Evaluation of Bicluster Analysis Results in Capture Fisheries Using the BCBimax Algorithm

by Cynthia Wulandari

Submission date: 02-Nov-2022 03:01PM (UTC+0700)
Submission ID: 1942257891
File name: JURNAL_PUBLIKASI_CYNTHIA_WULANDARI.pdf (780.82K)
Word count: 5512
Character count: 28844

Evaluation of Bicluster Analysis Results in Capture Fisheries Using the BCBimax Algorithm

Cynthia Wulandari[1], I Made Sumertajaya[2], Muhammad Nur Aidi[3]

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

[1]27cynthiawulandari@apps.ipb.ac.id, [2]imsjaya@apps.ipb.ac.id,

[3]muhammadai@apps.ipb.ac.id

Abstract – Biclustering is a simultaneous clustering technique by finding sub-matrixes that have the same similarity between rows and columns. One of the biclustering algorithms that is relatively fast and can be used as a reference for the comparison of several algorithms is the BCBimax algorithm. The BCBimax algorithm works by finding a sub-matrix containing element 1 of the formed binary data matrix. The selection of thresholds in the binarization process and the minimum combination of rows and columns are essential in finding the optimal bicluster. Capture fisheries have an important role in supporting sustainable growth in Indonesia, so information on the potential of fish species that have similarities in several provinces is needed in optimally mapping the potential. The BCBimax algorithm found 11 optimal biclusters in grouping capture fisheries data. The median of each variable is used as a threshold in the binarization process, and the minimum combination of row 2 and maximum column 2 is chosen to find the optimal bicluster result. The optimal average value of Mean Square Residual bicluster obtained is 0.405403 with the similarity of bicluster results (Liu and Wang index) which is different for each bicluster combination produced. All the bicluster results grouped the provinces and types of fish that had the same potential simultaneously.

Keywords: Bi-clustering; BCBimax Algorithm; Mean Square Residu; ASR; Liu and Wang index;

I. INTRODUCTION

Data can provide helpful information if it can be appropriately explored. The process of extracting and finding data patterns will affect the diversity of information obtained. The more varied information is expected to impact new insights that have not been explored before positively. Grouping can be a simple step in finding patterns in data, such as its application to multiple variable data. One of the analytical methods that can be used in overcoming the problem of multiple variable data is cluster analysis [1]. Cluster analysis is a method of grouping objects based on their similar characteristics [2]. The limitations of classical cluster analysis in classifying objects based on rows or columns only cause the diversity of information from other dimensions not to be appropriately clustered [3]. Rows and columns that have the same characteristic pattern are expected to be able to contribute more to providing helpful information. Therefore, a clustering technique was developed that handles the limitations of classical cluster analysis in conducting two-way clustering called Biclustering [4].

The concept of biclustering work was first introduced by [5]. However, [6] is the first to be applied to gene expression data. Biclustering works by looking for sub-matrixes and identifying rows and columns with similar characteristics [7]. This technique is done by simultaneously looking for a subset of rows with the same behavior along a subset of columns [8]. Currently, many algorithms have been successfully developed outside bioinformatics, such as the BCBimax algorithm, ISA, Plaid Models, OPSM, and various other algorithms [9]. Biclustering research on text mining, market data analysis [10], prediction and identification of abnormal energy consumption [11], to the latest research on water consumption pattern analysis [3] shows that the biclustering algorithm makes an exciting contribution to data clustering efforts.

However, the biclustering algorithm is not specific, and there are no precise rules in choosing the appropriate algorithm for specific criteria or data cases [10]. The advantages and success of the algorithm in previous studies in finding the optimal bicluster can be a good indicator in determining the algorithm to be used. For example, [12] states that the BCBimax algorithm is fast and precise in generating optimal biclusters and can be used as a reference for the comparison of several algorithms in various cases using several stages of evaluation.

Indonesia is a country that has the most significant and diverse potential for marine natural resources. One of the marine sectors that has an essential role in supporting sustainable growth in Indonesia is capture fisheries. Capture fisheries have a relatively severe complexity of problems, especially in mapping the potential, which is not considered adequate [13]. [14] revealed that the problem of sustainable capture fisheries in Indonesia is inseparable from several prominent issues, including rampant illegal, Unregulated and Unreported (IUU) fishing, symptoms of overfishing, weak or ineffective systems for utilizing fish resources, as well as potential data collection that has not been correctly entered. One of the efforts that can be made to determine the potential of capture fisheries is to simultaneously group fish that has the same potential in some provinces. The BCBIMAX algorithm, which is relatively fast and appropriate in finding a bicluster, is expected to group optimally captured fisheries data with the same potential fish species to all provisions in Indonesia.

The primary purpose of this study is to apply the BCBIMAX algorithm in grouping fish species according to 34 provinces in Indonesia. The results of the two-way grouping, based on fish and provincial types in Indonesia, are expected to provide additional information related to the effectiveness of policies that will be made in mapping the potential of capture fisheries in Indonesia. In addition, the application of BCBIMAX in the fisheries sector is also expected to be an additional new knowledge about Biclustering research outside the bioinformatics sector..

II. METHODS

A. *Data*

The data used in this study is secondary data obtained from the Ministry of Maritime Affairs and Fisheries (KKP) official website of the Republic of Indonesia in 2020. This unit of observation is 34 provinces in Indonesia, with 18 variables referring to the amount of production for each type of fish in Indonesia. The research unit is in the form of the production volume of fish species in Tons and has a numerical measurement scale. The data in this study were collected around September 2021 to capture fisheries classified as primary commodities and have the potential to be developed. The constituent elements of the data matrix in this study are presented in Table 1

TABLE I
CONSTITUENT ELEMENTS OF THE DATA MATRIX ON CAPTURE FISHERIES

Provinsi (Rows)	**Variable (Column)**
34 Provinces in Indonesia	Pomfret (C1), Skipjack (C2), Milk Shark (C3), Squid (C4), Octopus (C5), Snapper (C6), Grouper (C7), Kuwe (C8), Mackerel Scad (C9), Lobster (C10), Stingrays (C11), Rejungan (C12), Marlin (C13), Spanish mackerel (C14), Anchovy (C15), Mackarel tuna (C16), Tuna (C17), and Shrimp (C18)

B. *Research Stages*

1) *Pre-Processing:* The Bcbimax algorithm will be worked on data matrices, so ensuring the data is in matrix form is important. Next will be a scaling process and several stages of data exploration to see the initial characteristics of the data matrix, such as:
   a. Scaling creating a Heatmap from a scaling data matrix
   b. Making a PCA biplot which is presented in quadrants. PCA biplot is made to see an initial picture of the characteristics of each province on the variables used in capture fisheries.
2) *Algorithm BCbimax*
   The BCBimax algorithm divides several stages in finding a sub-matrix containing element one as the optimal bicluster result, such as:
   1. Suppose there is a binary data matrix $E_{R\times C}$. Divide the binary matrix randomly into two columns sets, **CU** and **CV.** Provided that the first row in column **CU** must contains element 1.
   2. Divide rows in the binary data matrix into three sets, namely:
      a. **RU** are rows containing element 1 in column **CU**
      b. **RW** are rows containing element 1 in columns **CU** dan **CV**
      c. **RV** are rows containing element 1 in column **CV**

The result of dividing the number of rows above will produce a data sub-matrix containing 0 elements, such as **GU** rows in the **CV** column set and **RV** rows in the CU column set. Then the submatrix containing element 0 will be deleted.

3. Two new sub-matrixes are constructed from step 2, namely the sub-matrixes **U=(RU** ∪ **RW**, **CU**) and **V=(RW** ∪ **RV**, **CU** ∪ **CV).** Step 2 will be repeated on the sub-matrix **U** and **V**, this process will continue until the minimum sub-matrix containing element 1 is found.
4. The submatrix containing element one will be stored as the bicluster result of the BCBimax algorithm. To avoid overlapping, the sub-matrix obtained as a result of biclustering will be stored and immediately deleted from the binary data matrix so that searching for a new bicluster can be resumed.
5. Processes 2 to 4 will be repeated until no sub-matrix containing element 1.

This study uses the biclust package referred to by [15] and has been provided by the software R in obtaining bicluster results.

3) Stages of Bicluster Analysis and Optimal Bicluster Selection

The BCBimax algorithm considers two essential stages in the search for the optimal sub-matrix (bicluster), including:

1. Binarization process

   The working concept of BCBimax is to find a submatrix containing element one so that the binarization process becomes a crucial initial stage in finding the optimal Bicluster. The binarization process is carried out with the condition that each data matrix element whose value exceeds the threshold will be set to a value of 1, and vice versa is 0. Choosing a threshold value that is too large will result in all data matrix elements being a value of 0 so that the number of biclusters found is less than optimal or not even found Bicluster. This study applies several experimental scenarios, especially in the selection of the threshold value in the binarization process, including:

   a. Set threshold value.
   b. Using the median of data if the threshold is unknown.
   c. Using the Threshold system.

   The threshold that produces a binary matrix with a non-dominant proportion of element values 0 will be selected in the optimal bicluster. The minimum combination of rows and columns that will be used will also affect the bicluster that will be formed.
2. The minimum combination of row and column thresholds is used.

   After the binarization process, the optimal bicluster can be searched by setting a minimum threshold combination of rows and columns to be tested. All combinations of these minimum thresholds will produce several biclusters to find the optimal bicluster. The dimensions of the bicluster formed will also not exceed the threshold that has been set.

The number of biclusters that are too large will cause overlapping. This will complicate the stages of interpretation of the results and the usefulness of the information obtained. Therefore, taking a sub-matrix from the combination matrix of the number of biclusters formed will be needed to determine the optimal bicluster. The selection of the optimal sub-matrix is based on the usefulness of the information needed in each research objective. The selected sub-matrix will be evaluated to find the optimal bicluster. The selected optimal cluster will be analyzed further regarding the characteristics of the bicluster results, such as the presence of overlapping and the membership of the formed bicluster.

4) *Bicluster Evaluation*

The optimal bicluster results can be determined by considering several aspects, such as the evaluation technique used and the usefulness of the information obtained from the grouping results. [16] used the Mean Square Residue (MSR) as an intra-bicluster function in evaluating the resulting bicluster. Mean Square Residue (MSR) is defined in the following equation:

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

where m_{IJ} is the average across biclusters, m_{Ij} the average in column j, m_{iJ} is the average in row i, $|I'|$ adalah is the dimension of the bicluster row, $|J'|$ is the bicluster column dimension. $E_{MSR}(I',J')$ hows the variation of the

interaction between rows and columns in the bicluster [16], [17]. Experimental combinations of min rows and max columns that produce more than one bicluster can be evaluated using the average residual (ASR) introduced by [18] to evaluate the total number of biclusters (n) generated [19], with the following equation:

$$ASR = \frac{1}{n}\left(\sum_{i=1}^{n} E_{MSR_i}(I', J')\right) \quad (2)$$

The quality of the bicluster will be better if the ASR value or the average MSR obtained is smaller or closer to 0. In addition to using ASR, the success of BCBimax in producing biclusters can be evaluated, too, using the Liu and Wang Index. The Liu and Wang index will see the similarity of all bicluster groups from each minimum combination of rows and columns. The Liu and wang index equation is written as follows [20].

$$I_{Liu\&Wang}(M_{opt}, M) = \frac{1}{K_{opt}} \sum_{i=1}^{K_{\text{opt}}} \max\left(\frac{[G_i \cap G_j] + [C_i \cap C_j]}{[G_i \cup G_j] + [C_i \cup C_j]}\right) \quad (3)$$

where M indicates all the resulting bicluster groups, M_{opt} ndica s the optimal bicluster of M elected based on the mean value of the residue divided by the smallest volume, K_{opt} is the number of biclusters in M_{opt}, $[G_i \cap G_j]$ is the number of rows (G) of the optimal bicluster (M_{opt}) which intersects the row in M, $[C_i \cap C_j]$ is the number of columns of the optimal bicluster (M_{opt}) that intersects with the column in M, $[G_i \cup G_j]$ is the number of rows combined in M_{opt} and M, $[C_i \cup C_j]$ a is the number of concatenated column (C) in M_{opt} and M.

III. RESULTS

A. Data Exploration

The first step in this research is to perform scaling on the data matrix that will be used. An initial description of the scale data matrix is presented in the heatmap diagram in Fig. 1. The heatmap matrix in Fig. 1 shows the volume of capture fisheries production in 2020 divided into three color categories, namely provinces with high fish production volume (red), medium (white), and low fish production volume (blue). The darkening color of the heat map indicates that the province has the highest production volume of fish species.

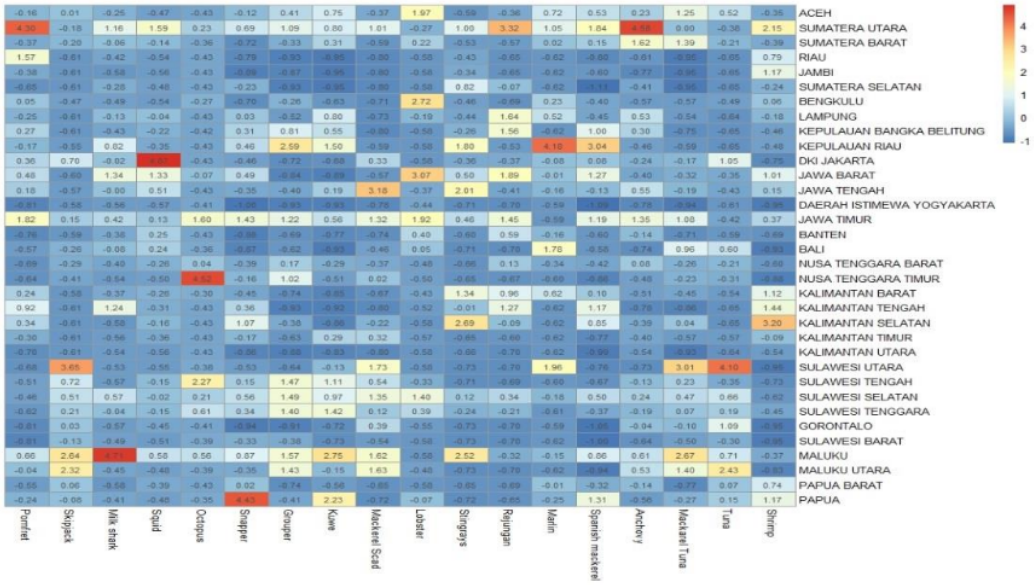


Fig. 1 Heatmap of Scaled Data Matrix

For example of the heatmap matrix in Fig. 1, DKI Jakarta produces the highest squid production volume (deep red color map) compared to several other provinces. As for the dark blue color, Special Region of Yogyakarta Province, mackerel fish has a low potential to be produced in the province. Overall, the heat map matrix above also illustrates that almost all provinces in Indonesia in 2020 had low to moderate production volumes of fish species, judging by the dominance of blue and white on the heat map. All variables used also have the potential to be produced in several provinces in Indonesia, such as (Pomfret, North Sumatra), (Skipjack, Maluku, and North Maluku) and others. However, not all provinces in Indonesia have the potential to produce this type of fish. Provincial information with the characteristics of certain fish species can provide an initial picture of the fish potential of the province concerned. The heat map diagram only shows an initial overview of the existing data matrix but cannot visualize the characteristics of fish species throughout the province.

The PCA biplot will display the diversity of the principal components of all data variables in a two-dimensional map. The variety of the data matrix presented in the PCA Biplot can be an initial picture of the potential for fish species based on the existing provinces. However, it is not a reference in comparing the results of the bicluster that will be formed. The PCA Biplot visualization is presented in Fig. 2

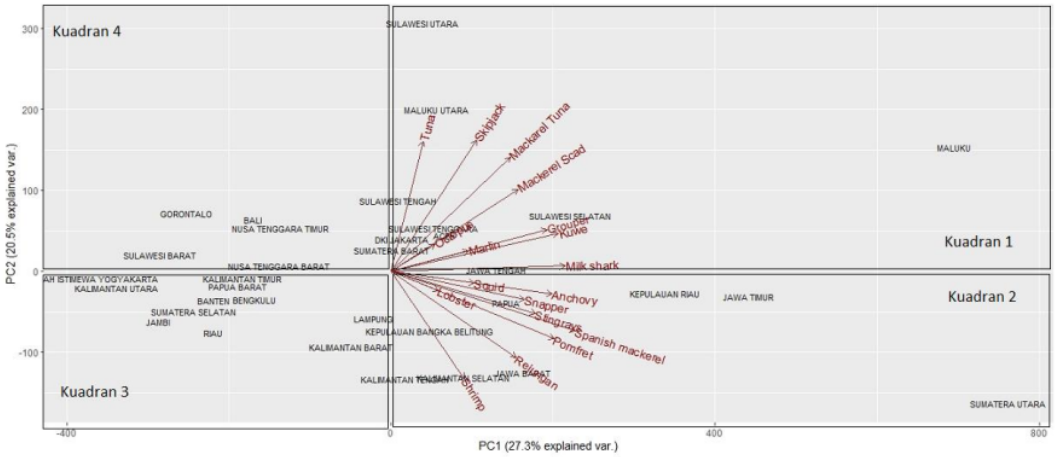


Fig. 2 PCA Biplot Results of Scale In Quadrant Data Matrix

The PCA biplot in Fig. 2 presents an initial description of the potential characteristics of fish species in several provinces into four quadrants. All provinces in quadrant one can be said to produce many types of fish, such as Tuna, Skipjack, Mackarel tuna, Mackerel Scad, Octopus, Marlin, Grouper, Kuwe, and milk shark. However, this type of fish has low potential to be produced in several provinces, including in quadrant 3. Fish species such as shrimp, rejungan, pomfret, mackerel, stingray, snapper, lobster, squid, anchovies, and lobster in quadrant 2 are also primarily produced in the province, which is related. On the other hand, the provinces included in quadrant 4 have a low potential to produce all types of fish in quadrant 2. The PCA biplot formed can only describe the data diversity of 47.8%, so several provinces with the highest production of fish species have not been well described in the biplot. This study uses the PCA biplot only as an initial picture to see the characteristics of each province that has potential for fish species.

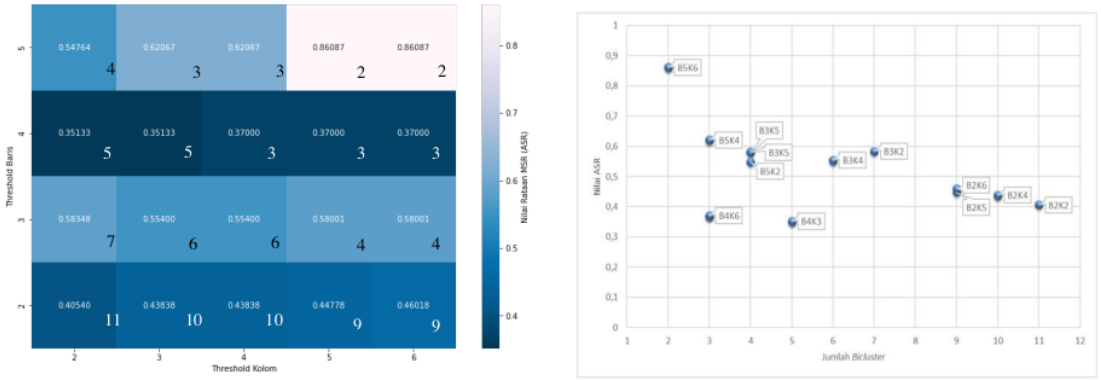
B. Biclustering Analysis Using The Bcbimax Algorithm

Finding the optimal bicluster using the BCBimax algorithm in this study, carried out with several experimental scenarios, especially in the binarization process and obtained the following results:

1) The Binarization Process Uses a System Threshold (1.878689)
   The threshold system produces a binary matrix with the proportion of the number of 0 values found is 0.93. The proportion of this value is quite large and impacts the absence of bicluster produced.

2) Data Scaling Matrix Median
   The experimental scenario using the median Scaling data matrix gives a threshold value that is too small so that all elements of the data matrix are set to a value of 1. This will, of course, make BCBimax only group the binary matrix into one bicluster. Choosing the median data matrix as the optimal bicluster is not considered adequate because the bicluster results will explain that all types of fish will have great potential to be produced in 34 provinces in Indonesia. On the other hand, this is contrary to the potential characteristics of groups of fish species described using the PCA Biplot.
3) The Median Of Each Variable
   The last experimental scenario that was carried out was to manually transform the data matrix using the median of each variable. Each type of fish certainly has various potentials, and choosing the median of each variable can be a solution for binarization efforts. Bicluster generated using the threshold of the median of each variable produces a variety of bicluster numbers for each combination of row and column thresholds used.

The BCBimax algorithm will look for several biclusters from the binary matrix that has been formed. This study used 100 minimum combinations of tested rows and columns, with experimental values ranging from 1 to 10 for each row and column. Trial combinations outside the range of these numbers did not produce any biclusters. One hundred combinations of minimum rows and columns that have been tried resulted in several biclusters that were less informative in answering the purpose of the biclustering analysis in this study. Therefore, the selection was carried out by selecting a combination of several biclusters (sub-matrix) that have been tried. The selected combination sub-matrix will be evaluated by calculating the average MSR (ASR) value in selecting the optimal bicluster. The distribution of the average MSR value (ASR) and the number of biclusters formed from the selected combination sub-matrix are presented in the form of two plots in Fig. 3.



(a) (b)

Fig. 3 (a) ASR Value Heatmap (b) Scatter Plot Number of Biclusters from Combination of Row and Column Thresholds

The optimal bicluster can be determined by looking at the smallest ASR value. The scatter plot shows that the smallest mean MSR (ASR) is in the minimum combination of row 4 and column 3 of (0.35133), with five biclusters formed. Further analysis of the membership characteristics of the five biclusters that have been carried out shows that these biclusters provide less informative information, especially in answering the purpose of biclustering in this study. The five biclusters formed could only cluster in 61% of the provinces. Some provinces considered to have potential (in the heatmap Fig. 2) for mapping are not clustered in the bicluster combination of min row 4 and minimum column 3. The easternmost regions of Indonesia, such as Papua, with the highest snapper production, are not clustered in the five biclusters. Suppose, seen from the scatterplot in Fig. 4(b), the ASR value of the minimum combination of row 2 and maximum column 2 results in a comparison of the ASR values that is not too far compared to the combination (4,3) with the 11 number of biclusters formed.

Further analysis was carried out to see the membership of the 11 biclusters formed. The percentage of provinces that can be grouped into 11 biclusters is around 73.5%. This is expected to be able to provide additional information about the province and the diversity of fish species that cannot be grouped in combination (4,3). The Liu and Wang

index will help further analyze the proportion of similarity of bicluster membership from all combination sub-matrixes that have been formed, as well as provide additional information in determining the optimal bicluster to be selected.

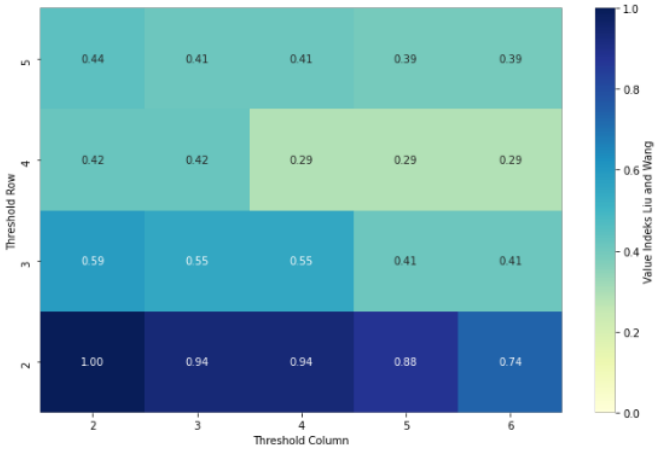


Fig. 4 Heatmap of Liu and Wang Index Values according to Combination of Row and Column Thresholds

The stages of intracluster evaluation in determining the optimal bicluster can be calculated using the Liu and Wang Index. The Liu and Wang index value close to 1 will indicate that the membership characteristics of all biclusters formed will be more similar. The Liu and Wang index presented in the heatmap Fig. 5 shows that all the biclusters formed have different member similarities. It can be seen from the proportion of similarity generated. The smallest ASR value with a minimum combination of row 4 and column 3 gives the proportion of bicluster members' similarity of 42% to the bicluster (2.2). It can be interpreted that the membership characteristics in the bicluster (4,3) have been described in the bicluster (2,2) of 42%. Therefore, considering the close comparison of ASR values, the characteristics of the bicluster membership (4.3), described in the bicluster (2.2), and the membership results are more informative in answering the objectives of the bicluster analysis in this study. The researcher determined the minimum combination of row 2 and maximum column 2 as the optimal bicluster result.

TABLE II
MEMBERSHIP CHARACTERISTICS OF OPTIMAL BICLUSTER RESULTS SELECTED BY BCBIMAX ALGORITHM

Bicluster	Bicluster size	Membership	
		Province (code)	Variable
1	5 × 11	Aceh, East Java, South Sulawesi, Southeast Sulawesi, North Sumatra	Skipjack, Milk shark, Snapper, Grouper Kuwe, Mackerel Scad, Lobster ,Rejungan, Spanish mackerel, Anchovy, Mackarel Tuna,
2	2 × 12	West Java, Maluku	Pomfret, Milk shark, Squid, Octopus, Snapper, Stingrays, Rejungan, Marlin, Spanish mackerel, Tuna, Shrimp
3	2 × 11	West Nusa Tenggara, Central Sulawesi	Skipjack, Squid, Octopus, Grouper, Kuwe, Mackerel Scad, Lobster, Anchovy, Mackarel Tuna, Tuna
4	2 × 11	Central Java, Lampung	Pomfret, Milk shark, Squid, Kuwe, Lobster, Stingrays, Rejungan, Marlin, Anchovy, Shrimp.
5	2 × 8	Bali, West Sumatra	Skipjack, Milk shark, Squid, Octopus, Lobster, Marlin, Mackarel Tuna, Tuna
6	2 × 7	DKI Jakarta, West Kalimantan	Pomfret, Milk shark, Squid, Stingrays, Rejungan, Marlin, Spanish mackerel.

Bicluster	Bicluster size	Membership	
		Province (code)	Variable
7	2 × 7	South Kalimantan, Bangka Belitung Islands.	Pomfret, Squid, Snapper, Grouper, Stingrays, Rejungan, Spanish mackerel.
8	2 × 6	Papua, West Papua	Skipjack, Snapper, Marlin, Spanish mackerel,Tuna, Shrimp
9	2 × 6	North Maluku, North Sulawesi	Skipjack, Octopus, Kuwe, Mackerel Scad, Mackarel Tuna, Tuna
10	2 × 5	Central Kalimantan, Riau Islands	Pomfret, Milk shark, Snapper, Stingrays, Spanish mackerel.
11	2 × 3	Gorontalo, East Nusa Tenggara	Mackerel Scad, Mackarel Tuna,Tuna

The membership characteristics of the optimal bicluster results selected using the BCBimax algorithm are presented in Table II. From the table, it is known that not all provinces are grouped into 11 biclusters formed; only about 73.5% of provinces are grouped into biclusters. The Bicluster formed shows that there is no overlap between provinces. Suppose further analysis is carried out through the heatmap diagram in Fig. 2, looking at the provinces with the highest production of fish species. In that case, the nine provinces that are not clustered are provinces that have low fish production. Provinces that have high production of fish species such as (North Sumatra, Anchovy), (Maluku, milk shark), (Dki Jakarta, Squid), (Papua, Snapper), (North Sulawesi, and Tuna) are well grouped into 11 biclusters formed. In general, the Bicluster in combination (2.2) grouped fish species that had high yields in the respective provinces. Only a few species of fish are not potentially clustered in the Bicluster. Provinces with high potential for fish species that are not clustered in the bicluster can be caused because there is no similarity between the rows and columns formed.

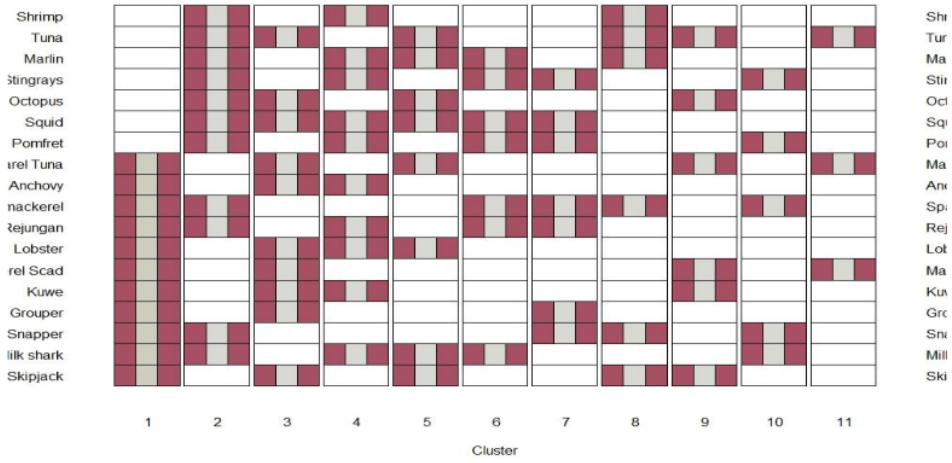


Fig. 5 Bicluster Membership Chart Based on Fish Type

Further analysis to find out the potential similarity of fish species in each bicluster will be more easily illustrated in the graph presented in Fig. 6. Each row in the graph will show the type of fish, and the column will represent the bicluster group. The red rectangle depicts the type of fish whose bicluster province groups have in common, so they are grouped into one bicluster. The average value for fish species in all provinces is depicted by a small rectangle with a greenish-grey scale. The brighter the green color, the more provinces produce this type of fish. For example, in Bicluster 1, as many as 14.70% of the provinces of Indonesia have in common as producers of Skipjack, milk shark, Snapper, Grouper, Kuwe, Mackerel Scad, Lobster, Rejungan, Mackerel, anchovies, and tuna fish species, seen from the lighter green color resulting. On the other hand, the other biclusters only have a similarity of about 5.88%, or only two provinces with the same fish potential. Several provinces also produce the same type of fish in several biclusters. For example, all provinces in bicluster 1, 3, 5, 8, and 9 produce the same type of fish, Skipjack,

and several other fish species shown in Fig. 6. The distribution map of the bicluster results can also provide a clearer picture of the distribution of provinces based on their bicluster groups.

Fig. 6 Distribution Map of Optimal Bicluster Results by Province

The optimal bicluster distribution map by province is presented in Fig. 7, showing that all biclusters formed are scattered throughout Indonesia. Eastern regions such as Papua and West Papua are identical with fish species included in bicluster 8. The distribution map associated with the bicluster membership graph in Fig. 6 can provide interesting information. First, all fish species in bicluster 1 (Skipjack, milk shark, Snapper, Grouper, Kuwe, Mackerel Scad, Lobster, Rejungan, mackerel, anchovies, and tuna) have the potential to be found in three regions of Indonesia, namely Sumatra, Java, and Sulawesi. The mapping of the potential of all fish species in Bicluster 1 for the Sumatran region can be carried out by two provinces, namely Aceh Province and North Sumatra. The two provinces can be an opportunity for other areas of Sumatra that still have low fish species production so that they can be mapped optimally. The case is different for the Java region; according to Bicluster 1, a province on the island of Java that has the potential for related fish species is in East Java. This can be additional information for several provinces on the island of Java so that the supply of related fish species, which are still relatively low, can be appropriately allocated by East Java. The potential for fish species in the Sulawesi region is dominated by all fish species in biclusters 1, 3, and 9. The fish species that dominate the three biclusters are skipjack, Kuwe, Mackerel Scad, and tuna. Heatmap Fig. 2 shows that the Sulawesi region has the most potential for producing octopus and skipjack tuna, and the BCBimax algorithm has successfully clustered it in biclusters 3 and 9. Second, the Papua region dominates the fish species in bicluster 8, such as skipjack, snapper, Marlin, Spanish mackerel, tuna, and Shrimp. Heatmap Fig. 2 also shows that Papua has the highest production volume of snapper and has been successfully clustered in bicluster 8. Third, Gorontalo and NTT produced the fewest groups of fish species compared to other biclusters.

Overall, it can be said that the BCBimax Algorithm classifies all areas that have the highest potential for fish species into 11 biclusters. The fish species with the highest production volume in heatmap Fig. 2 are almost clustered in the 11 biclusters, except for Octopus which is not clustered in NTT Province. This could be because Gorontalo Province does not have the potential to produce Octopus, so even though Octopus has the highest production volume in NTT, it cannot be clustered in the bicluster. BCBimax clusters around 73% or 25 provinces with the highest fish potential. Meanwhile, provinces such as Banten, South Sumatra, Kalimantan, Riau, Bengkulu, West Sulawesi, Special Region of Yogyakarta, North Kalimantan, and West Sulawesi were not grouped into the bicluster results because they have a low potential for fish species.

IV. CONCLUSION

The BCBimax algorithm determines 11 biclusters as the optimal bicluster results, with an AS[11] value of 0.405403. This study uses each variable's median threshold as the binarization stage in forming a binary data matrix. Bicluster with a minimum combination of rows and columns (2,2) is determined as the optimal bicluster result. All provinces with the highest potential for fish species are clustered well by BCBimax, which is spread over 11 biclusters. The results in this study provide conclusions, the evaluation stages used, and the usefulness of information obtained from the formed bicluster can be used as an indicator in choosing the optimal bicluster. In addition, the selection of thresholds in the binarization process needs to be considered. Inappropriate threshold values will result in a less than optimal bicluster to group cases..

REFERENCES

[1] A. M. Tamonob, A. Saefuddi, and A. H. Wigena, "Nonlinear Principal Component Analysis and Principal Component Analysis With Successive Interval in K-Means Cluster Analysis," *Forum Stat. Dan Komputasi*, vol. 20, no. 2, pp. 68–77, 2015.

[2] 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, pp. 1–12, 2021, doi: 10.28919/cmbn/5391.

[3] M. G. Silva, S. C. Madeira, and R. Henriques, "Water Consumption Pattern Analysis Using Biclustering: When, Why and How," *Water (Switzerland)*, vol. 14, no. 12, pp. 1–35, 2022, doi: 10.3390/w14121954.

[4] E. N. Castanho, H. Aidos, and S. C. Madeira, "Biclustering fMRI time series: a comparative study," *BMC Bioinformatics*, vol. 23, no. 1, pp. 1–30, 2022, doi: 10.1186/s12859-022-04733-8.

[5] 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.

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

[7] C. A. Putri, R. Irfani, and B. Sartono, "Recognizing poverty pattern in Central Java using Biclustering Analysis," *J. Phys. Conf. Ser.*, vol. 1863, no. 1, 2021, doi: 10.1088/1742-6596/1863/1/012068.

[8] V. A. Padilha and R. J. G. B. Campello, "A systematic comparative evaluation of biclustering techniques," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–25, 2017, doi: 10.1186/s12859-017-1487-1.

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

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

[11] F. Divina, F. A. G. Vela, and M. G. Torres, "Biclustering of smart building electric energy consumption data," *Appl. Sci.*, vol. 9, no. 2, 2019, doi: 10.3390/app9020222.

[12] 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.

[13] A. Fahrudin, S. H. Wisudo, and B. Juanda, "PERIKANAN TANGKAP DI INDONESIA : POTRET DAN TANTANGAN KEBERLANJUTANNYA Capture Fisheries in Indonesia : Portraits and Challenges of Sustainability," pp. 145–162, 2019.

[14] Bappenas, "Kajian Strategi Pengelolaan Perikanan Berkelanjutan," *Kementeri. PPN/Bapenas Direktorat Kelaut. dan Perikan.*, p. 120, 2014.

[15] S. Dolnicar, S. Kaiser, K. Lazarevski, and F. Leisch, "Biclustering: Overcoming data dimensionality problems in market segmentation," *J. Travel Res.*, vol. 51, no. 1, pp. 41–49, 2012, doi: 10.1177/0047287510394192.

[16] B. S. Biswal, A. Mohapatra, and S. Vipsita, "A review on biclustering of gene expression microarray data: Algorithms, effective measures and validations," *Int. J. Data Min. Bioinform.*, vol. 21, no. 3, pp. 230–268, 2018, doi: 10.1504/IJDMB.2018.097683.

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

[18] J. Yang, W. Wang, H. Wang, and P. Yu, "δ-clusters: Capturing subspace correlation in a large data set," *Proc. - Int. Conf. Data Eng.*, pp. 517–528, 2002, doi: 10.1109/icde.2002.994771.

[19] Y. Lee, J.-H. Lee, and C.-H. Jun, "Validation measures of bicluster solutions," *Ind. Eng. Manag. Syst.*, vol. 8, no. 2, pp. 101–108, 2009, [Online]. Available: http://kiie.org/iems/contents/vol8no2/8-2-04.pdf

[20] X. Liu and L. Wang, "Computing the maximum similarity bi-clusters of gene expression data," *Bioinformatics*, vol. 23, no. 1, pp. 50–56, 2007, doi: 10.1093/bioinformatics/btl560.

Evaluation of Bicluster Analysis Results in Capture Fisheries Using the BCBimax Algorithm

ORIGINALITY REPORT

3%	1%	2%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	scidok.sulb.uni-saarland.de Internet Source	<1%
2	Kusdiantoro Kusdiantoro, Achmad Fahrudin, Sugeng Hari Wisudo, Bambang Juanda. "PERIKANAN TANGKAP DI INDONESIA: POTRET DAN TANTANGAN KEBERLANJUTANNYA", Jurnal Sosial Ekonomi Kelautan dan Perikanan, 2019 Publication	<1%
3	Communications in Computer and Information Science, 2012. Publication	<1%
4	Ruizhi Wang, Duoqian Miao, Gang Li, Hongyun Zhang. "Rough Overlapping Biclustering of Gene Expression Data", 2007 IEEE 7th International Symposium on BioInformatics and BioEngineering, 2007 Publication	<1%
5	archive.org Internet Source	<1%

6	knepublishing.com Internet Source	<1 %
7	citeseerx.ist.psu.edu Internet Source	<1 %
8	Submitted to Padjadjaran University Student Paper	<1 %
9	www.ncbi.nlm.nih.gov Internet Source	<1 %
10	Rui Henriques, Sara C. Madeira. "Biclustering with Flexible Plaid Models to Unravel Interactions between Biological Processes", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015 Publication	<1 %
11	S.C. Madeira. "Biclustering algorithms for biological data analysis: a survey", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1/2004 Publication	<1 %
12	www.aporc.org Internet Source	<1 %
13	F. Divina. "Biclustering of Expression Data with Evolutionary Computation", IEEE Transactions on Knowledge and Data Engineering, 5/2006 Publication	<1 %

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		