



# Biclustering Performance of Iterative Signature Algorithm and Plaid Model after Imputation on Indonesian Macroeconomic Indicators

Yani Prihantini Hiola\*, I Made Sumertajaya, and Indahwati

*Statistics and Data Science Study Program, School of Data Science, Mathematics and Informatics,  
IPB University, Indonesia*

## Abstract

Biclustering is a two-way clustering method that identifies local patterns simultaneously across rows and columns of a data matrix. However, missing values may alter data structures and affect biclustering results. Studies evaluating the interaction between imputation methods and biclustering algorithms remain limited. This study evaluates the performance of the Iterative Signature Algorithm (ISA) and Plaid Model following imputation using Hot Deck, K-Nearest Neighbor (KNN), and Expectation Maximization (EM). The novelty of this study lies in assessing how the interaction between imputation methods and biclustering algorithms affects bicluster recovery and quality. Missing values were generated under MCAR at 5% and 10% proportions with 100 repetitions. Bicluster quality was evaluated using Mean Squared Residue (MSR), Transposed Virtual Error (VET), and Sub-Matrix Correlation Score (SCS), while bicluster consistency was assessed using the Jaccard Index (JI). ISA consistently achieved higher JI values, indicating better preservation of bicluster structures, whereas the Plaid Model produced lower MSR, VET, and SCS values, indicating more homogeneous biclusters. KNN generally showed the most consistent performance across scenarios. These findings suggest that imputation methods and biclustering algorithms should be selected jointly according to the analytical objective to obtain reliable biclustering results from incomplete macroeconomic data.

**Keywords:** Biclustering; Iterative Signature Algorithm; Jaccard Index; Mean Squared Residue; Missing Data imputation; Plaid Model.

Copyright © 2026 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (<https://creativecommons.org/licenses/by-sa/4.0>)

## 1. Introduction

Many datasets are unlabeled, making the identification of meaningful hidden patterns a major analytical challenge. Clustering is commonly used to address this problem by grouping objects based on similarity [1]. However, conventional clustering only groups objects and may fail to capture local patterns involving subsets of rows and columns [2]. To overcome this limitation, biclustering was developed as a two-way clustering approach that simultaneously groups rows and columns to identify local patterns [3]. Because no universal guideline exists for selecting the most appropriate biclustering algorithm for a particular dataset [4], an effective algorithm should be flexible, robust to noise and missing data, and capable of producing statistically valid

---

\*Corresponding author. E-mail: [yanihiola@apps.ipb.ac.id](mailto:yanihiola@apps.ipb.ac.id)

biclusters [5]. This study focuses on two biclustering algorithms, namely the Iterative Signature Algorithm (ISA) and the Plaid Model. These algorithms were selected based on previous studies conducted in Indonesian regional contexts, where ISA was evaluated against BCBimax for capture fisheries data [6], while the Plaid Model was compared with BCBimax for identifying regional food-security patterns [7]. Therefore, this study compares ISA and the Plaid Model to evaluate their performance in macroeconomic indicator data.

Biclustering cannot be directly applied to data containing missing values. Biclustering algorithms, including ISA and the Plaid model, assume a complete matrix in the bicluster search [8]. Based on this, it is necessary to estimate the missing data first so that the biclustering process can proceed and does not hinder valid interpretation of the data structure. According to [9], one way to handle missing data is by ignoring the missing values so that only the available values are analyzed [10]. However, this method can reduce the sample size, making the analysis less efficient [11]. Therefore, in this study, missing data are handled using imputation techniques, namely estimating the missing values using estimated values obtained based on the available information.

In this study, the imputation methods used are Hot Deck, K-Nearest Neighbor (KNN), and Expectation Maximization (EM). Previous research conducted by [12] concluded that the best imputation accuracy was obtained from the KNN method compared to Hot Deck under a 5% missing data proportion. Another study conducted by [13] found that the Hot Deck imputation method outperformed the KNN and Predictive Mean Matching (PMM) methods under missing data proportions ranging from 5% to 50%. In addition, there is another model-based approach, namely the EM method, which is very helpful in designing robust identification algorithms for handling missing data [14]. Based on the study conducted by [15], EM shows the best performance in performing imputation compared to Multiple Imputation, Mean, and KNN methods on datasets with missing data proportions ranging from 5% to 45%. These findings indicate that the performance of imputation methods may vary depending on the proportion of missing data and the characteristics of the dataset used in the study. Therefore, the imputation methods used in this study are Hot Deck, K-Nearest Neighbor, and Expectation Maximization.

In the imputation process, the Hot Deck, K-Nearest Neighbor (KNN), and Expectation Maximization (EM) methods produce missing data estimates with different characteristics. Hot Deck is a donor-based imputation method, KNN preserves local similarity among observations, whereas EM employs a probabilistic approach through distribution parameter estimation. These differences in working principles may produce distinct imputed data patterns and potentially influence the local structure of the data.

ISA and Plaid Model are biclustering algorithms that have different characteristics and mechanisms for bicluster formation. ISA is an iterative biclustering algorithm designed to identify submatrices with similar patterns [16], whereas Plaid Model is based on an additive statistical model (layer model) that models data as a combination of several biclusters [17]. Given the differences in how these two algorithms operate, there are indications that each algorithm exhibits varying sensitivity in identifying biclusters after the imputation process, which in turn affects the consistency of the identified bicluster structure when compared to the bicluster results obtained from the complete dataset (the ideal scenario).

However, previous studies have generally focused only on comparing imputation methods or biclustering algorithms separately. Research examining the influence of imputation methods on the performance of multiple biclustering algorithms remains limited. In fact, biclustering is highly dependent on local patterns and relationships between rows and columns, so changes in data structure caused by the imputation process may affect the quality of the resulting biclusters.

Therefore, this study offers novelty by comparing the performance of the ISA and Plaid Model biclustering algorithms after the imputation process using Hot Deck, KNN, and EM methods. The evaluation is conducted to determine the robustness of biclustering algorithms in preserving bicluster structures in data containing missing values.

Based on these considerations, this study addresses two research questions: how are Indonesian provinces grouped based on macroeconomic indicators using the Iterative Signature Algorithm (ISA) and Plaid Model, and how do these biclustering algorithms perform after imputation using Hot Deck, K-Nearest Neighbor (KNN), and Expectation Maximization (EM) methods?

In evaluating the performance of biclustering, data on Indonesia’s macroeconomic development indicators are used as a case study. The selection of this case study is motivated by the characteristics of macroeconomic indicator data, which are often complex, heterogeneous, and characterized by localized patterns across provinces. Accordingly, biclustering analysis was applied to identify localized structures that may not be captured by classical clustering methods, as such methods generally focus on global patterns. Data from Badan Pusat Statistik (2025) [18] show that Indonesia, as an archipelagic country, has uneven economic growth across regions. BPS recorded that Indonesia’s economic growth in 2024 reached 5.03%, slightly slowing down compared to 2023 at 5.05%. The contribution to economic development is also dominated by the group of provinces on Java Island. This is reflected in the National Gross Domestic Product value recorded at 57.02%, meaning that more than half of the total national GDP is contributed by Java Island. In identifying and understanding the interrelationships between regions in the context of economic development, the biclustering method is used to capture local patterns between provinces and macroeconomic indicators, so that the results of this analysis can be used as a basis for formulating more targeted and equitable development policies.

Based on this background, this study evaluates the performance of ISA and the Plaid Model after imputation using Hot Deck, KNN, and EM. Bicluster quality was assessed using MSR, VEt, and SCS, whereas bicluster consistency was evaluated using the Jaccard Index at global, bicluster, and member levels.

## 2. Methods

In this section, information regarding the research data and several methods used in the study is presented. The methods described begin from the stages prior to the biclustering process up to the final evaluation of the role of imputation in biclustering.

### 2.1. Data

The data used in this study are secondary data, where each row represents a province and each column represents a macroeconomic indicator for the year 2024 obtained from the official website of [Badan Pusat Statistik \(BPS\)](#). The observational units used in this study consist of 38 provinces in Indonesia, with a total of 26 variables. Further explanations can be found in Appendix 7. [Table 1](#) presents the grouping of variables based on their substantive characteristics by referring to the economic meaning of each variable.

**Table 1:** Grouping of Variables Based on Economic Indicators

Group	Variables	Total
Economic Production and Growth	$X_1, X_3, X_{21}, X_{22}$	4
Tourism and Service Sector	$X_2, X_{15}, X_{16}$	3
Labor and Income	$X_4, X_8, X_{24}$	3
Agricultural Resilience and Sector	$X_5$	1
Human Development/Welfare	$X_6, X_7, X_9, X_{23}$	4
Basic Quality of Life	$X_{10}, X_{25}, X_{26}$	3
Regional Fiscal Structure	$X_{11}, X_{12}, X_{13}, X_{14}$	4
GRDP Expenditure Structure	$X_{17}, X_{18}, X_{19}, X_{20}$	4
Total		26

All variables were continuous macroeconomic indicators and were standardized using z-score before analysis. No additional preprocessing was performed.

## 2.2. Hot Deck

The Hot-Deck method imputes missing values using observed values from donor observations that are most similar to incomplete observations (recipients) [19, 20]. In this study, donors were selected from complete cases and similarity was measured using Euclidean distance [21]:

$$d_{ij} = \sqrt{\sum_{p=1}^q (x_{ip} - x_{jp})^2}$$

where  $d_{ij}$  denotes the Euclidean distance between observations  $i$  and  $j$ . Missing values were replaced using the nearest donor observation. Euclidean distance was selected because it performs well for continuous numerical variables and is computationally efficient [22]. However, Euclidean distance may become less effective in high-dimensional data because distances among observations tend to become more similar. Nevertheless, the number of variables used in this study is still considered manageable, allowing Euclidean distance to remain appropriate for capturing similarity among provinces. Alternative distance measures such as Manhattan distance or Mahalanobis distance may be considered in future studies.

## 2.3. K-Nearest Neighbor

The K-Nearest Neighbor (KNN) method imputes missing values using information from the (K) most similar observations [23, 24]. Similarity was measured using Euclidean distance, and missing values were estimated using weighted mean estimation (WME):

$$x_J = \frac{\sum_{k=1}^K w_k v_k}{\sum_{k=1}^K w_k}$$

where  $x_J$  is the estimated value,  $v_k$  is the observed value from the  $k$ -th nearest neighbor,  $w_k = \frac{1}{(d_{ij})^2}$  is the corresponding weight, and  $k$  denotes the number of neighbors. Because the choice of  $k$  influences imputation accuracy, values from 1 to 10 were evaluated using RMSE and MAE [25, 26].

## 2.4. Expectation Maximization

The Expectation Maximization (EM) algorithm is an iterative maximum likelihood approach for handling incomplete data [27]. The algorithm alternates between an Expectation step (E-step), which estimates missing values based on the observed data, and a Maximization step (M-step), which updates parameter estimates. This process is repeated until convergence is achieved.

EM assumes a multivariate normal distribution and iteratively estimates missing values using the mean vector and covariance matrix of the observed data. Although macroeconomic indicators may not fully satisfy this assumption, EM remains widely used because it provides stable parameter estimation and effectively handles incomplete data. Previous studies have shown that EM frequently produces accurate imputations even when normality is not fully satisfied [28].

## 2.5. Iterative Signature Algorithm

The Iterative Signature Algorithm (ISA) is a biclustering algorithm that identifies subsets of rows and columns that exhibit coherent local patterns through an iterative updating process [16, 29, 30]. ISA begins with an initial random seed and iteratively updates row and column memberships based on threshold values until convergence is achieved.

The ISA algorithm updates row and column memberships iteratively based on threshold values. A column is selected as a member of the bicluster if its score satisfies the following condition:

$$a_{xj}^C > t_C \sigma_C$$

and

$$\frac{1}{|J|} \sum_{u=1}^{|J|} a_{xy}^C > \frac{t_C}{\sqrt{|J|}}$$

Similarly, a row is selected as a member of the bicluster if:

$$a_{Iy}^G > t_G \sigma_G$$

and

$$\frac{1}{|I|} \sum_{u=1}^{|I|} a_{xy}^G > \frac{t_G}{\sqrt{|I|}}$$

where  $t_C$  and  $t_G$  represent the column and row threshold values, respectively,  $\sigma_C$  and  $\sigma_G$  denote the standard deviations of column and row scores, while  $a_{xy}^C$  and  $a_{xy}^G$  represent the elements of the normalized column and row matrices.

In the bicluster search process using ISA, the first step was to determine the optimal row and column threshold values. In this study, the optimal thresholds were determined through a grid search procedure using threshold values ranging from 0.10 to 0.85 with intervals of 0.05 and 100 initial seeds. A grid search procedure was employed because the row and column thresholds directly influence bicluster size, membership composition, and coherence, making threshold selection a critical factor in ISA performance [29, 30]. The optimal threshold combination was selected based on MSR, VEt, SCS, the number of biclusters formed, and the resulting row-column membership structure to ensure coherent and interpretable biclusters.

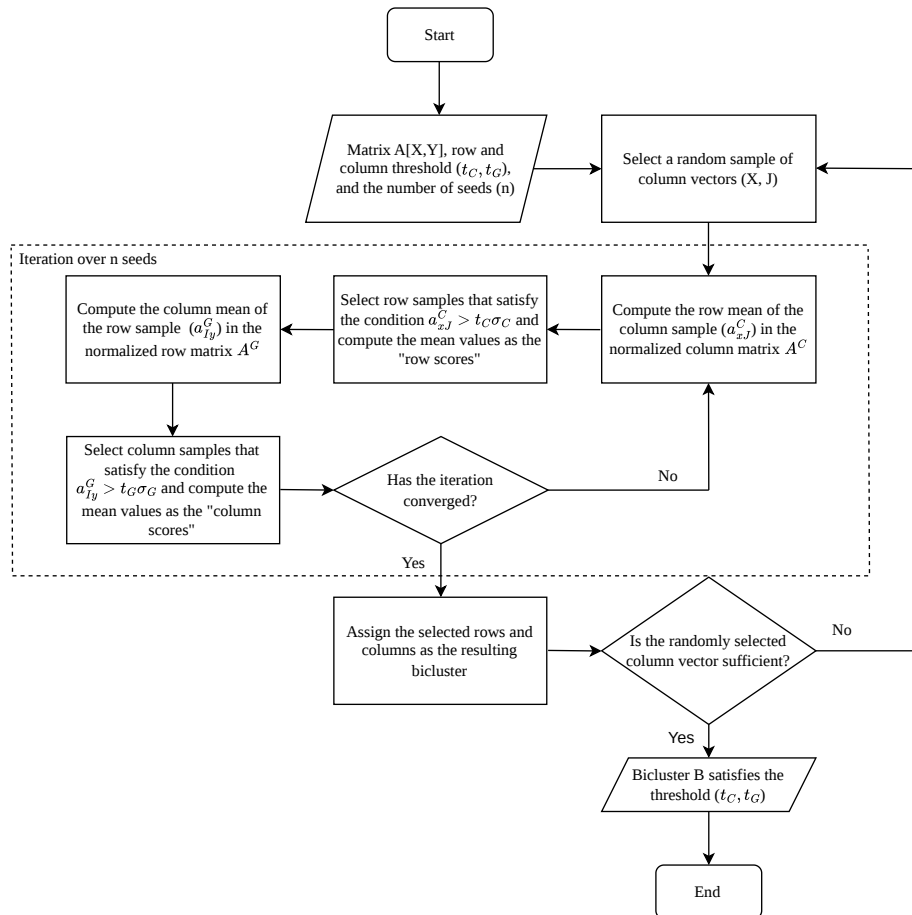


Fig. 1: Iterative Signature Algorithm Flowchart (modified from [31])

### 2.6. Plaid Model

The Plaid Model is a model-based biclustering algorithm that divides the data into several submatrices in the form of  $k$ -layers, where each layer consists of subsets of rows and subsets of columns [17, 32]. In this context, a layer represents a bicluster, while the first layer functions as the background layer representing the overall structure of the dataset. Each bicluster contributes additively to the model, allowing the Plaid Model to adapt flexibly to local patterns within the data.

Mathematically, the Plaid Model can be expressed as follows:

$$Y_{ij} = \mu_0 + \sum_{k=1}^K \theta_{ijk} \rho_{ik} \kappa_{jk} + \varepsilon_{ij}$$

where  $Y_{ij}$  represents the observed data value at the  $i$ -th row and  $j$ -th column,  $\mu_0$  is the overall background effect,  $\theta_{ijk}$  is the effect of the  $k$ -th layer,  $\rho_{ik}$  and  $\kappa_{jk}$  are row and column membership indicators for the  $k$ -th bicluster, respectively, and  $\varepsilon_{ij}$  represents the residual error.

In this study, the Plaid Model was developed using four model specifications, namely constant, row constant, column constant, and two-way models. Multiple model specifications were evaluated because different Plaid formulations capture different structural effects within the data, allowing the most appropriate representation of bicluster patterns to be identified [32]. To improve the quality of the estimation results, a back-fitting procedure with two iterations was applied to refine parameter estimation, while a shuffle procedure with five random permutations was used as a significance test to ensure that the identified layers were not formed purely due to random noise. The biclustering process was restricted to a maximum of 10 layers for each model specification.

In addition, the Plaid Model used release parameters that functioned as thresholds for evaluating the contribution of elements to residual reduction. Elements with contributions below the specified release values were removed to maintain the information density within biclusters. The row release and column release parameters were optimized using the same grid search range as applied in the ISA algorithm. Release parameters were optimized because they determine the inclusion of rows and columns in biclusters and therefore directly affect bicluster density, coherence, and interpretability [32]. Parameter tuning was performed by first selecting the optimal release parameters within each model specification and subsequently choosing the best-performing specification based on bicluster quality, bicluster size, and membership structure.

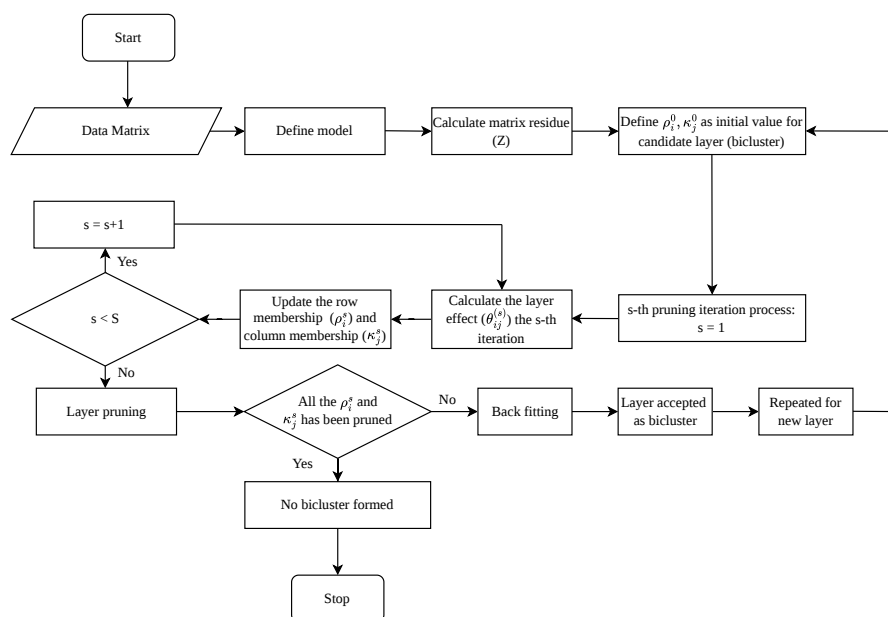


Fig. 2: Plaid Model Flowchart (modified from [33])

## 2.7. Biclustering Evaluation

Biclustering performance was evaluated using internal and external indices [5]. Internal evaluation assessed bicluster quality using MSR, VEt, and SCS [34], whereas external evaluation assessed bicluster consistency using the Jaccard Index (JI). MSR measures bicluster homogeneity, VEt evaluates the preservation of shifting and scaling patterns, and SCS assesses correlation structures within biclusters. Lower values indicate better bicluster quality [34]. Because MSR is more sensitive to shifting patterns, VEt and SCS were additionally employed to provide a more comprehensive evaluation.

The MSR is defined as:

$$MSR_{I,J} = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} (r_{ij})^2$$

with

$$r_{ij} = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}$$

where  $a_{iJ}$ ,  $a_{Ij}$ , and  $a_{IJ}$  denote the row mean, column mean, and overall bicluster mean, respectively.

The VEt is defined as:

$$VEt(\mathcal{B}) = \frac{1}{|I||J|} \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} |\hat{b}_{*ij} - \hat{r}_{*i}|$$

where  $\hat{b}_{ij}$  and  $\hat{r}_i$  denote the standardized bicluster value and virtual condition value, respectively. The SCS is defined as:

$$S(\mathcal{B}) = \min(S_{row}(I, J), S_{col}(I, J))$$

where  $S_{row}$  and  $S_{col}$  are calculated using Pearson correlation among rows and columns within the bicluster [34].

The optimal parameter combination for both ISA and the Plaid Model was selected based on bicluster quality evaluated using MSR, VEt, and SCS, while also considering the resulting bicluster membership structure and the number of biclusters formed to ensure coherent and interpretable biclusters.

External evaluation was performed using the Jaccard Index (JI), which measures the similarity and consistency of biclustering structures between complete and imputed data. In biclustering applications, the Jaccard Index is commonly used to quantify the overlap of row and column memberships between two biclusters [17]. The adaptation performed in this study does not modify the original Jaccard similarity principle proposed by [35], but extends its application to evaluate biclustering similarity at global, bicluster, and membership levels to assess both structural similarity and membership consistency between biclustering results.

At the global level, JI measures the similarity of the overall bicluster structures between complete and imputed data and is defined as follows:

$$JI_{Global} = \frac{|S_A \cap S_B|}{|S_A \cup S_B|}$$

where  $S_A$  and  $S_B$  represent the sets of all bicluster cells obtained from complete data and imputed data, respectively.

At the bicluster level, JI evaluates the similarity between individual biclusters and is defined as follows:

$$JI(Bicres_1, Bicres_2) = \frac{1}{g} \sum_{i=1}^g \sum_{j=1}^t \left( \frac{|BC_i(Bicres_1) \cap BC_j(Bicres_2)|}{|BC_i(Bicres_1) \cup BC_j(Bicres_2)|} \right)$$

where  $g$  and  $t$  denote the number of biclusters in the first and second biclustering results, respectively, while  $BC_i$  and  $BC_j$  represent individual biclusters from each biclustering result.

At the member level,  $JJ$  evaluates the consistency of row and column memberships between complete and imputed data using the following equations:

$$JJ_{Row} = \frac{|R_A \cap R_B|}{|R_A \cup R_B|}$$

$$JJ_{Column} = \frac{|C_A \cap C_B|}{|C_A \cup C_B|}$$

$$JJ_{Member} = \frac{JJ_{Row} + JJ_{Column}}{2}$$

where  $R_A$  and  $R_B$  denote the row memberships, while  $C_A$  and  $C_B$  denote the column memberships of biclusters obtained from complete and imputed data, respectively.

## 2.8. Simulation Design

A simulation study was conducted under the Missing Completely at Random (MCAR) mechanism with missing-data proportions of 5% and 10%. Each scenario was repeated 100 times, generating 200 simulated datasets, which were subsequently imputed using Hot Deck, KNN, and EM before being analyzed with the ISA and Plaid biclustering algorithms.

In the 10% missing-data scenario, six datasets contained insufficient observed values for distance calculation, particularly for KNN imputation. To ensure consistency and comparability across all imputation methods, these datasets were excluded from the analysis, resulting in 94 usable datasets instead of 100. Each imputed dataset was subsequently analyzed using the ISA and Plaid biclustering algorithms. The following presents a flowchart that contains the stages of the research.

## 3. Results and Discussion

This section presents the biclustering results obtained using the ISA algorithm and the Plaid Model on complete and imputed datasets.

### 3.1. Exploratory Data Analysis

Exploratory analysis was conducted using boxplots and correlation heatmaps on standardized data to examine variable distributions, outliers, and inter-variable relationships. Data exploration was performed on data that had been standardized using z-scores.

Based on Fig. 4, exploratory analysis indicated substantial variability across provinces and several strong correlations among macroeconomic indicators, supporting the application of biclustering analysis.

### 3.2. Bicluster Profile on Complete Data

#### 3.2.1. Iterative Signature Algorithm (ISA)

Based on the parameter tuning procedure described in Section 2.5, the optimal row and column thresholds for ISA using 100 seeds, were identified as 0.45 and 0.75, respectively (seed = 456). This parameter combination produced three biclusters with coherent structures and adequate row and column memberships. The resulting bicluster membership structure is presented in Table 2.

Based on Table 2, BC1 is characterized by indicators from the Labor and Income, Regional Fiscal Structure, and GRDP Expenditure Structure groups and mainly consists of provinces in Kalimantan and Eastern Indonesia. This pattern suggests that regional economic activity remains closely linked to labor-market conditions and government expenditure, indicating that public spending continues to play an important role in supporting economic development in these regions. BC2 is dominated by indicators related to Economic Production and Growth

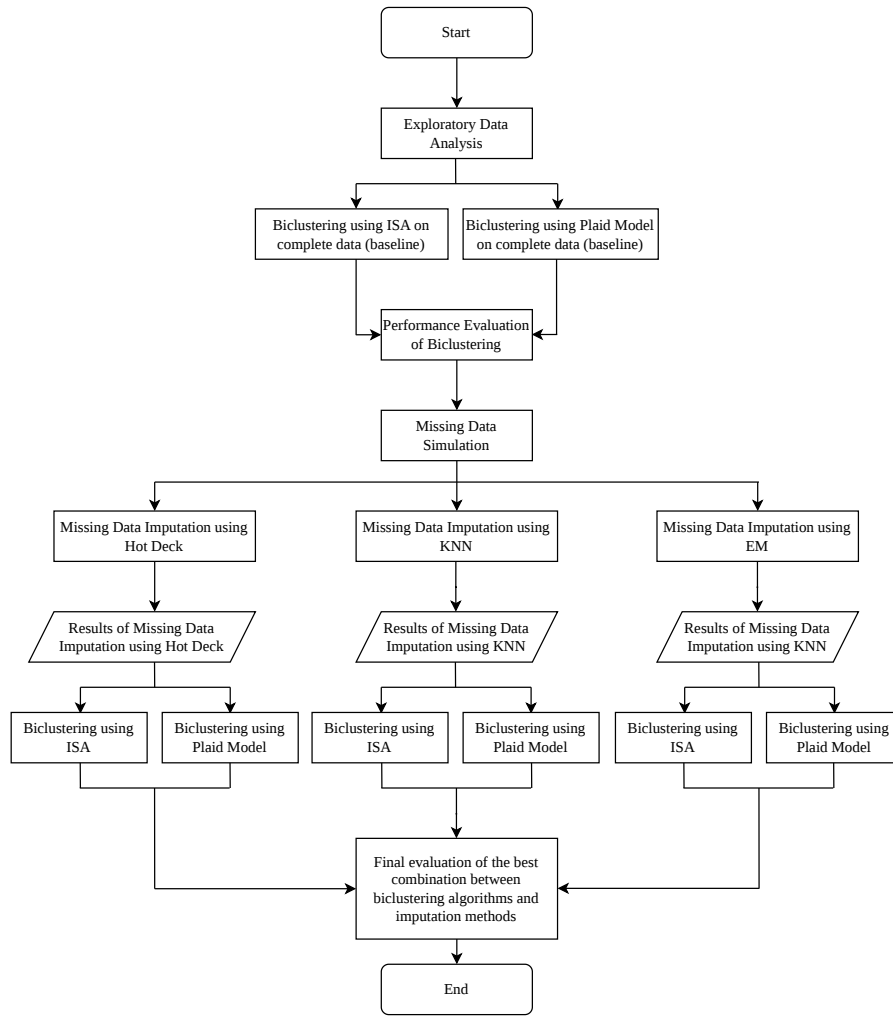


Fig. 3: Research Flowchart

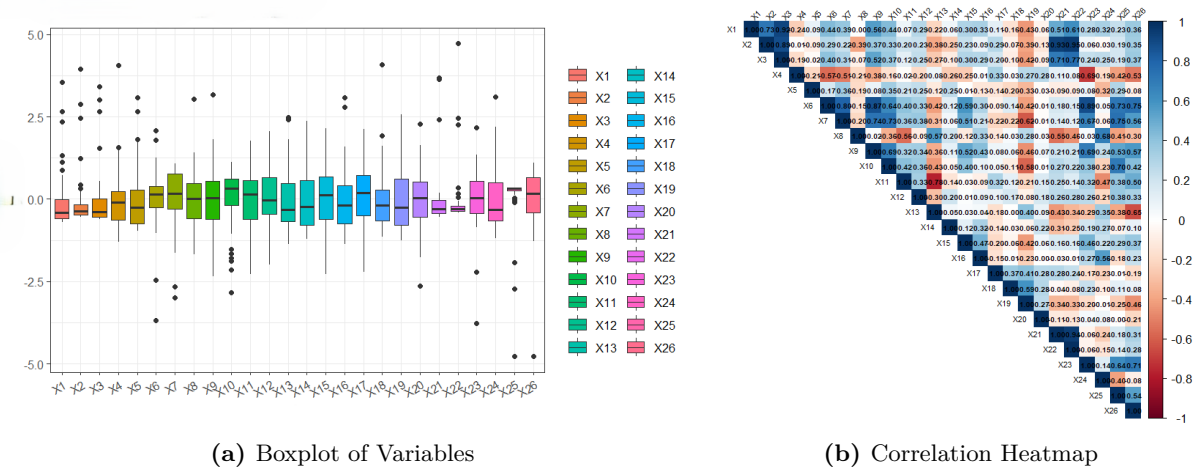


Fig. 4: Exploratory Data

together with a tourism-related indicator and includes several major economic provinces on Java and Sumatera. The bicluster reflects a development pattern driven by productive economic activities, regional output, and tourism demand, highlighting the contribution of the service sector to regional economic growth. BC3 is characterized by Human Development/Welfare, Basic

**Table 2:** Bicluster Membership Structure of ISA

BC	BC Size	Row (Province)	Column (Variable)
1	10 × 5	Kalimantan Tengah, Kalimantan Selatan, Kalimantan Utara, Maluku Utara, Papua Barat, Papua Barat Daya, Papua, Papua Selatan, Papua Tengah, Papua Pegunungan	X4, X8, X13, X19, X24
2	6 × 5	Sumatera Utara, Jawa Barat, Jawa Tengah, Jawa Timur, Papua Barat, Papua Tengah	X1, X2, X3, X21, X22
3	15 × 7	Sumatera Utara, Sumatera Barat, Sumatera Selatan, Jambi, Kepulauan Riau, DKI Jakarta, DI Yogyakarta, Banten, Bali, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Selatan	X6, X7, X9, X10, X15, X16, X23

Quality of Life, and Tourism and Service Sector indicators and includes several major urban and tourism-oriented provinces such as DKI Jakarta, DI Yogyakarta, Bali, and Kepulauan Riau. This pattern indicates that improvements in welfare and quality of life tend to coexist with stronger tourism and service-sector performance, suggesting a close relationship between human development and service-based economic activities.

### 3.2.2. Plaid Model

The Plaid Model was developed using four model specifications: constant, row-constant, column-constant, and two-way. Parameter optimization was conducted for the row release and column release parameters, while back-fitting (two iterations), shuffle testing (five permutations), and a maximum of 10 layers were applied to improve estimation quality and assess bicluster significance. Based on the evaluation procedure described previously, the optimal configuration was obtained under the row-constant model with a row release of 0.50 and a column release of 0.75, producing five biclusters. The resulting bicluster membership structure is presented in [Table 3](#).

**Table 3:** Bicluster Membership Structure of Plaid Model

BC	BC Size	Row (Province)	Column (Variable)
1	4 × 4	Jawa Barat, Jawa Tengah, Jawa Timur, Papua Tengah	X2, X3, X21, X22
2	2 × 5	Papua Selatan, Papua Pegunungan	X6, X7, X23, X25, X26
3	2 × 4	Nusa Tenggara Timur, Papua Pegunungan	X4, X17, X19, X20
4	4 × 2	Papua Barat Daya, Papua, Papua Selatan, Papua Tengah	X8, X14
5	11 × 2	Sumatera Barat, Bengkulu, Jawa Tengah, DI Yogyakarta, Jawa Timur, Nusa Tenggara Timur, Kalimantan Timur, Kalimantan Utara, Sulawesi Tenggara, Maluku, Papua Tengah	X13, X24

Based on [Table 3](#), BC1 is characterized by indicators from the Economic Production and Growth group together with a tourism-related indicator and includes several major economic provinces on Java and Papua Tengah. This pattern suggests that regional economic growth is closely associated with tourism-related activities, highlighting the contribution of the service sector to productive economic performance. BC2 is characterized by Human Development/Welfare and Basic Quality of Life indicators and consists of Papua Selatan and Papua Pegunungan. The bicluster indicates that welfare conditions in these newly established provinces are primarily reflected through human development and living-standard indicators, emphasizing the close relationship between human welfare and quality of life during socioeconomic development. BC3 is characterized by Labor and Income indicators together with a GRDP Expenditure Structure indicator and includes Nusa Tenggara Timur and Papua Pegunungan. This pattern suggests that labor-market conditions and household income remain closely linked to expenditure-

driven economic activities, indicating the importance of labor participation and consumption in supporting regional economies. BC4 is characterized by Labor and Income and Regional Fiscal Structure indicators and mainly consists of provinces in the Papua region. The bicluster suggests that labor conditions remain closely associated with government fiscal intervention, highlighting the continuing role of public expenditure in supporting regional development. BC5 is characterized by Regional Fiscal Structure and Labor and Income indicators and includes provinces from several regions of Indonesia. This pattern indicates that fiscal capacity and labor-market conditions are closely interconnected across diverse regional contexts and jointly contribute to regional economic performance. Based on the biclustering results obtained using the ISA algorithm and the Plaid Model, a comparison of algorithm performance was conducted based on the mean values of MSR, VEt, and SCS, as well as the number of biclusters generated and the percentage of row and column memberships successfully grouped.

**Table 4:** Performance Comparison of the ISA Algorithm and the Plaid Model

Algorithm	Average Evaluation			Number of Biclusters	Membership (%)	
	MSR	VEt	SCS		Row	Column
ISA	0,3314	0,4462	0,3077	3	63,16	65,38
Plaid Model	0,1369	0,1327	0,0086	5	42,10	65,38

Based on Table 4, the Plaid Model produced lower MSR, VEt, and SCS values, indicating more homogeneous biclusters. This result is consistent with its additive-layer mechanism and pruning procedure, which retain only strongly contributing elements. In contrast, ISA generated broader memberships (63.16% vs 42.10%), reflecting its tendency to capture wider local patterns at the expense of bicluster homogeneity.

### 3.3. Performance of Imputation Methods on Bicluster Structure

As previously explained in the simulation design (Section 2.8), a total of 594 simulated datasets were analyzed to evaluate biclustering performance after imputation. Bicluster consistency was assessed using the Jaccard Index (JI), whereas bicluster quality was evaluated using MSR, VEt, and SCS. For KNN imputation, the optimal number of neighbors was  $k = 4$  for the 5% missing-data scenario and  $k = 3$  for the 10% scenario.

**Table 5:** Final Simulation Evaluation of Imputation Performance in Biclustering Algorithms

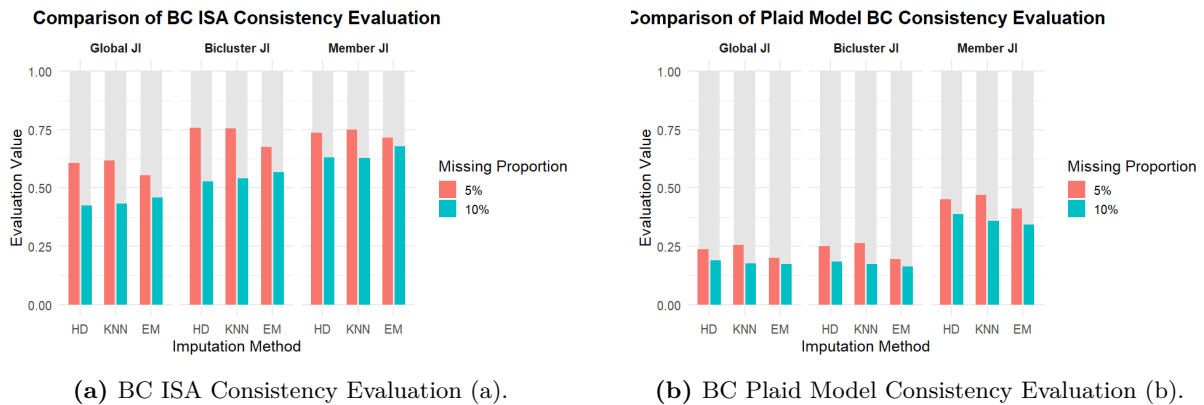
Algorithm	Proportion	Imputation	Bicluster Consistency Evaluation			Bicluster Quality Evaluation		
			Average JI Global	Average JI Bicluster	Average JI Member	Average MSR	Average Transpose VE	Average SCS
ISA	5%	HD	0.6078	<b>0.7565</b>	0.7377	<b>0.4653</b>	0.5358	0.3587
		KNN	<b>0.6167</b>	0.7561	<b>0.7496</b>	0.4715	0.5383	0.3650
		EM	0.5545	0.6761	0.7146	0.5110	<b>0.5113</b>	<b>0.3414</b>
	10%	HD	0.4246	0.5286	0.6300	0.5638	0.5493	0.3541
		KNN	0.4330	0.5413	0.6282	<b>0.5551</b>	0.5422	0.3523
		EM	<b>0.4580</b>	<b>0.5683</b>	<b>0.6783</b>	0.5764	<b>0.5271</b>	<b>0.3397</b>
Model Plaid	5%	HD	0.2368	0.2490	0.4503	0.1447	0.2079	0.0697
		KNN	<b>0.2556</b>	<b>0.2631</b>	<b>0.4693</b>	<b>0.1439</b>	<b>0.1711</b>	<b>0.0572</b>
		EM	0.1995	0.1959	0.4113	0.1547	0.1950	0.0722
	10%	HD	<b>0.1888</b>	<b>0.1845</b>	<b>0.3865</b>	0.1489	0.1827	0.0699
		KNN	0.1764	0.1735	0.3590	<b>0.1481</b>	<b>0.1792</b>	<b>0.0684</b>
		EM	0.1741	0.1618	0.3423	0.1512	0.2101	0.0717

Based on Table 5, bicluster structures obtained after imputation were not identical to those from the complete data, as indicated by Jaccard Index values below one in all scenarios. Nevertheless, ISA consistently achieved higher JI values than the Plaid Model, indicating better preservation of bicluster memberships after imputation. This difference is likely related to the underlying mechanisms of the algorithms. ISA directly searches for coherent local patterns and is

therefore more tolerant of moderate perturbations introduced by imputation, whereas the Plaid Model relies on the estimation of additive bicluster layers that are more sensitive to changes in the data structure.

As expected, increasing the missing-data proportion from 5% to 10% reduced bicluster consistency for both algorithms. The loss of information altered the row-column relationships required to reconstruct the original bicluster structures, although the magnitude of this reduction varied across imputation methods and biclustering algorithms. For ISA, KNN produced the highest consistency measures at the 5% missing-data proportion, whereas EM performed best at the 10% proportion. The neighborhood-based mechanism of KNN effectively preserves local similarity structures when missingness is relatively low, while the model-based estimation of EM becomes advantageous under higher missingness because it utilizes information from the overall data distribution to reconstruct the underlying structure. A different pattern was observed for the Plaid Model. KNN consistently produced the lowest MSR, VEt, and SCS values across both missing-data proportions, suggesting that preserving local variability is important for accurately estimating bicluster layers. In contrast, the smoothing effect of EM may reduce local contrasts among observations, leading to less distinguishable biclusters and lower quality measures. Meanwhile, Hot Deck showed relatively better membership recovery under the 10% missing-data scenario, although it did not produce the best bicluster quality.

The contrasting results obtained by ISA and the Plaid Model suggest a trade-off between bicluster recovery and bicluster homogeneity. ISA tends to retain broader bicluster memberships, resulting in higher Jaccard Index values and better recovery of the original bicluster structure. In contrast, the Plaid Model identifies more selective biclusters, leading to lower MSR, VEt, and SCS values and consequently better bicluster quality. This difference may also be influenced by the heterogeneous nature of the macroeconomic data, which comprise economic, social, fiscal, and tourism indicators with varying relationships across provinces. Overall, the results indicate that the effectiveness of an imputation method depends not only on the proportion of missing data but also on the compatibility between the imputation mechanism, the biclustering algorithm, and the characteristics of the analyzed data.

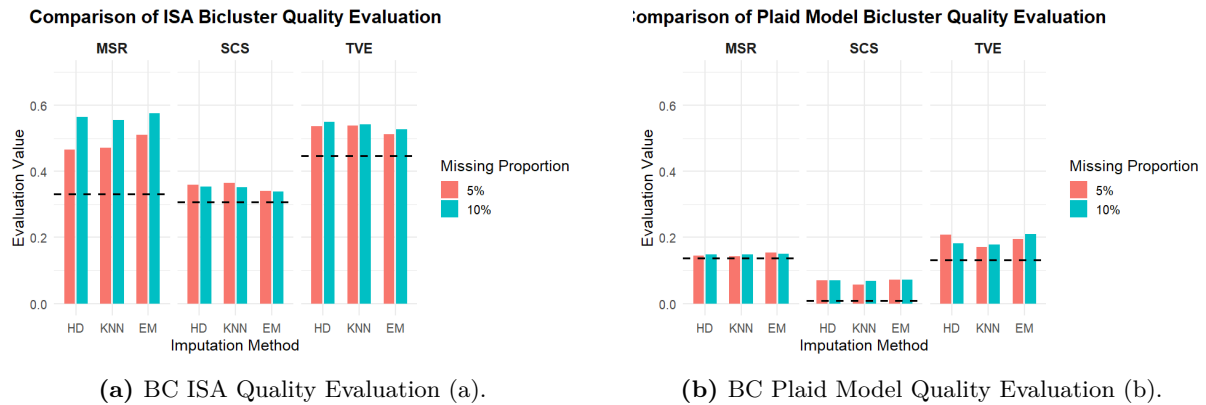


**Fig. 5:** Comparison of BC Consistency Evaluation: (a) ISA Consistency Evaluation, and (b) Plaid Model Consistency Evaluation.

To assess simulation variability, 95% confidence intervals were calculated for all evaluation metrics across the simulation repetitions. Overall, the confidence intervals were relatively narrow, indicating stable performance estimates and suggesting that the comparative conclusions regarding the imputation methods and biclustering algorithms were not substantially influenced by random simulation variability.

### 3.4. Aligned Rank Transform

ART ANOVA was performed to evaluate the effects of missing-data proportion, imputation method, and biclustering algorithm on bicluster quality (MSR) and bicluster consistency (JI



**Fig. 6:** Comparison of BC Quality Evaluation: (a) ISA Quality Evaluation, and (b) Plaid Model Quality Evaluation.

Bicluster), as well as their interaction effects.

**Table 6:** Table of ART ANOVA Results

MSR as the Response Variable		
Factor	p-value	Decision
Proportion	$< 2, 2 \times 10^{-16}$	Reject H0
Imputation	$1, 7270 \times 10^{-5}$	Reject H0
Algorithm	$< 2, 2 \times 10^{-16}$	Reject H0
Proportion*Imputation	0,8100	Fail to Reject H0
Proportion*Algorithm	$1, 0415 \times 10^{-9}$	Reject H0
Imputation*Algorithm	0,1774	Fail to Reject H0
Proportion*Imputation*Algorithm	0,6464	Fail to Reject H0
JI Bicluster as the Response Variable		
Factor	p-value	Decision
Proportion	$< 2, 2 \times 10^{-16}$	Reject H0
Imputation	0,3047	Fail to Reject H0
Algorithm	$< 2, 2 \times 10^{-16}$	Reject H0
Proportion*Imputation	$3, 7479 \times 10^{-5}$	Reject H0
Proportion*Algorithm	$5, 8950 \times 10^{-11}$	Reject H0
Imputation*Algorithm	0,3959	Fail to Reject H0
Proportion*Imputation*Algorithm	0,0864	Fail to Reject H0

The ART ANOVA results indicate that missing-data proportion and biclustering algorithm significantly affected both MSR and bicluster-level JI. Several interaction effects were also detected, particularly between missing-data proportion and biclustering algorithm. These findings suggest that biclustering performance is influenced not only by the amount of missing information but also by the interaction between imputation conditions and the characteristics of the biclustering algorithm.

## 4. Conclusion

This study applied the Iterative Signature Algorithm (ISA) and Plaid Model biclustering algorithms to identify groups of Indonesian provinces based on macroeconomic indicators. The results show that both algorithms were able to identify meaningful bicluster structures, although they exhibited different characteristics. ISA generated biclusters with broader memberships and a greater ability to recover bicluster structures, whereas the Plaid Model produced more selective biclusters with higher internal coherence and overall quality. These findings indicate that different

biclustering algorithms may reveal different perspectives of the underlying relationships between provinces and macroeconomic indicators.

The performance evaluation conducted after the imputation process showed that the effectiveness of biclustering results depends on the interaction between the imputation method and the biclustering algorithm. Overall, ISA achieved higher bicluster consistency, while the Plaid Model demonstrated better bicluster quality and greater stability under increasing levels of missing data. Among the imputation methods considered, KNN generally provided the most consistent performance across simulation scenarios. These results highlight the importance of jointly considering imputation and biclustering procedures when analyzing incomplete macroeconomic data.

This study is limited to the MCAR missing-data mechanism, relatively low missing-data proportions (5% and 10%), three imputation methods, and two biclustering algorithms. Therefore, the findings should be interpreted within these experimental settings. Future research may extend this work by investigating MAR and MNAR mechanisms, considering higher levels of missingness, incorporating additional imputation and biclustering techniques, and evaluating biclustering performance across different datasets and application domains.

## **CRedit Authorship Contribution Statement**

**Yani Prihantini Hiola:** Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing–Original Draft Preparation, Visualization. **I Made Sumertajaya:** Conceptualization, Methodology, Formal Analysis, Writing–Review & Editing, Supervision, Project Administration. **Indahwati:** Methodology, Formal Analysis, Writing–Review & Editing, Supervision.

## **Declaration of Generative AI and AI-assisted technologies**

Generative AI tools such as ChatGPT, Gemini, and DeepL Translator were utilised exclusively for language refinement and grammar editing during the preparation of this work.

## **Declaration of Competing Interest**

The authors declare no competing interests.

## **Funding and Acknowledgments**

This research received no external funding.

## **Data and Code Availability**

The data analyzed in this study are publicly available and were compiled from various dynamic tables and official statistical publications accessed through the website of Badan Pusat Statistik<sup>1</sup>. Additionally, the R scripts developed for data preprocessing, missing-data simulation, imputation, biclustering analysis, and performance evaluation are available from the corresponding author upon reasonable request.

## **References**

- [1] Ardhendu Mandal, Debasmita Saha, and Manas Sarkar. “Follicle Segmentation Using K-Means Clustering from Ultrasound Image of Ovary”. In: *Advances in Intelligent Systems and Computing*. Vol. 1255. Springer Science and Business Media Deutschland GmbH, 2021, pp. 545–553. DOI: [10.1007/978-981-15-7834-2\\_51](https://doi.org/10.1007/978-981-15-7834-2_51).

---

<sup>1</sup><https://www.bps.go.id>

- [2] Antonio M. Sudoso. “Exact and Heuristic Algorithms for Constrained Biclustering”. In: *Computational Optimization and Applications* (2026). DOI: [10.1007/s10589-026-00764-6](https://doi.org/10.1007/s10589-026-00764-6).
- [3] Qiyiwen Zhang, Changgee Chang, and Qi Long. “Robust Knowledge-Guided Biclustering for Multi-Omics Data”. In: *Briefings in Bioinformatics* 25.1 (2023). DOI: [10.1093/bib/bbad446](https://doi.org/10.1093/bib/bbad446).
- [4] Binda Wang, Yunwen Miao, Hongya Zhao, Jian Jin, and Yizeng Chen. “A Biclustering-Based Method for Market Segmentation Using Customer Pain Points”. In: *Engineering Applications of Artificial Intelligence* 47 (2016), pp. 101–109. DOI: [10.1016/j.engappai.2015.06.005](https://doi.org/10.1016/j.engappai.2015.06.005).
- [5] Eduardo N. Castanho, Helena Aidos, and Sara C. Madeira. “Biclustering Data Analysis: A Comprehensive Survey”. In: *Briefings in Bioinformatics* 25.4 (2024). DOI: [10.1093/bib/bbae342](https://doi.org/10.1093/bib/bbae342).
- [6] Cynthia Wulandari. “Evaluasi Hasil Analisis Bicluster Provinsi dan Potensi Perikanan Tangkap di Indonesia antara Algoritme BCBimax dan ISA”. Tesis. Bogor, Indonesia: Institut Pertanian Bogor, 2023. <https://repository.ipb.ac.id/handle/123456789/116689>.
- [7] Nur Hikmah. “Evaluasi Kinerja Biclustering Algoritme BCBimax dan Model Plaid dalam Melihat Pola Ketahanan Pangan Wilayah Indonesia”. Tesis. Bogor, Indonesia: Institut Pertanian Bogor, 2024. <https://repository.ipb.ac.id/handle/123456789/144244>.
- [8] J. Li, J. Reisner, H. Pham, S. Olafsson, and S. Vardeman. “Biclustering with Missing Data”. In: *Information Sciences* 510 (2020), pp. 304–316. DOI: [10.1016/j.ins.2019.09.047](https://doi.org/10.1016/j.ins.2019.09.047).
- [9] Francis Huang and Brian Keller. “Working with Missing Data in Large-Scale Assessments”. In: *Large-scale Assessments in Education* 13.1 (2025). DOI: [10.1186/s40536-025-00248-9](https://doi.org/10.1186/s40536-025-00248-9).
- [10] Lijuan Ren, Tao Wang, Aicha Sekhari Seklouli, Haiqing Zhang, and Abdelaziz Bouras. “A Review on Missing Values for Main Challenges and Methods”. In: *Information Systems* 119 (2023), p. 102268. DOI: [10.1016/j.is.2023.102268](https://doi.org/10.1016/j.is.2023.102268).
- [11] Lateef Babatunde Amusa and Twinomurinzi Hossana. “An Empirical Comparison of Some Missing Data Treatments in PLS-SEM”. In: *PLOS ONE* 19.1 (2024), e0297037. DOI: [10.1371/journal.pone.0297037](https://doi.org/10.1371/journal.pone.0297037).
- [12] Indry Handayani, Dadan Kusnandar, and Wirda Andani. “Pendugaan Data Hilang Menggunakan Perbandingan Imputasi Hot-Deck dan K-Nearest Neighbor”. In: *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)* 12.4 (2023), pp. 319–324. <https://jurnal.untan.ac.id/index.php/jbmstr/article/view/68323>.
- [13] Tsasya Raudhatunnisa and Nori Wilantika. “Performance Comparison of Hot-Deck Imputation, K-Nearest Neighbor Imputation, and Predictive Mean Matching in Missing Value Handling, Case Study: March 2019 SUSENAS Kor Dataset”. In: *Proceedings of The International Conference on Data Science and Official Statistics* 2021.1 (2022), pp. 753–770. DOI: [10.34123/icdsos.v2021i1.93](https://doi.org/10.34123/icdsos.v2021i1.93).
- [14] Loc Nguyen. “Handling Missing Data with Expectation Maximization Algorithm”. In: *Global Research and Development Journals* (2021). DOI: [10.70179/9473zk43](https://doi.org/10.70179/9473zk43).
- [15] Tan Duy Le, Razvan Beuran, and Yasuo Tan. “Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare”. In: *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2018, pp. 247–251. DOI: [10.1109/KSE.2018.8573344](https://doi.org/10.1109/KSE.2018.8573344).
- [16] I Made Sumertajaya Sumertajaya, Wiwik Andriyani Lestari Ningsih, Asep Saefuddin, and Embay Rohaeti. “Biclustering Performance Evaluation of Cheng and Church Algorithm and Iterative Signature Algorithm”. In: *JTAM (Jurnal Teori dan Aplikasi Matematika)* 7.3 (2023), pp. 643–653. DOI: [10.31764/jtam.v7i3.14778](https://doi.org/10.31764/jtam.v7i3.14778).

- [17] Titin Siswantining, Achmad Eriza Aminanto, Devvi Sarwinda, and Olivia Swasti. “Bi-clustering Analysis Using Plaid Model on Gene Expression Data of Colon Cancer”. In: *Austrian Journal of Statistics* 50.5 (2021), pp. 101–114. DOI: [10.17713/ajs.v50i5.1195](https://doi.org/10.17713/ajs.v50i5.1195).
- [18] Badan Pusat Statistik. *Indikator Ekonomi Januari 2024*. BPS-Statistics Indonesia Publication. 2024. <https://www.bps.go.id/id/publication/2024/03/28/78f4412bc3d4b2909067c841/indikator-ekonomi-januari-2024.html>.
- [19] Rebecca R. Andridge and Roderick J. A. Little. “A Review of Hot Deck Imputation for Survey Non-response”. In: *International Statistical Review* 78.1 (2010), pp. 40–64. DOI: [10.1111/j.1751-5823.2010.00103.x](https://doi.org/10.1111/j.1751-5823.2010.00103.x).
- [20] Inayatul Izzati Diana Yusuf, Budi Susetyo, and La Ode Abdul Rahman. “Perbandingan Metode Hot-Deck, Regression dan K-Nearest Neighbor Imputation dalam Pendugaan Data Hilang pada Dapodik Tahun 2020”. In: *Xplore: Journal of Statistics* 12.1 (2023), pp. 22–35. DOI: [10.29244/xplore.v12i1.1056](https://doi.org/10.29244/xplore.v12i1.1056).
- [21] Rustam Mussabayev. “Optimizing Euclidean Distance Computation”. In: *Mathematics* 12.23 (2024), p. 3787. DOI: [10.3390/math12233787](https://doi.org/10.3390/math12233787).
- [22] Achmad Fikri Sallaby and Azlan Azlan. “Analysis of Missing Value Imputation Application with K-Nearest Neighbor (K-NN) Algorithm in Dataset”. In: *The IJICS (International Journal of Informatics and Computer Science)* 5.2 (2021), p. 141. DOI: [10.30865/ijics.v5i2.3185](https://doi.org/10.30865/ijics.v5i2.3185).
- [23] Edgar Acuña and Caroline Rodriguez. “The Treatment of Missing Values and Its Effect on Classifier Accuracy”. In: *Classification, Clustering, and Data Mining Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 639–647. DOI: [10.1007/978-3-642-17103-1\\_60](https://doi.org/10.1007/978-3-642-17103-1_60).
- [24] Haoyu Meng. “A Comparative Study on Missing Value Imputation Techniques in Machine Learning”. In: *SHS Web of Conferences* 218 (2025), p. 02014. DOI: [10.1051/shsconf/202521802014](https://doi.org/10.1051/shsconf/202521802014).
- [25] Lorenzo Beretta and Alessandro Santaniello. “Nearest Neighbor Imputation Algorithms: A Critical Evaluation”. In: *BMC Medical Informatics and Decision Making* 16.S3 (2016). DOI: [10.1186/s12911-016-0318-z](https://doi.org/10.1186/s12911-016-0318-z).
- [26] Khaled M. Fouad, Mahmoud M. Ismail, Ahmad Taher Azar, and Mona M. Arafa. “Advanced Methods for Missing Values Imputation Based on Similarity Learning”. In: *PeerJ Computer Science* 7 (2021), e619. DOI: [10.7717/peerj-cs.619](https://doi.org/10.7717/peerj-cs.619).
- [27] Kenneth Lange and Hua Zhou. “A Legacy of EM Algorithms”. In: *International Statistical Review* 90.S1 (2022). DOI: [10.1111/insr.12526](https://doi.org/10.1111/insr.12526).
- [28] F. V. Nelwamondo, S. Mohamed, and T. Marwala. “Missing Data: A Comparison of Neural Network and Expectation Maximization Techniques”. In: *Current Science* 93.11 (2007), pp. 1514–1521. <https://www.jstor.org/stable/24099079>.
- [29] Sven Bergmann, Jan Ihmels, and Naama Barkai. “Iterative Signature Algorithm for the Analysis of Large-Scale Gene Expression Data”. In: *Physical Review E* 67.3 (2003). DOI: [10.1103/PhysRevE.67.031902](https://doi.org/10.1103/PhysRevE.67.031902).
- [30] Jan Ihmels, Sven Bergmann, and Naama Barkai. “Defining Transcription Modules Using Large-Scale Gene Expression Data”. In: *Bioinformatics* 20.13 (2004), pp. 1993–2003. DOI: [10.1093/bioinformatics/bth166](https://doi.org/10.1093/bioinformatics/bth166).
- [31] Alexander L. Richards, Peter Holmans, Michael C. O’Donovan, Michael J. Owen, and Lesley Jones. “A Comparison of Four Clustering Methods for Brain Expression Microarray Data”. In: *BMC Bioinformatics* 9.1 (2008). DOI: [10.1186/1471-2105-9-490](https://doi.org/10.1186/1471-2105-9-490).

- [32] Laura Lazzeroni and Art Owen. “Plaid Models for Gene Expression Data”. In: *Statistica Sinica* 12.1 (2002), pp. 61–86. <https://www3.stat.sinica.edu.tw/statistica/j12n1/j12n14/j12n14.htm>.
- [33] I Made Sumertajaya, Nur Hikmah, and Farit Mochamad Afendi. “Comparative Analysis of BCBimax and Plaid Biclustering Algorithm for Pattern Recognition in Indonesia Food Security”. In: *BAREKENG: Jurnal Ilmu Matematika dan Terapan* 20.1 (2025), pp. 0335–0346. DOI: [10.30598/barekengvol20iss1pp0335-0346](https://doi.org/10.30598/barekengvol20iss1pp0335-0346).
- [34] Beatriz Pontes, Ral Giráldez, and Jesús S. Aguilar-Ruiz. “Quality Measures for Gene Expression Biclusters”. In: *PLOS ONE* 10.3 (2015), e0115497. DOI: [10.1371/journal.pone.0115497](https://doi.org/10.1371/journal.pone.0115497).
- [35] Paul Jaccard. “Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines”. In: *Bulletin de la Société Vaudoise des Sciences Naturelles* 37.140 (1901), pp. 241–272. DOI: [10.5169/seals-266440](https://doi.org/10.5169/seals-266440).

## Appendix

### Appendix A. List of Variables Used in the Study

**Table 7:** List of Variables Used in the Study

Variable	Variable Name	Unit
$X_1$	Realization of Domestic Investment	Billion Rp
$X_2$	Number of Domestic Tourist Trips	Trips
$X_3$	Gross Regional Domestic Product at Constant Prices 2010	Billion Rp
$X_4$	Labor Force Participation Rate	Percent
$X_5$	Farmer Exchange Rate	Index
$X_6$	Human Development Index	Index
$X_7$	Percentage of Non-Poor Population	Percent
$X_8$	Provincial Minimum Wage	Rupiah
$X_9$	Per Capita Expenditure	Thousand Rp
$X_{10}$	Percentage of Population with Adequate Food Consumption	Percent
$X_{11}$	Provincial Education Expenditure Budget	Percent
$X_{12}$	Provincial Health Expenditure Budget	Percent
$X_{13}$	Provincial Economic Expenditure Budget	Percent
$X_{14}$	Provincial Housing and Public Facilities Expenditure Budget	Percent
$X_{15}$	Occupancy Rate of Star Hotels	Percent
$X_{16}$	Occupancy Rate of Non-Star Hotels and Other Accommodations	Percent
$X_{17}$	Distribution of GRDP at Current Prices by Household Consumption Expenditure	Percent
$X_{18}$	Distribution of GRDP at Current Prices by NPISH Consumption Expenditure	Percent
$X_{19}$	Distribution of GRDP at Current Prices by Government Consumption Expenditure	Percent
$X_{20}$	Distribution of GRDP at Current Prices by Gross Fixed Capital Formation	Percent
$X_{21}$	Number of Micro-Scale Industrial Companies by Province	Unit
$X_{22}$	Number of Small-Scale Industrial Companies by Province	Unit
$X_{23}$	Mean Years of Schooling	Years
$X_{24}$	Average Monthly Net Wage/Salary of Employees by Province and Main Occupation	Rupiah
$X_{25}$	Percentage of Households Using Electricity as Main Lighting Source	Percent
$X_{26}$	Percentage of Households with Access to Proper Drinking Water Sources	Percent