the 8th row and 2nd column thresholds using the Mean of all data. This combination yields the optimal bicluster, with an ASR of 0.56909. Five biclusters are informative because the number of biclusters is not too large, even as the ASR value increases.

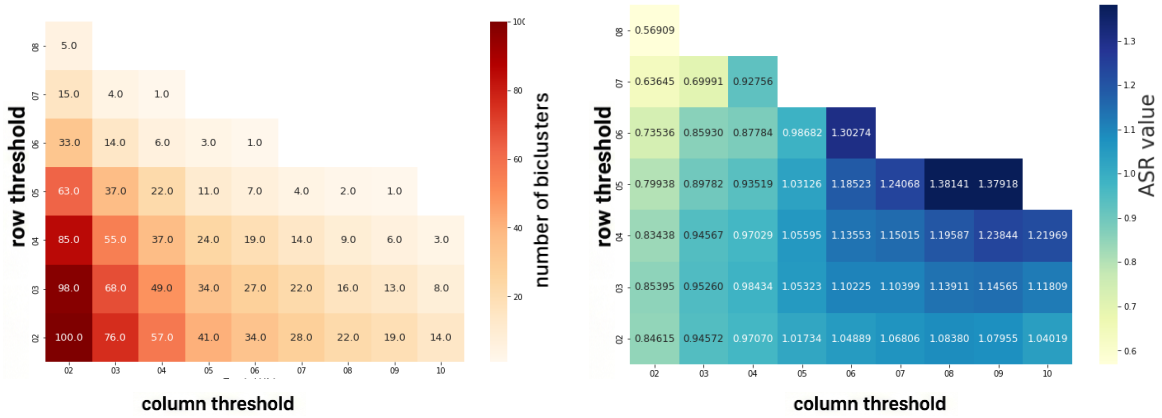

Figure 5. Number of biclusters (left) and ASR value (right) Mean threshold of all data

The larger ASR value can be due to the Bicluster's membership, which is grouped as having potential for high fruit and vegetable production. This can be caused by the influence of outlier data that can affect the resulting binary data matrix. The membership of the rows and columns produced by the five biclusters will be evaluated, and further analysis will be carried out, including the overlapping between the resulting BC.

3.2.5. Binarization using system thresholds

Figure 5 shows that the BCBimax algorithm, using the mean of each variable and the Mean of all data, performs quite well in clustering the potential of fruits and plants across several provinces in Indonesia. The next consideration is to see the binarisation process's group results using the system threshold. The system threshold is commonly used when there are no special considerations. The system threshold works by first determining the minimum and maximum values of the data described in equation (5). The results of grouping the binary data matrix using the system threshold are presented in the heatmap of Figure 6. The heatmap in Figure 6 shows that the BCBimax algorithm produces a relatively small number of BCBimax, ranging from 1 to 4. The ASR value using the system threshold is also smaller than the median or mean.

The BCBimax algorithm finds one Bicluster at the threshold of row 3 and column 2 and reports it as the optimal bicluster result, with an ASR value of 0.09533. The ASR value of 0.09533 at the system threshold is the smallest compared to the Median and Mean thresholds, so it is a good choice for grouping in the BCBimax algorithm. The combination of row 2 and column 3 thresholds will also be considered in selecting the optimal Bicluster from each threshold.

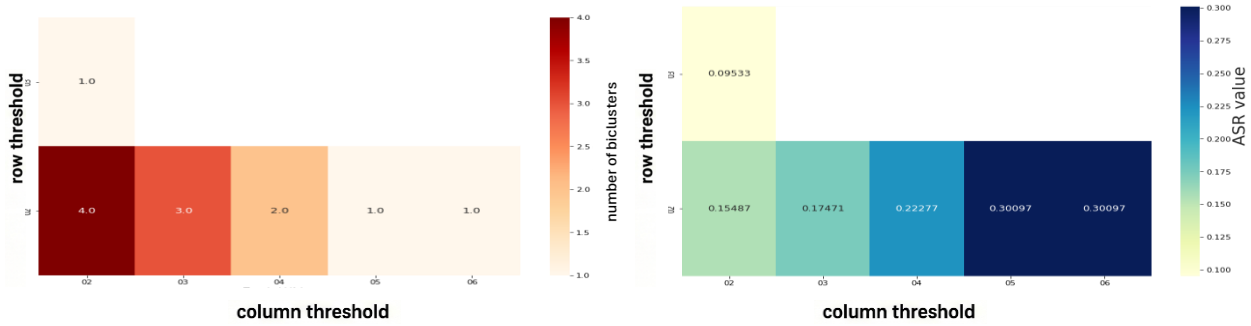

Figure 6. Number of biclusters (left) and ASR value (right) of the System threshold

4. DISCUSSION

After determining the optimal BC for each selected threshold and calculating the resulting ASR, further analysis will be conducted on the membership of the generated biclusters. In this advanced analysis, the membership derived from the median threshold will not be used due to the excessive number of biclusters produced (100 BCs). Therefore, for further analysis, the Mean of each variable, the Mean of the entire dataset, and the system threshold will be used. As previously explained, the Mean of each variable and the Mean of the entire dataset produce the same membership. The resulting biclusters contain many overlapping memberships. The membership results and the formed overlaps are presented in Table 2.

Table 2. Membership Table of BC for the Selected Threshold as Optimal BC

Province	X1	X20	X3	X9	X4	X14	X6	X18
Aceh	BC1	BC1	BC2, BC3	BC2	BC3	BC3		
Sumatera Utara	BC1	BC1	BC2,BC3, BC5	BC2,BC4	BC3, BC4	BC3, BC4,BC5		
Sumatera Barat	BC1	BC1						
Sumatera Selatan	BC1	BC1	BC2,BC5	BC2,BC4	BC4	BC3, BC4,BC5		
Lampung	BC1	BC1	BC2,BC5	BC2,BC4	BC4	BC3, BC4,BC5		
Jawa Barat	BC1	BC1	BC2,BC4,BC5	BC2,BC4	BC3, BC4	BC3, BC4,BC5	BC1	BC1
Jawa Tengah	BC1	BC1	BC2,BC4,BC5	BC2,BC4	BC3, BC4	BC3, BC4,BC5	BC1	BC1
Jawa Timur	BC1	BC1	BC2,BC4,BC5		BC3	BC3, BC4,BC5	BC1	BC1
Jambi			BC2,BC5	BC2,BC4	BC4	BC3, BC4,BC5		
Kalimantan Barat			BC2,BC4,BC5	BC2,BC4	BC3, BC4	BC3, BC4,BC5		
Sulawesi Tengah			BC3		BC3			
Riau				BC4		BC4		
Sulawesi Selatan			BC3		BC3			

Several overlapping regions are shown in Table 2, indicated by yellow labels. Table 2 shows that certain areas with specific variables are grouped into multiple BCs when using the mean thresholds of all variables and of the entire dataset. Meanwhile, the BCBimax algorithm only groups three regions with two variables into a single BC when using the system threshold. The system threshold, which captures only one BC, may result in less informative insights. This is evident as several other regions

with high potential could not be grouped using the system threshold [18][19]. Therefore, the Mean of each variable or the entire dataset is the basis for determining the optimal BC.

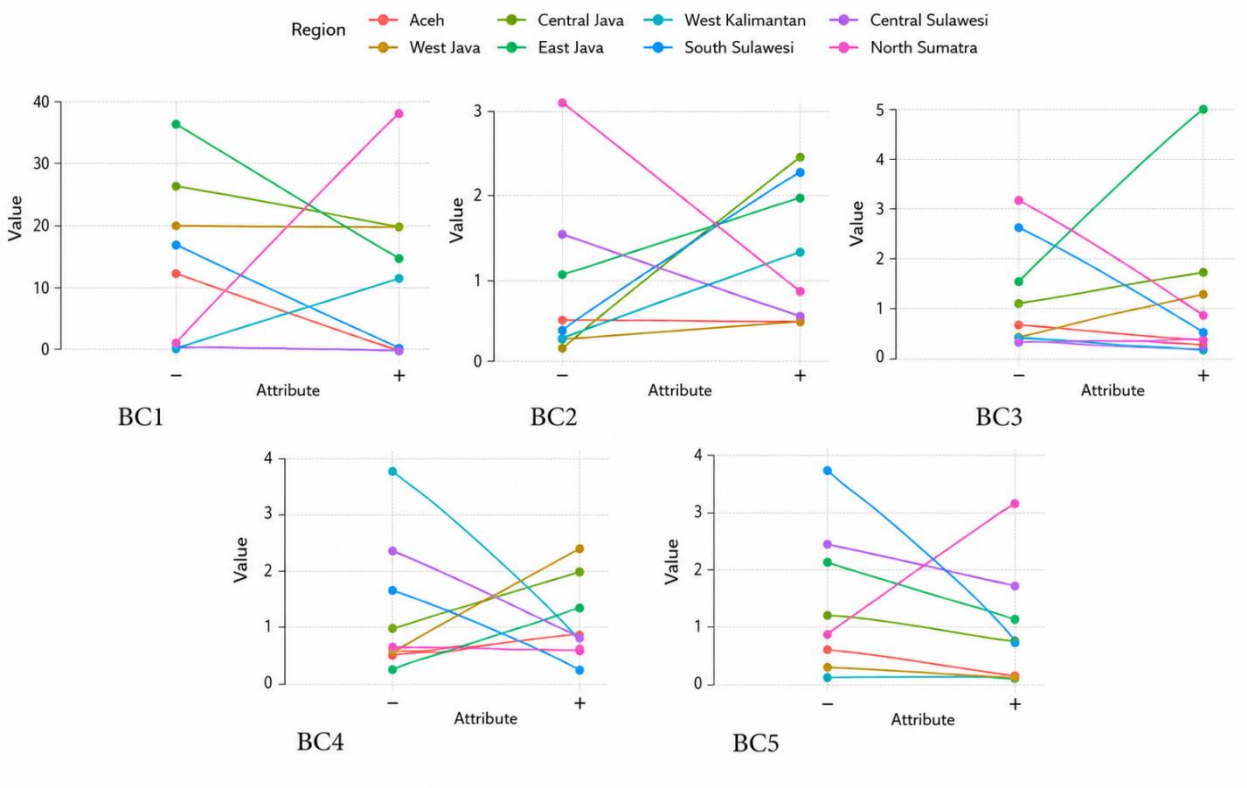

Figure 7. Profiling Plot Bicluster

Further analysis can be conducted by considering the profiling depicted in Figure 7. Profiling province subset membership against variable subsets can illustrate the level of homogeneity within the formed biclusters [20]. A bicluster becomes more homogeneous when its profiles coincide and align more closely. Figure 7 shows that the five optimal biclusters exhibit low homogeneity. The BCBimax algorithm identifies homogeneous membership in BC2, BC3, and BC5. This is evident from the profiling plots for several overlapping regions, such as Central Java and South Sulawesi, in BC2. BC5 also demonstrates relatively homogeneous bicluster membership, although North Sumatra and South Sulawesi exhibit negatively co-expressed traits (opposite trends). The BCBimax algorithm captures memberships with high homogeneity despite many regions overlapping.

The results of this study indicate that the threshold plays a crucial role in determining the optimal BC in the BCBimax algorithm. Additionally, data characteristics can be a key consideration when selecting the appropriate threshold. A previous study by Wulandari et al. (2023) on clustering fishery production found that the median was the best threshold. However, the heatmap reveals different results upon examining the data scaling applied. Fishery production tends to be moderate, aligning with the median of each variable. Moreover, outliers are distributed across all variables in the study.

In the 2023 fruit and vegetable production data, after scaling, production levels were generally low, with outliers observed only in three regions. Extremely low fruit and vegetable production results

in overly large BCs, as all variables in each region are set to 1. This causes many variables to be assigned the value 1. Consequently, the BCBimax algorithm captures excessively large BCs, up to 100 BCs. Having too many BCs results in significant overlap in memberships. Meanwhile, the system threshold only clusters one BC, capturing memberships with exceptionally high production potential. The complete comparison of evaluation values using ASR for each threshold is presented in Table 3.

Based on Table 3, the Mean of each variable or the entire dataset serves as the optimal BC result when considering thresholds. Therefore, in this study, the threshold's influence in the binarization process is crucial for clustering using the BCBimax algorithm. A heatmap can be a valuable reference when selecting a threshold, especially for datasets that contain outliers. Future research could examine the impact of outliers on the binarization process and threshold selection in the BCBimax algorithm.

Table 3. Results of the analysis of the influence of thresholds in finding optimal Bc on the BCBimax algorithm.

No	Binarization Threshold	Optimal Threshold Combination	ASR Value	Bicluster Count
1.	The median of each variable	Row 2 column 2	0.46948	100 BC
2.	Global data median	Row 10 column 10	0.69754	63 BC
3.	The Mean of each variable	Row 8, column 2	0.56909	5 BC
4.	Mean of all data	Row 8, column 2	0.56909	5 BC
5.	System Threshold	Row 3, column 2	0.09533	1BC

The findings of this study demonstrate that threshold selection substantially influences the structure, interpretability, and stability of biclusters generated by the BCBimax algorithm. The median-based thresholds tend to produce excessively large biclusters with high overlapping memberships, indicating that the binary transformation becomes less selective when the majority of data values are relatively low. This condition causes many observations to be converted into identical binary patterns, resulting in a loss of clustering specificity. Similar observations were reported in [9], which emphasized that binary pattern extraction strongly depends on the effectiveness of the binarization stage in preserving meaningful relationships within the data. In addition, Anthony and Ratsaby explained in [11] that inappropriate cutpoint selection may reduce the robustness and interpretability of binary classification results.

The Mean-based thresholds provide a more balanced biclustering structure because they reduce the dominance of low-valued observations and produce a more selective binary matrix. This result indicates that the Mean threshold is more sensitive to the presence of outliers and high-production regions, enabling the BCBimax algorithm to identify biclusters that are more informative and manageable in number. The resulting five biclusters with moderate ASR values suggest a trade-off between bicluster compactness and cluster diversity. This finding is consistent with the study in [6], which stated that biclustering quality should not only be evaluated based on coherence measures but also on the interpretability and redundancy of the resulting biclusters. Furthermore, Castanho et al. in [3] highlighted that meaningful biclustering results require a balance between homogeneity and structural diversity.

Another important finding is that the system threshold produces the smallest ASR value but only forms a single bicluster. Although a low ASR generally indicates stronger coherence within the bicluster, the resulting clustering structure becomes less informative because it captures only a limited subset of regions and variables. This suggests that relying solely on evaluation metrics without considering the number and diversity of biclusters may lead to suboptimal interpretations. Therefore,

this study contributes by demonstrating that optimal bicluster selection should simultaneously consider ASR values, bicluster quantity, overlap structure, and interpretability. These findings extend previous studies such as [7], which primarily focused on evaluation values without comprehensively analyzing the interaction between threshold characteristics and bicluster overlap patterns.

5. CONCLUSION

This study demonstrates that threshold selection plays a critical role in determining the quality, interpretability, and structure of biclusters generated by the BCBimax algorithm. The results indicate that different thresholding strategies produce substantially different biclustering characteristics, particularly in terms of ASR values, bicluster quantity, and overlap structure. Median-based thresholds tend to generate excessively large biclusters with high overlapping memberships, while the system threshold produces highly compact biclusters but with limited interpretability due to the formation of only one bicluster. Among the evaluated approaches, the Mean threshold of each variable and the Mean threshold of all data provide the most balanced biclustering results by generating informative biclusters with moderate overlap and stable ASR values. These findings suggest that threshold selection should not rely solely on evaluation metrics such as ASR, but also consider bicluster diversity and interpretability. Furthermore, this study highlights the importance of understanding data characteristics, particularly the presence of outliers and skewed distributions, before performing binarization in biclustering analysis. The novelty of this study lies in the comparative evaluation of multiple thresholding strategies in the BCBimax algorithm using ASR, bicluster overlap, and bicluster quantity as integrated evaluation criteria. Future studies may further investigate the influence of outliers, alternative normalization methods, and adaptive thresholding techniques to improve biclustering performance in high-dimensional agricultural and regional datasets.

REFERENCES

[1] N. Trianasari, I. M. Sumertajaya, Erfiani, and I. W. Mangku, "Application of beta mixture distribution in data on gpa proportion and course scores at the mbti telkom university," *Commun. Math. Biol. Neurosci.*, vol. 2021, 2021, doi: 10.28919/cmbn/5391.

[2] A. P. da Silva, R. Jude, and R. A. Gallardo, "Infectious Bronchitis Virus: A Comprehensive Multilocus Genomic Analysis to Compare DMV/1639 and QX Strains," *Viruses*, vol. 14, no. 9, 2022, doi: 10.3390/v14091998.

[3] E. N. Castanho, H. Aidos, and S. C. Madeira, "Biclustering fMRI time series: a comparative study," *BMC Bioinformatics*, vol. 23, no. 1, 2022, doi: 10.1186/s12859-022-04733-8.

[4] J. Lai *et al.*, "Factors associated with mental health outcomes among health care workers exposed to coronavirus disease 2019," *JAMA Netw. Open*, vol. 3, no. 3, p. e203976, 2020, doi: 10.1001/jamanetworkopen.2020.3976.

[5] S. C. Madeira and A. L. Oliveira, "Biclustering Algorithms for Biological Data Analysis: A Survey."

[6] 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.

[7] C. Wulandari, I. M. Sumertajaya, and M. N. Aidi, "Evaluation of Bicluster Analysis Results in Capture Fisheries Using the BCBimax Algorithm," *JUITA J. Inform.*, 2023, [Online]. Available:

https://api.semanticscholar.org/CorpusID:258593633
[8] M. Lejeune, V. Lozin, I. Lozina, A. Ragab, and S. Yacout, "Recent advances in the theory and practice of Logical Analysis of Data," *European Journal of Operational Research*, vol. 275, no. 1. Elsevier B.V., pp. 1–15, 2019. doi: 10.1016/j.ejor.2018.06.011.
[9] D. S. Rodriguez-Baena, A. J. Perez-Pulido, and J. S. Aguilar-Ruiz, "A biclustering algorithm for extracting bit-patterns from binary datasets," *Bioinformatics*, vol. 27, no. 19, pp. 2738–2745, 2011, doi: 10.1093/bioinformatics/btr464.
[10] E. Mayoraz and M. Moreira, "LNAI 1704 - Combinatorial Approach for Data Binarization."
[11] M. Anthony and J. Ratsaby, "Robust cutpoints in the logical analysis of numerical data," *Discret. Appl. Math.*, vol. 160, no. 4–5, pp. 355–364, 2012, doi: 10.1016/j.dam.2011.07.014.
[12] G. Kaur and K. Singh, "A comparative study of image segmentation using thresholding techniques," 2000.
[13] P. Guruprasad, "OVERVIEW OF DIFFERENT THRESHOLDING METHODS IN IMAGE PROCESSING."
[14] S. Jardim, J. António, and C. Mora, "Image thresholding approaches for medical image segmentation-short literature review," in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 1485–1492. doi: 10.1016/j.procs.2023.01.439.
[15] E. Acuña and C. Rodriguez, "The treatment of missing values and its effect in the classifier accuracy."
[16] M. Z. Rodriguez *et al.*, "Clustering Algorithms: A Comparative Approach," 2016, [Online]. Available: http://arxiv.org/abs/1612.08388
[17] 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.
[18] I. M. S. Sumertajaya, W. A. L. Ningsih, A. Saefuddin, and E. Rohaeti, "Biclustering Performance Evaluation of Cheng and Church Algorithm and Iterative Signature Algorithm," *JTAM (Jurnal Teor. dan Apl. Mat.*, vol. 7, no. 3, p. 643, 2023, doi: 10.31764/jtam.v7i3.14778.
[19] I. M. Sumertajaya, N. Hikmah, and F. M. Afendi, "Comparative Analysis of Bcbimax and Plaid Biclustering Algorithm for Pattern Recognition in Indonesia Food Security," *Barekeng*, vol. 20, no. 1, pp. 335–346, 2026, doi: 10.30598/barekengvol20iss1pp0335-0346.
[20] W. H. Organization, "Mental health at work: policy brief," World Health Organization, Geneva, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240053052.