Development of generalized principal component analysis using multiple imputation genetic algorithm

Fahrezal Zubedi[1,2], I Made Sumertajaya[1], Khairil Anwar Notodiputro[1], Utami Dyah Syafitri[1]
[1]Statistics and Data Science Study Program, School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia
[2]Statistics Study Program, Faculty of Mathematics and Natural Science, State University of Gorontalo, Gorontalo, Indonesia

Article Info

Article history:

Received Mar 27, 2025
Revised Oct 17, 2025
Accepted Nov 8, 2025

Keywords:

Biplot
Correlation
Dimensionality reduction
Missing values
Trends in human development

ABSTRACT

In this study, we propose an innovative method called the integrated GPCA-MIGA, which integrates the multiple imputation genetic algorithm (MIGA) and generalized principal component analysis (GPCA) to perform missing value imputation and data dimensionality reduction simultaneously. The approximated original data produced by GPCA serves as the basis for MIGA to update missing values in the next iteration. At the same time, GPCA refines the low-dimensional representation using the latest imputation results from MIGA, thereby balancing the accuracy of missing value imputation and the stability of dimensionality reduction. The objective of this study is to evaluate the performance of the integrated GPCA-MIGA and analyze trends in human development at the district/city level in Indonesia. The findings of this study show that the integrated GPCA-MIGA effectively reduces the dimensionality of data containing missing values compared to other methods. The integrated GPCA-MIGA method was applied to human development data. The results were then visualized using a biplot, which revealed that human development trends in Jayawijaya from 2019 to 2022 indicate progress in school enrollment rates for ages 16–18 years.

Corresponding Author:

I Made Sumertajaya
Statistics and Data Science Study Program, School of Data Science, Mathematics and Informatics
IPB University
Meranti Rd, Dramaga District, Bogor, 16680 West Java, Indonesia
Email: imsjaya@apps.ipb.ac.id

1. INTRODUCTION

One of the most used methods for dimensionality reduction is principal component analysis (PCA). This method summarizes many variables into a few principal components without significant loss of information. However, PCA has limitations when dealing with data that exhibits correlations across observations (rows) [1], [2]. To address this limitation, generalized principal component analysis (GPCA), also known as generalized low rank approximation of matrices (GLRAM), was developed. GPCA is capable of simultaneously reducing the dimensions of both correlated variables and correlated observations. In addition, GPCA can reduce the data dimension of a collection of matrices simultaneously [3]. This method was first introduced in [4] to enhance the efficiency of image compression and has been shown to produce higher visual quality and more efficient computation time compared to PCA.

Research on GPCA has been applied across various fields, accompanied by developments in methodological aspects. For example, GPCA has been used to identify genes with overlapping patterns, enabling the recognition of gene interactions and functions. In this context, GPCA has proven effective in compressing images of gene expression patterns [5]. This advantage was further examined in comparative studies, which demonstrated that GPCA outperforms other dimensionality reduction methods such as PCA and multilinear PCA, particularly in pattern recognition tasks like image classification and object identification [6]. In other words, GPCA is not only effective in practice but also supported by a strong and mathematically proven theoretical foundation. To address new challenges in the application of GPCA, several further developments have been carried out. One of them is the combination of GPCA with methods such as top-push constrained feature learning (TFL) to improve the accuracy of face image recognition through dimensionality reduction [7]. Another development is randomized GPCA, which aims to reduce the computational complexity of GPCA [8]. However, most developments of GPCA have continued to focus on improving accuracy and computational efficiency, without explicitly addressing or integrating the handling of missing value problems commonly found in empirical data.

The main problem in this study is how to effectively apply GPCA to data containing missing values. The presence of missing values can reduce analysis quality, both in terms of accuracy and interpretability [9], [10]. Conventional imputation methods, such as mean or median imputation, are often used for practicality but have fundamental limitations. Such methods tend to underestimate variance and alter the correlation structure among variables [11]. As a result, subsequent analyses, such as dimensionality reduction with GPCA, may yield representations that do not adequately reflect the structure of the original data. To address this issue, various imputation methods have been developed, one of which is the multiple imputation genetic algorithm (MIGA). Introduced in [12], MIGA combines multiple imputation principles with genetic algorithms (GA) to optimally estimate the missing values. This method is designed to preserve the statistical structure of data, such as mean, skewness, and covariance matrices, which are often distorted by conventional imputation methods. MIGA has been shown to outperform other algorithms such as expectation maximization (EM), auxiliary regressions, and k-nearest neighbors imputation (K-NNI). However, when GPCA is applied after MIGA (non-integrated), the dimensionality reduction process follows the imputed data that remain fixed, rather than data that adapt dynamically to the reduction model. This condition may lead to a larger number of retained dimensions and a decrease in the total explained variance. Therefore, this study proposes a new method called the integrated GPCA-MIGA, which is designed to perform dimensionality reduction on data containing missing values efficiently. In this method, the covariance structure generated by GPCA is directly utilized to update the MIGA imputation process in the subsequent iteration, thereby achieving a balance between imputation accuracy and reduction stability.

The empirical issue addressed in this study is how to analyze and visualize the trends of human development at the district/city level in Indonesia during the 2019 to 2022 period. During this period, the growth rate of Indonesia's human development index (HDI) tended to slow down. Additionally, disparities in human development achievements among district/city in Indonesia persist [13]–[16]. This situation indicates the need for a more in-depth and comprehensive analysis of the dynamics of human development at the district/city level. However, human development indicator data is characterized by high dimensionality, with correlations both among indicators and among district/city, and contain missing values, which complicate direct analysis and visualization. To address issues in empirical data, the integrated GPCA-MIGA is applied to simultaneously reduce the dimensions of district/city and variables, even when the data contains missing values. The reduced data are then visualized using the biplot approach, which enables yearly mapping of district/city positions and their associated indicators through a variety of informative visual displays. Based on the description above, this study has two main objectives: i) to evaluate the performance of the integrated GPCA-MIGA and ii) to analyze trends in human development at the district/city level in Indonesia from 2019 to 2022.

2. METHOD

2.1. Data

This study utilized both simulated and empirical data. The simulated data consists of four matrices ($A_j \in \mathbb{R}^{500\times100}, j = 1, 2, 3, 4$), each organized into 25 observation clusters and 5 variable clusters. Observations and variables within the same cluster are correlated, while those across clusters are uncorrelated. The four simulated matrices are constructed to be mutually correlated. The process of data generation is described as follows:

i) Generate a symmetric and positive definite covariance matrix $S \in \mathbb{R}^{20\times20}$ where all diagonal elements are 1 and all off-diagonal elements are 0.8.
ii) Apply cholesky decomposition to a matrix S to derive a matrix L and its transpose matrix L^t [17].
iii) Generate a matrix $Z \in \mathbb{R}^{500\times20}$ using a partition approach. Every division matrix is produced from multivariate normal (MVN) distribution, a distinct mean vector for each column, and the covariance matrix (diagonal).
iv) Transform every division matrix from matrix Z_p $\mathbb{R}^{500\times20}$ to matrix $A \in \mathbb{R}^{500\times20}$ employing the outcome of Cholesky decomposition via:

Submatrix 1 $A = L_{20\times20} \times Z_{20\times20} \times L^t_{20\times20}$
⋮
Submatrix 25 $A = L_{20\times20} \times Z_{20\times20} \times L^t_{20\times20}$

v) Repeat steps 3-4 four times, then concatenating the results horizontally, forming five clusters of variables. Variables within the same cluster are correlated, but those between clusters are not correlated. The matrix $A_1 \in \mathbb{R}^{500\times100}$ will be formed.

vi) Repeat steps 3-5 three times to generate X_1, X_2, and X_3. Under the condition of correlated matrices A_1, A_2, A_3 and A_4, as detailed below, each subsequent matrix A_{j+1} is obtained by adding X_j to the preceding matrix A_j, for j=1, 2, 3.

The simulated data were randomly removed under the missing completely at random (MCAR) mechanism with missing value percentages of 5, 10, 15, and 20% in each simulated data, while ensuring that no row or column had all its entries missing. This simulation analysis aims to evaluate the performance of the integrated GPCA-MIGA in reducing the dimensionality of data containing missing values. The performance of this method was compared with that of MIGA+GPCA (non-integrated), mean imputation+GPCA, and median imputation+GPCA. Furthermore, the results of these four approaches were compared with GPCA applied to complete data, which serves as the baseline representing the ideal condition without any information loss. The overall workflow of the simulation analytical procedure is systematically and structurally designed, as illustrated in Figure 1.

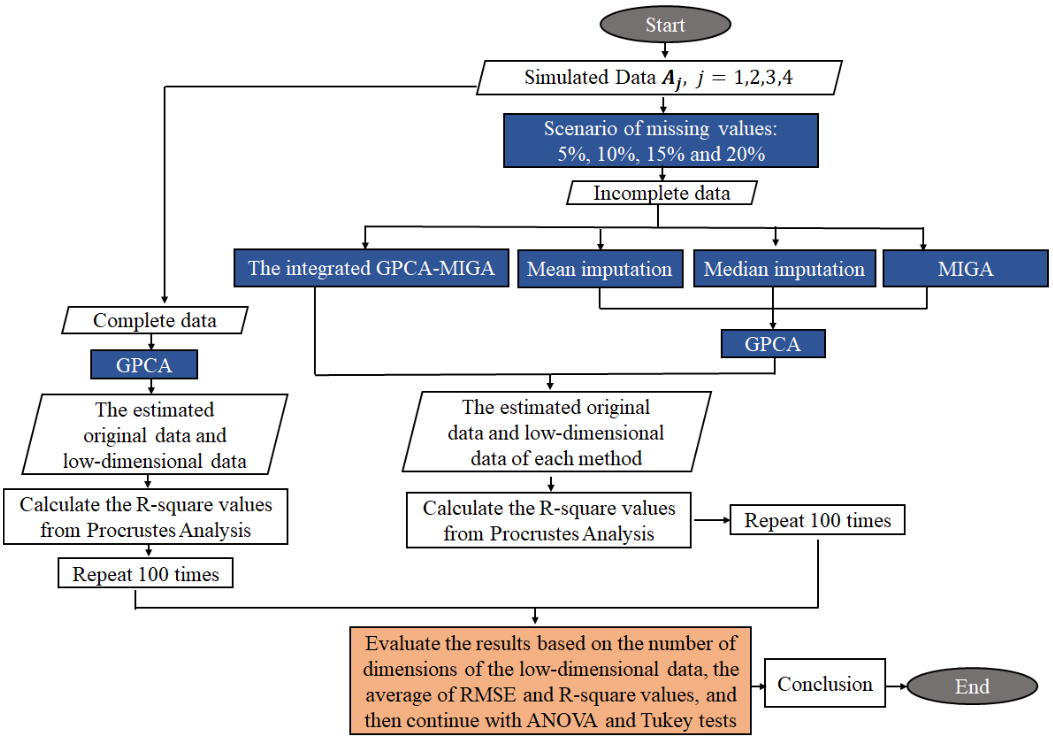

Start
Simulated Data A_j, j = 1,2,3,4
Scenario of missing values: 5%, 10%, 15% and 20%
Incomplete data
The integrated GPCA-MIGA
Mean imputation
Median imputation
MIGA
Complete data
GPCA
GPCA
The estimated original data and low-dimensional data
The estimated original data and low-dimensional data of each method
Calculate the R-square values from Procrustes Analysis
Calculate the R-square values from Procrustes Analysis
Repeat 100 times
Repeat 100 times
Evaluate the results based on the number of dimensions of the low-dimensional data, the average of RMSE and R-square values, and then continue with ANOVA and Tukey tests
Conclusion
End

Figure 1. Flowchart of the simulation data analysis

This study employed empirical data consisting of indicators related to the dimensions of human development at the district/city level in Indonesia for the period 2019 to 2022. These data were sourced from Statistics Indonesia (BPS) publications, which are available on the official websites of each district/city. The data consists of 20 indicators covering a total of 514 district/city in Indonesia. These indicators comprehensively represent the three dimensions of human development, namely long life and healthy life, knowledge, and a decent standard of living. The long life and health life dimension includes indicators such as the percentage of households with clean drinking water sources (X_1), the percentage of households with access to adequate drinking water (X_2), the percentage of households that do not have defecation facilities (X_3), and morbidity (X_4). The knowledge dimension includes indicators such as school enrollment rates for the age groups 7-12 years (X_5), 13-15 years (X_6), 16-18 years (X_7), gross participation rates at the elementary (X_8), junior high (X_9), and senior high school levels (X_{10}) and net participation rates at elementary (X_{11}), junior high (X_{12}), and senior high school levels (X_{13}). The decent standard of living dimension includes indicators such as the percentage of formal workers (X_{14}), the percentage of poor people (X_{15}), the open unemployment rate (X_{16}), the average wages of workers and employees per month (X_{17}), the gross regional domestic product per capita based on current prices (X_{18}), the percentage of informal workers (X_{19}), and the Gini ratio (X_{20}) [13]–[16]. The flowchart of the empirical data analysis is presented in Figure 2.

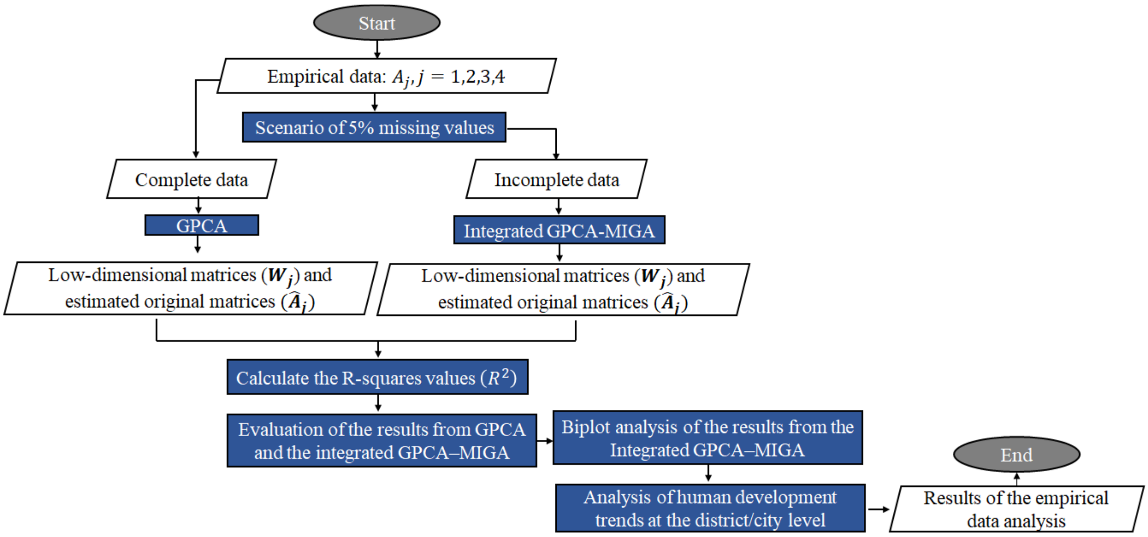

Start
Empirical data: $A_j, j = 1,2,3,4$
Scenario of 5% missing values
Complete data
Incomplete data
GPCA
Integrated GPCA-MIGA
Low-dimensional matrices ($\boldsymbol{W}_j$) and estimated original matrices ($\hat{\boldsymbol{A}}_j$)
Low-dimensional matrices ($\boldsymbol{W}_j$) and estimated original matrices ($\hat{\boldsymbol{A}}_j$)
Calculate the R-squares values (R^2)
Evaluation of the results from GPCA and the integrated GPCA–MIGA
Biplot analysis of the results from the Integrated GPCA–MIGA
End
Analysis of human development trends at the district/city level
Results of the empirical data analysis

Figure 2. Flowchart of the empirical data analysis

2.2. Generalized principal component analysis

GPCA is a dimensionality reduction method that simultaneously minimizes the number of dimensions of both observations and variables. To achieve this, GPCA is applied to a set of data matrices $\{A_j\}_{j=1}^{n} \in \mathbb{R}^{r\,x\,c}$, with the aim of obtaining a low-dimensional representation $\{W_j\}_{j=1}^{n} \in \mathbb{R}^{k\,x\,l}$, which is expressed as (1) [18].

$$W_j = P^t A_j Q \tag{1}$$

Where $P \in \mathbb{R}^{r\,x\,k}$ represents the final reduced form for the observations dimension, while $Q \in \mathbb{R}^{c\,x\,l}$ represents the final reduced form for the variable dimension. Conversely, to approximate a set of data matrices, the (2) is used.

$$\hat{A}_j \approx P\, W_j\, Q^t \tag{2}$$

The optimal P and Q matrices are obtained through an iterative procedure until a convergence criterion based on root mean square error (RMSE) is met. Once convergence is reached, the low-dimensional representation and the approximated original data matrices are computed using the final P and Q matrices [19]. The detailed steps of the GPCA are presented in Algorithm 1, specifically in steps 11 to 23 of the integrated GPCA–MIGA algorithm [20].

GPCA also facilitates analysis of the positions of observations and variables using the biplot concept, a two-dimensional visualization that simplifies the interpretation of multivariate relationships. Regions (district/city) aligned with indicator vectors are interpreted as having above-average values, those in the opposite direction as below-average, and those near the center as close to the average [21]–[23]. After performing the GPCA, the principal component scores for observations and variables are calculated as follows: $\hat{A}_j = G\, H^t$. The positions of observations and variables are plotted using principal component scores as shown in (3) and (4).

$$G = P\, W^{\alpha} \tag{3}$$

$$H = (H^t)^t = (W^{\alpha} Q^t)^t \tag{4}$$

Where $\alpha = 0.5$, G contains observation coordinates, and H contains variable coordinates [1].

2.3. Multiple imputation genetic algorithm

The MIGA was developed to address the problem of missing values in multivariate data by integrating imputation techniques with GA as an optimization method [24], [25]. The basic idea is to impute missing values by generating candidate imputations that not only approximate the original data but also preserve the essential statistical properties of multivariate data, such as means, covariances, and skewness. The core insight of MIGA is that GA, with their evolutionary nature, can be employed to explore a wide solution space for imputations, ensuring that missing data imputation is not merely the act of filling empty entries but a process of maintaining the distributional characteristics and inter-variable relationships necessary for valid and reliable statistical analysis. The main contribution of MIGA lies in the formulation of a multiobjective fitness function based on the Minkowski distance, which simultaneously preserves means, covariances, and skewness. The fitness function f_r is defined as (5).

$$f_r := min\left(\left(D_r(\tilde{x}_A, \tilde{x}_C) + D_r\left(\tilde{S}, I\right) + D_r(b_A, b_C)\right)\right) \tag{5}$$

Where $D_r(\tilde{x}_A, \tilde{x}_C)$ represents the distance between the relative means of the complete data and the imputed data, while $D_r\left(\tilde{S}, I\right)$ represents the distance between the relative covariance matrix ($\tilde{S}$) and the identity matrix (I), thus ensuring that the covariance and correlation structures are preserved. Similarly, $D_r(b_A, b_C)$ represents the distance between the skewness vectors of the complete data and the imputed data [12], [26].

2.4. The integrated GPCA-MIGA algorithm

The integrated GPCA-MIGA algorithm imputes missing values and reduces data dimensionality within a unified iterative framework. In each iteration, MIGA imputes missing values through evolutionary operations including selection, crossover, and mutation, followed by GPCA, which performs data dimensionality reduction. The approximate original data obtained from GPCA contains elements in the positions of previously missing values. These elements serve as the basis for MIGA to update the imputed missing values in the next iteration. GPCA then refines the low-dimensional representation using the latest imputation results from MIGA. The iterative process terminates when the relative change of all imputed missing entries between two consecutive iterations is less than or equal to 0.001. The detailed steps of the integrated GPCA-MIGA algorithm are presented in Algorithm 1.

Algorithm 1. The integrated GPCA-MIGA

```
Input   : $A_1$, $A_2$, $A_3$, $A_4$, $c_1 = c_2 = 5, c_3 = 10, l = 100$
Output  : $W_1, W_2, W_3, W_4, \hat{A}_1, \hat{A}_2, \hat{A}_3, \hat{A}_4, P, Q$, RMSE, fitness values
1.  Create matrices $X_{A_1}, X_{A_2}, X_{A_3}$ and $X_{A_4}$ which consist of complete observations (rows) from matrices $A_1, A_2, A_3, A_4$.
2.  Compute the column-wise mean, skewness, and the covariance matrix for each of $X_{A_1}$, $X_{A_2}$, $X_{A_3}$, and $X_{A_4}$.
3.  Create matrices $X_{C_1}, X_{C_2}, X_{C_3}$ and $X_{C_4}$ consisting of observations (rows) from $A_1, A_2, A_3, A_4$ with at least one missing element.
4.  Create vector indices $m$ of $X_{C_1}, X_{C_2}, X_{C_3}$ and $X_{C_4}$ within matrices $A_1, A_2, A_3, A_4$.
5.  Generate an initial population containing $l$ individuals based on the distribution of variables for each matrix.
6.      Compute fitness values for all individuals in the population.
7.      Select $c$ individuals from the population with the smallest fitness values.
8.      Perform mutation on the selected $c_1$ individuals and then repeat it $c_3$ times per individual.
9.      Perform crossover on the selected $c_1$ individuals and then repeat it for the $c_2 - 1$ remaining individuals.
10.     Recalculate the fitness values of individuals after crossover and mutation. The individual with the smallest fitness value is used as the imputation for missing values.
11.     Standardize data matrices $A_1, A_2, A_3, A_4$ to form $S_1, S_2, \ldots, S_n$.
12.     Initialize matrix $P$ as an identity matrix $P_0 = (I, 0)^t$.
13.     $i = 0, \mathrm{RMSE}(i) = \infty$
14.         Compute matrix $M_R$ using the formula: $M_R = \sum_{j=1}^{n} S_j^t P_i P_i^t S_j$.
15.         Determine the $v$ eigenvectors $\{\beta_j^Q\}_{j=1}^{v}$from $M_R$ that correspond to a cumulative variance proportion (≤90%), resulting in $Q_i = \left[\beta_1^Q, \ldots, \beta_v^Q\right]$.
16.         Compute matrix $M_L$ using the formula $M_L = \sum_{j=1}^{n} S_j Q_i Q_i^t S_j^t$.
17.         Determine the $v$ eigenvectors $\{\beta_j^P\}_{j=1}^{v}$ from $M_L$that correspond to a cumulative variance proportion (≤90%), resulting in $P_i = [\beta_1^P, \ldots, \beta_v^P]$.
```

```
18.     Compute RMSE as $RMSE(i) = \sqrt{\frac{1}{n}\sum_{j=1}^{n} ||S_j - P_i P_i^t S_j Q_i Q_i^t||_F^2}$.
19.     Repeat steps 13-17 until $RMSE(i-1) - RMSE(i) \leq 0{,}001$.
20.     Derive the matrices $P$ and $Q$ from concluding iteration.
21.     Create the reduced-dimensional data matrix using $W_j = P^t S_j Q$, for each $j$ from $i$ to $n$.
22.     Approximate the original data matrix as $\hat{S}_j = PW_jQ^T$, for each $j$ from $i$ to $n$.
23.     Convert the scale of the data matrix $\hat{S}_j$ to the original scale $\hat{A}_j$.
24.     Save the imputed data values from the first iteration of the integrated GPCA-MIGA.
25.     Calculate the relative change for each missing value point.
26. Repeat steps 6 to 25 using the updated individuals. If the difference in relative change between the
    current and previous iterations of all imputed missing values is $\leq 1.10^{-3}$, then the iterative process
    stops.
27. Obtain the dimensionally reduced data from the integrated GPCA–MIGA and the approximated
    original data.
```

The explanation of the integrated GPCA-MIGA algorithm is followed by a performance evaluation that compares it with several alternative approaches to assess its effectiveness in handling missing value and performing dimensionality reduction. To clarify the distinct characteristics of each approach, a comparative analysis was conducted among the methods applied in this study. This comparison highlights the strengths and limitations of each technique, aiming to determine how the integration between imputation strategies and dimensionality reduction methods contributes to achieving accurate, stable, and computationally efficient data representations. The results of the comparative analysis for the four main approaches examined are summarized in Table 1 [27], [28].

Table 1. Comparative analysis of various imputation methods using the GPCA framework

Methods	Strengths	Limitations
Integrated GPCA-MIGA	– Convergence accelerates as GPCA provides a structured search direction for the GA. – Missing value imputation aligns with the structural patterns of the data, ensuring that dimensionality reduction results retain the original relationships among variables and observations. – The process converges based on the relative change of imputed values, ensuring the stability and consistency of the results.	– The method is not robust to outliers.
MIGA+GPCA	– MIGA adaptively imputes missing values through evolutionary operations, producing complete data that better reflect the original distribution before dimensionality reduction. – Although the processes are independent, this sequential framework yields more accurate and stable low-rank representations than mean and median imputation.	– The method is not robust to outliers. – Imputing missing values without considering structural patterns makes dimensionality reduction inconsistent with the original data structure.
Mean imputation+GPCA	– Mean imputation requires no iteration, allowing the GPCA dimensionality reduction process to remain fast and stable. – Mean imputation maintains the stability of the distribution's central tendency, thereby preventing GPCA from being biased toward variables with a high proportion of missing values.	– The method is not robust to outliers. – Mean imputation disregards the correlations among variables, potentially distorting the original multivariate structure.
Median imputation+GPCA	– Median imputation requires no iteration, allowing the GPCA dimensionality reduction process to remain fast and stable. – The median imputation better represents the central value in a skewed distribution, making the low-dimensional representation obtained from GPCA more stable.	– The method is not robust to outliers. – Median imputation disregards the correlations among variables, potentially distorting the original multivariate structure.

As an additional evaluation method, Procrustes analysis was conducted on the original data and the approximated original data. The fundamental concept of this analysis is that the original data (A) is kept fixed, while the approximated original data ($\hat{A}$) is transformed to achieve the highest possible alignment between the two datasets, making them as similar as possible. This transformation includes translation, rotation, and scaling (dilation) processes [29]. The similarity between the two data is indicated by the R^2 value. An R^2 value close to 1 indicates a higher degree of similarity between the two datasets. The formula used to compute R^2 is presented as (6).

$$R^2 = 1 - \frac{D_{TRD}(\boldsymbol{A},\hat{\boldsymbol{A}})}{trace\ (\boldsymbol{AA^t})} \tag{6}$$

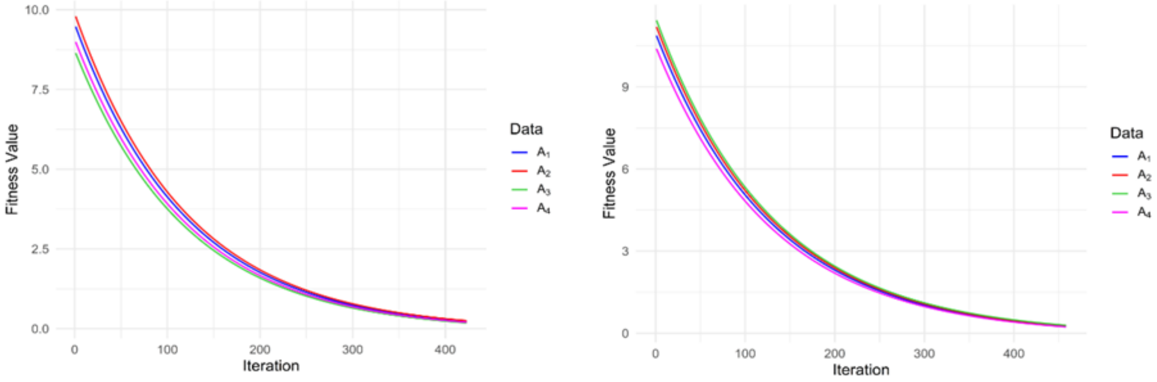
Where $D_{TRD}(A, \hat{A})$ represents the Procrustes distance used to measure the difference between the original data and the approximated original data [30], [31].

3. RESULTS AND DISCUSSION

3.1. Simulation results

The iterative process of the integrated GPCA-MIGA was terminated when the relative change of all imputed elements between two consecutive iterations was less than or equal to 0.001. This criterion indicates that the process had reached a stable or convergent state, where variations in the imputed elements between successive iterations became negligible. According to the simulation results, convergence was achieved at the 423rd iteration at a 5% level of missing data, at the 458th iteration at a 10% level, at the 486th iteration at a 15% level, and at the 527th iteration at a 20% level.

During each iteration, the MIGA generated several imputation candidates and evaluated them using the fitness function. The candidate with the smallest fitness value was selected as the optimal imputation result and integrated into the GPCA process for dimensionality reduction. The consistent decrease in fitness values with increasing iterations, as illustrated in Figure 3, indicates a gradual improvement in the dimensionality reduction process. Figure 3(a) presents the results for missing value percentages of 5%, and 10% as presented in Figure 3(b) only, as their convergence patterns reflect similar tendencies at other missing value levels.



(a) (b)

Figure 3. Fitness values at different percentages of missing values of (a) 5% and (b) 10%

The simulated data used in this study had an underlying structure of 25×5, comprising 25 clusters of observations and 5 clusters of variables. This configuration was used as a reference to assess how effective each approach was in performing dimensionality reduction on datasets containing missing values. The integrated GPCA-MIGA demonstrated the most consistent performance compared to the other methods. At missing value percentages of 5 and 10%, the integrated GPCA-MIGA preserved the original structure by producing dimensions of 25×5 across all replications, indicating its ability to reduce dimensionality stably even in the presence of missing values. As the percentage of missing values increased, the integrated GPCA-MIGA produced larger reduced dimensions, yet it remained more efficient than the other methods. In contrast, the other methods tended to generate larger dimensions even at relatively low percentages of missing values.

As shown in Table 2, the R^2 values from the Procrustes analysis between the original data and the approximated original data obtained from GPCA range from 0.991 to 0.993, indicating that GPCA can approximate the original data from its low-dimensional representation with a very small error level. In the data containing missing values, the integrated GPCA-MIGA produced the highest and most stable R^2 values across all levels of missingness compared to other methods. This indicates the ability of the method to approximate the original data from its low-dimensional representation with a very small error level. In addition, the R^2 values of the integrated GPCA-MIGA were closer to those of GPCA on complete data compared to other methods. In other words, the dimensionality reduction results obtained from the integrated GPCA-MIGA for data containing missing values still reflect the same structure and relationship patterns as the data before the missing values occurred.

Table 2. The R^2 values from the Procrustes analysis of different methods at various levels of missing value

Data pairs	Percentage of missing values (%)	GPCA	Integrated GPCA-MIGA	Mean imputation+GPCA	Median imputation+GPCA	MIGA+GPCA
A_1 and $\hat{A}_1$	0	0.991	-	-	-	-
	5	-	0.974	0.655	0.633	0.724
	10	-	0.962	0.628	0.612	0.706
	15	-	0.960	0.603	0.583	0.671
	20	-	0.944	0.562	0.532	0.646
A_2 and $\hat{A}_2$	0	0.993	-	-	-	-
	5	-	0.976	0.651	0.630	0.721
	10	-	0.961	0.625	0.610	0.703
	15	-	0.954	0.605	0.587	0.674
	20	-	0.941	0.565	0.531	0.641
A_3 and $\hat{A}_3$	0	0.992	-	-	-	-
	5	-	0.973	0.647	0.627	0.725
	10	-	0.959	0.621	0.607	0.698
	15	-	0.957	0.609	0.581	0.669
	20	-	0.946	0.557	0.536	0.641
A_4 and $\hat{A}_4$	0	0.991	-	-	-	-
	5	-	0.971	0.649	0.626	0.722
	10	-	0.963	0.623	0.609	0.696
	15	-	0.958	0.602	0.588	0.666
	20	-	0.940	0.555	0.535	0.643

Based on Table 3, GPCA applied complete data and yielded an average RMSE of 1.908, indicating a very low estimation error. All approaches exhibited an increase in RMSE as the percentage of missing values increased. In the 5% missing value scenario, the integrated GPCA-MIGA achieved the lowest average RMSE of 2.08, followed by MIGA+GPCA with 2.27, while mean imputation+GPCA with 2.42 and median imputation+GPCA with 2.54 resulted in larger estimation errors. This pattern was consistent at the 10, 15, and 20% missing value levels, with the integrated GPCA-MIGA consistently outperforming the other approaches, and its average RMSE remaining close to that of GPCA in every scenario.

GPCA yielded a very small standard deviation of 0.06 in the complete data, demonstrating stability in preserving the original data structure through dimensionality reduction. In the data with 5% missing values, the integrated GPCA-MIGA recorded a standard deviation of 0.17, which was more stable than MIGA+GPCA, mean imputation+GPCA, and median imputation+GPCA. This pattern remained consistent at a higher percentage of missing values. The standard deviation values for each method are presented in Table 3. Overall, these findings confirm that the integrated GPCA-MIGA is more accurate and more robust in reducing the dimensionality of data containing missing values. To complement the descriptive interpretation of the average RMSE and its standard deviation, statistical analysis (analysis of variance (ANOVA) and Tuckey test) was carried out.

Table 3. Average RMSE (standard deviation) of different methods at various levels of missing values

Methods	Percentage of missing values				
	0%	5%	10%	15%	20%
GPCA	1.91 (0.06)	-	-	-	-
Integrated GPCA-MIGA	-	2.08 (0.17)	2.57 (0.27)	3.01 (0.21)	4.28 (0.29)
Mean imputation+GPCA	-	2.42 (0.24)	2.69 (0.31)	3.32 (0.33)	4.47 (0.39)
Median imputation+GPCA	-	2.54 (0.26)	2.72 (0.32)	3.46 (0.33)	4.61 (0.40)
MIGA+GPCA	-	2.27 (0.26)	2.66 (0.32)	3.20 (0.32)	4.42 (0.34)

ANOVA was used to statistically evaluate the methods. This test aimed to determine the effects of the method, the percentage of missing values, and their interaction on the RMSE values. Based on Table 4, all factors have p-values <0.05, indicating significant differences in RMSE values among methods, among levels of missing values percentages, and in their interaction. Since the ANOVA results showed significant differences, a Tukey test was conducted to identify which groups differed significantly from each other.

Based on the results of the Tukey test, the integrated GPCA-MIGA outperformed the other methods in reducing the dimensionality of data containing missing values across various levels of missingness. Although the Tukey test computes all possible pairwise comparisons among the method groups. This study focuses only on comparisons involving the GPCA method as the baseline. Based on Table 5, the integrated GPCA-MIGA has the smallest mean difference compared to GPCA and is rejected at the 5% significance level, indicating that its performance is the closest to GPCA. In contrast, the other methods show much larger mean differences, indicating a significant increase in RMSE values compared to GPCA.

Table 4. Results of ANOVA on the effects of methods and percentages of missing values on RMSE

Source of variation	Df	Sum square	Mean square	F-value	p-values
Method	4	176.2	44.1	1782.52	2×10^{-16}
Percentage of missing values	3	1042.5	347.5	7346.64	2×10^{-16}
Method x Percentage of missing values	9	3.8	0.4	11.66	2.59×10^{-16}

Table 5. Results of the Tukey test comparing average RMSE differences among methods

Comparison	Mean difference	Adjusted p-value
GPCA with Integrated GPCA-MIGA	1.0776	0.0001
GPCA with MIGA+GPCA	1.2309	0.0001
GPCA with mean imputation+GPCA	1.3183	0.0001
GPCA with median imputation+GPCA	1.4263	0.0001

3.2. Application to real data

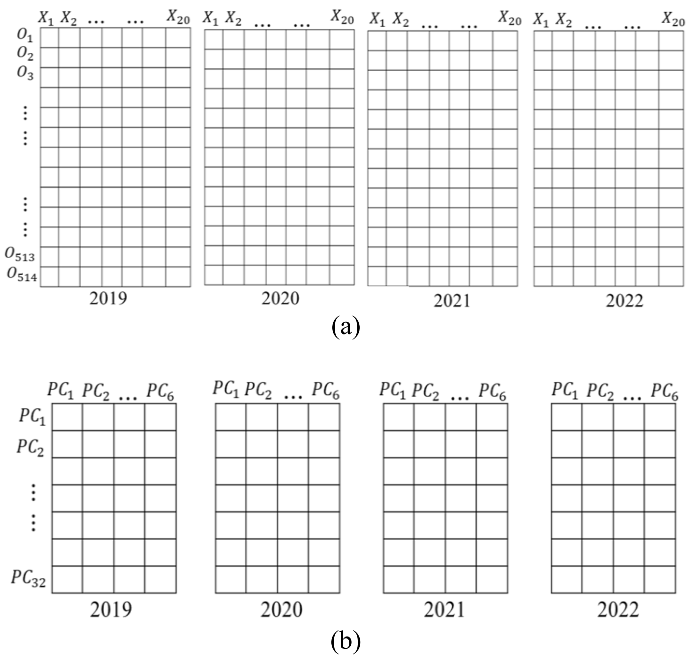
The percentage of missing values in each data is 5%. Based on the simulation results, the integrated GPCA-MIGA outperformed the other methods in reducing the dimensionality of data containing missing values. Therefore, the integrated GPCA-MIGA was subsequently applied to the empirical data. In this case, the integrated GPCA-MIGA stopped at the 401st iteration after meeting the criteria.

Based on the results obtained, an increasing number of the integrated GPCA-MIGA iterations leads to a decrease in fitness values. This indicates that the integrated GPCA-MIGA successfully optimizes the solutions progressively for each dataset. The RMSE value of the integrated GPCA-MIGA, which is 4.128 and close to the RMSE value of GPCA, which is 3.392, indicates that both methods can produce low-dimensional data with nearly equivalent error levels.

This suggests that the integrated GPCA-MIGA does not provide a significant difference in estimating the original data from the reduced data compared to GPCA. The dimension reduction process using matrices P and Q successfully retains most of the variation in the original data. In this process, data from 514 district/city were reduced to 32 principal components through matrix P, with a cumulative variance proportion of 88.11%. Meanwhile, data from 20 indicators were reduced to 6 principal components through matrix Q, with a cumulative variance proportion of 88.31%.

Figure 4 illustrated the visual of the matrix structure before (Figure 4(a)) and after (Figure 4(b)) dimensionality reduction. This method effectively summarizes the data into lower dimensions without significant loss of the total explained variance. The R^2 values from the Procrustes analysis for the integrated GPCA-MIGA are 0.889 for (A_1 and $\hat{A}_1$), 0.882 for (A_2 and $\hat{A}_2$), 0.874 for (A_3 and $\hat{A}_3$), and 0.871 for (A_4 and $\hat{A}_4$), demonstrating that the approximated original data from the integrated GPCA-MIGA closely the original data. The R^2 values range from 0.877 to 0.896 for GPCA. The small differences in R^2 values between the integrated GPCA-MIGA and GPCA for each data pair indicate that both methods exhibit similar performance in explaining the variability of the original data. The integrated GPCA-MIGA results are then visualized using the biplot approach.

As shown in Figure 5, the district/city are distributed into four quadrants, each defined by distinct sets of indicators. The district/city located in quadrant 1 are associated with indicators X_5, X_6, X_7, X_{10}, X_{14}, X_{16}, X_{17}, and X_{19}. Those in quadrant 2 are associated with indicators X_{12}, X_{15}, and X_{20}, while quadrant 3 are associated with indicators X_3, X_4, and X_{11}. Meanwhile, district/city in quadrant 4 are associated with indicators X_1, X_2, X_8, X_9, X_{10}, X_{13}, and X_{18}. The total cumulative variance explained by the biplot amounts to 52.05%, indicating that more than half of the data's information is represented by the extracted components. The district/city positioned close to one another on the biplot exhibit similar characteristics based on the principal components derived from the integrated GPCA-MIGA. For example, Bangka, Bangka Tengah, and Bangka Timur exhibit similar characteristics across the dimensions of human development.



(a)

(b)

Figure 4. The matrix structure of (a) before dimensionality reduction and (b) after dimensionality reduction

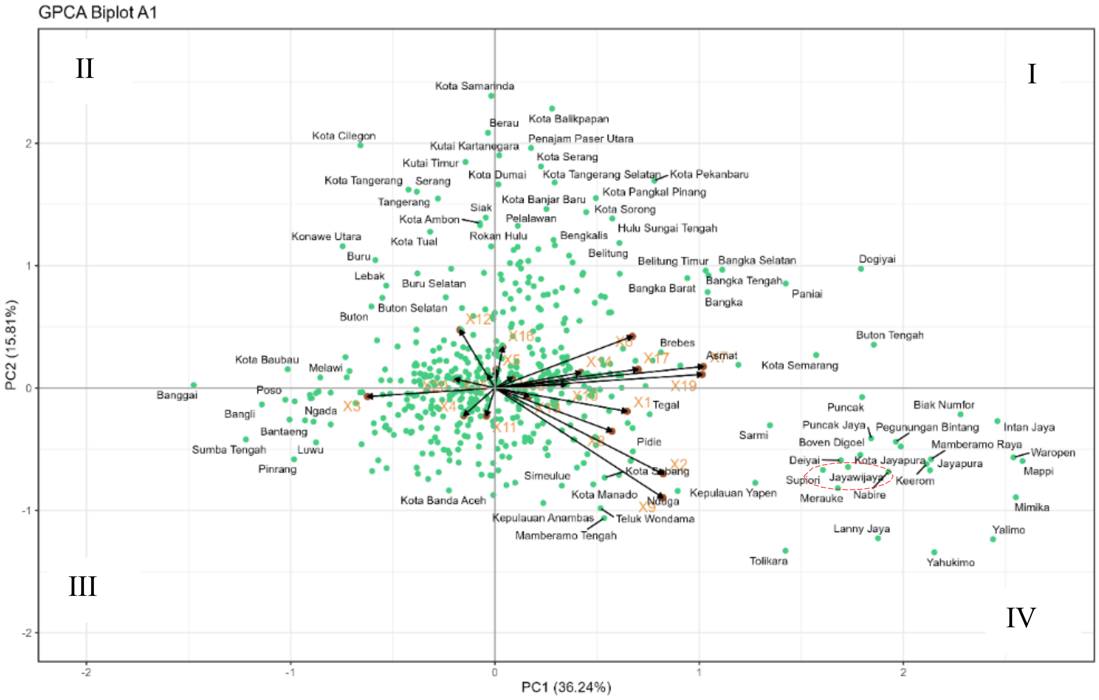


Figure 5. The integrated GPCA-MIGA results are illustrated in a biplot derived from the 2019 data

As shown in Figure 6, the district/city are distributed into four quadrants, each defined by distinct sets of indicators. The district/city located in quadrant 1 are associated with indicators X_5, X_6, X_8, X_9, X_{13}, X_{14}, X_{16}, X_{18}, and X_{19}. Those in quadrant 2 are associated with indicators X_4, X_{11}, X_{15}, and X_{20}, while quadrant 3 are associated with indicators X_2, X_3, and X_{10}. Meanwhile, district/city in quadrant 4 are associated with indicators X_1, X_7, and X_{17}. The total cumulative variance explained by the biplot amounts to 50.94%. The district/city such as Maros, Jeneponto, and Takalar exhibit similar characteristics across the dimensions of human development.

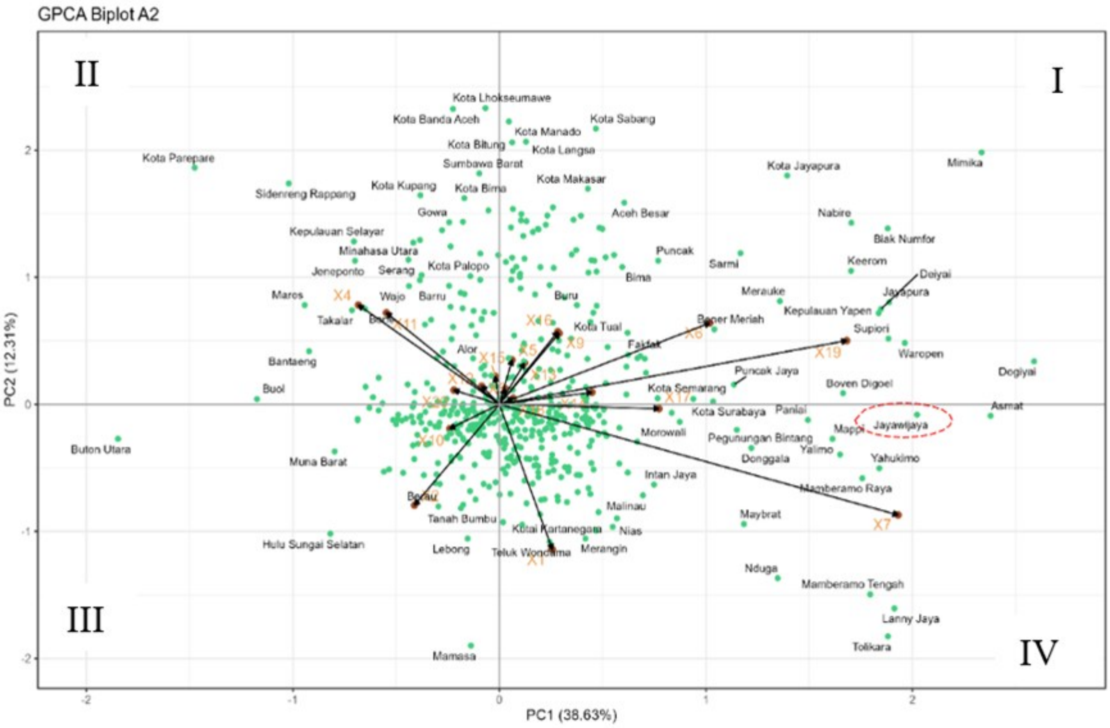


Figure 6. The integrated GPCA-MIGA results are illustrated in a biplot derived from the 2020 data

As shown in Figure 7, the districts/city are distributed into four quadrants, each defined by distinct sets of indicators. The districts/city located in quadrant 1 are associated with indicators X_3, X_6, X_7, X_9, X_{12}, X_{14}, and X_{18}. Those in quadrant 2 are associated with indicators X_1, X_2, X_{10}, X_{16}, and X_{19}, while quadrant 3 are associated with indicators X_4, and X_{17}. Meanwhile, districts/city in quadrant 4 are associated with indicators X_5, X_8, X_{11}, X_{13}, X_{15}, and X_{20}. The total cumulative variance explained by the biplot amounts to 51.62%. The districts/city such as Yalimo, Memberamo Raya, and Jayawijaya exhibit similar characteristics across the dimensions of human development.

As shown in Figure 8, the districts/city are distributed into four quadrants, each defined by distinct sets of indicators. The districts/city located in quadrant 1 are associated with indicators X_3, X_4, X_{10}, X_{15}, X_{18}, and X_{20}. Those in quadrant 2 are associated with indicators X_1, X_2, X_7, X_8, X_{11}, and X_{17}, while quadrant 3 are associated with indicators X_6, X_9, and X_{19}. Meanwhile, districts/city in quadrant 4 are associated with indicators X_{15}, X_{13}, and X_{16}. The total cumulative variance explained by the biplot amounts to 51.16%. The districts/city such as Buol and Sigi exhibit similar characteristics across the dimensions of human development.

The biplot results show that from 2019 to 2022, Jayawijaya was positioned quite far from the central point (0,0), indicating that its characteristics significantly differed from most other districts, which were more concentrated around the center. In 2019, Jayawijaya had adequate access to the indicator's households with clean drinking water sources (X_1) and households that have access to adequate drinking water (X_2). Additionally, the indicators of the gross participation rates at the elementary school (X_8) and junior high school level (X_9) were high in Jayawijaya. In 2020, Jayawijaya remained in Quadrant IV, with relatively high values for the indicators households with clean drinking water sources (X_1), school enrollment rates 16-18 years (X_7), and average wages of workers and employees per month (X_{17}). Although indicators X_2, X_8, and X_9 showed a decline in 2020, there was an increase in X_7 and X_{17}. In 2021, Jayawijaya was close to the indicators net participation rates at junior high school (X_{12}) and gross participation rates at the junior high school (X_9). This suggests an increase in participation at the junior high school level, reflecting a more equitable and inclusive focus on education. Additionally, X_7 remained high and moved even closer to Jayawijaya. The indicator households that do not have defecation (X_3) became one of the more closely associated with Jayawijaya in 2021. This indicates that there are significant problems with access to basic sanitation. In 2022, other indicators that were previously close to Jayawijaya, such as X_9, X_3 and X_{13}, appeared to have weaker associations or even moved further away. However, indicator of school enrollment rates for ages 16-18 years (X_7) remained high. This annual trend analysis for Jayawijaya illustrates both progress and persisting challenges, which can also be explored for another districts/city. On the one hand, Jayawijaya has achieved relatively good results in terms of education and the average wages of workers. On the other hand, there are still serious problems related to access to sanitation and adequate drinking water. For policymakers, this implies that the government needs to maintain and further improve achievements in education and average wages, while at the same time giving greater attention to the provision of household sanitation facilities and access to adequate drinking water.

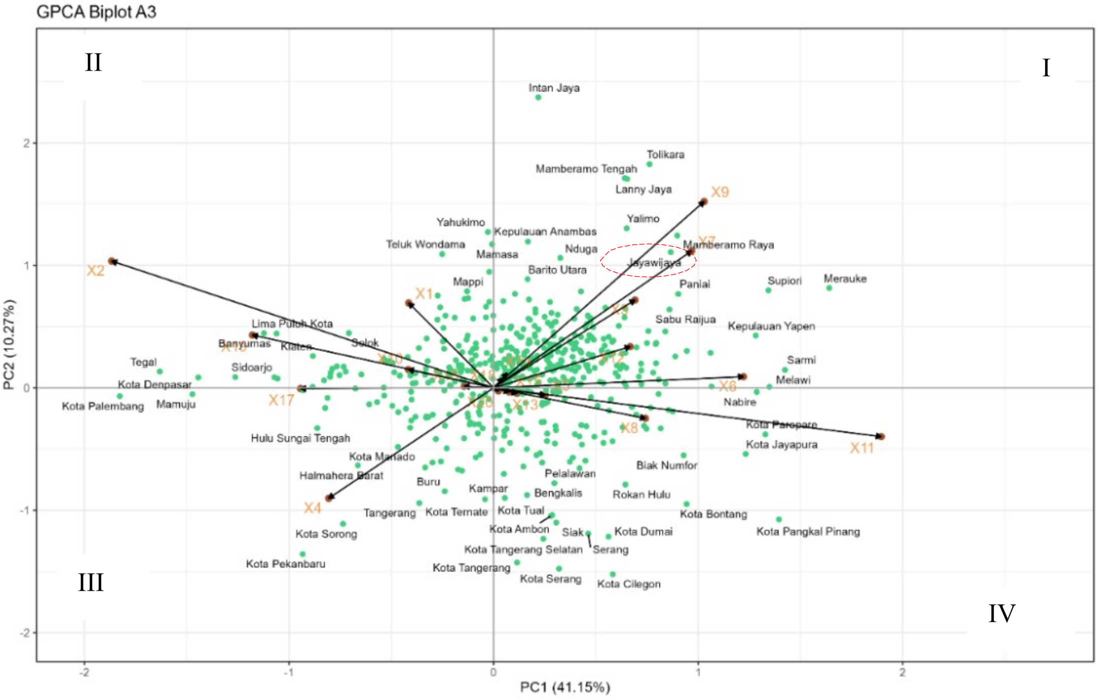


Figure 7. The integrated GPCA-MIGA results are illustrated in a biplot derived from the 2021 data

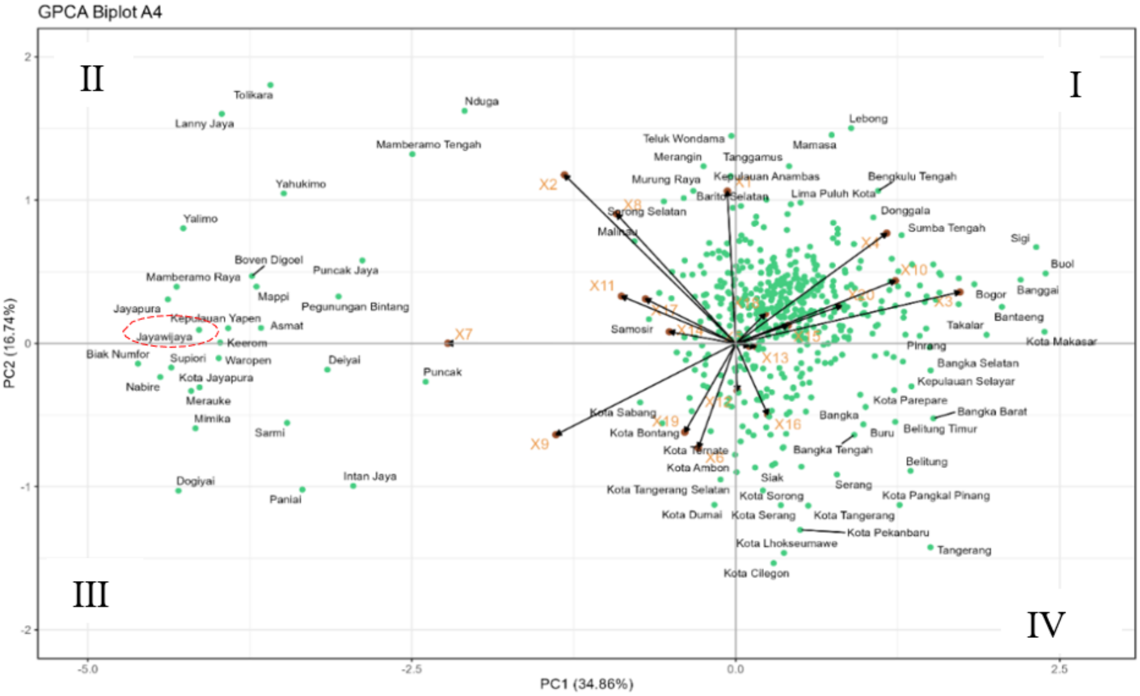


Figure 8. The integrated GPCA-MIGA results are illustrated in a biplot derived from the 2022 data

4. CONCLUSION

Based on the simulation results, the integrated GPCA-MIGA consistently outperformed other methods in reducing the dimensionality of data containing missing values, such as MIGA+GPCA, mean imputation+GPCA, and median imputation+GPCA. This method produced the lowest average RMSE and the highest R^2 values across all levels of missingness, indicating superior accuracy and robustness in estimating the original data from its low-dimensional representation. According to the biplot visualization results, the human development trends in Jayawijaya from 2019 to 2022 show progress in the indicator of school enrollment rates for ages 16-18 years. However, significant challenges remain in access to basic sanitation, as reflected in the high number of households that do not have defecation facilities in 2021. Thus, problems that occur in each district/city can be identified. The integrated GPCA-MIGA analysis and its visualization through biplots provide valuable insights for policymakers, as they allow a detailed examination of the characteristics of districts/city, thereby facilitating targeted interventions and more efficient resource allocation. This study has several limitations that require attention. First, the computation time was substantially longer when analyzing data with higher percentages of missing values. Second, the method is not yet sufficiently robust in the presence of outliers. Future research may focus on developing more efficient algorithms to handle datasets with high percentage of missing values and implementing robust approaches to reduce the method's sensitivity to outliers.

FUNDING INFORMATION

This study was supported by the Directorate General of Higher Education, Research, and Technology, Ministry of Education, Culture, Research, and Technology, through the main and derivative research program implementation contracts No: 027/E5/PG.02.00.PL/2024 and 22138/IT3.D10/PT.01.03/P/B/2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Fahrezal Zubedi		✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	
I Made Sumertajaya	✓	✓		✓		✓	✓		✓	✓		✓		
Khairil Anwar Notodiputro	✓	✓		✓		✓	✓		✓	✓		✓		
Utami Dyah Syafitri	✓	✓		✓		✓	✓		✓	✓		✓		

C	:	**C**onceptualization	I	:	**I**nvestigation	Vi	:	**Vi**sualization
M	:	**M**ethodology	R	:	**R**esources	Su	:	**Su**pervision
So	:	**So**ftware	D	:	**D**ata Curation	P	:	**P**roject administration
Va	:	**Va**lidation	O	:	Writing - **O**riginal Draft	Fu	:	**Fu**nding acquisition
Fo	:	**Fo**rmal analysis	E	:	Writing - Review & **E**diting			

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest related to this study.

DATA AVAILABILITY

The data that support this study were sourced from Statistics Indonesia (BPS) publications, which are publicly accessible on the official websites of each district/city; for example, data for Jayawijaya district can be accessed at https://jayawijayakab.bps.go.id/id.

REFERENCES

[1] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[2] A. F. M. Alkarkhi and W. A. A. Alqaraghuli, "Principal components analysis," in *Easy Statistics for Food Science with R*, vol. 12, no. 6, 2019, pp. 125–141, doi: 10.1016/B978-0-12-814262-2.00008-X.

[3] J. Ye, "Generalized low rank approximations of matrices," *Machine Learning*, vol. 61, no. 1–3, pp. 167–191, 2005, doi: 10.1007/s10994-005-3561-6.

[4] J. Ye, R. Janardan, and Q. Li, "GPCA: an efficient dimension reduction scheme for image compression and retrieval," *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 354–363, 2004, doi: 10.1145/1014052.1014092.

[5] J. Ye, R. Janardan, and S. Kumar, "Biological image analysis via matrix approximation," in *Encyclopedia of Data Warehousing and Mining, Second Edition*, Pennsylvania, United States: IGI Global, 2011, doi: 10.4018/9781605660103.ch027.

[6] H. Itoh, A. Imiya, and T. Sakai, "Dimension reduction and construction of feature space for image pattern recognition," *Journal of Mathematical Imaging and Vision*, vol. 56, no. 1, pp. 1–31, 2016, doi: 10.1007/s10851-015-0629-1.

[7] Y. Chen *et al.*, "Face identification with top-push constrained generalized low-rank approximation of matrices," *IEEE Access*, vol. 7, pp. 160998–161007, 2019, doi: 10.1109/ACCESS.2019.2947164.

[8] K. Li and G. Wu, "A randomized generalized low rank approximations of matrices algorithm for high dimensionality reduction and image compression," *Numerical Linear Algebra with Applications*, vol. 28, no. 1, 2021, doi: 10.1002/nla.2338.

[9] Y. Dong and C. Y. J. Peng, "Principled missing data methods for researchers," *SpringerPlus*, vol. 2, no. 1, pp. 1–17, 2013, doi: 10.1186/2193-1801-2-222.

[10] F. M. F. Lobato, V. W. Tadaiesky, I. M. D. Araújo, and A. L. Santana, "An evolutionary missing data imputation method for pattern classification," *GECCO 2015-Companion Publication of the 2015 Genetic and Evolutionary Computation Conference*, pp. 1013–1019, 2015, doi: 10.1145/2739482.2768451.

[11] Y. Sun, J. Li, Y. Xu, T. Zhang, and X. Wang, "Deep learning versus conventional methods for missing data imputation: a review and comparative study," *Expert Systems with Applications*, vol. 227, 2023, doi: 10.1016/j.eswa.2023.120201.

[12] J. C. F.- García, R. Neruda, and G. H.- Pérez, "A genetic algorithm for multivariate missing data imputation," *Information Sciences*, vol. 619, pp. 947–967, 2023, doi: 10.1016/j.ins.2022.11.037.

[13] BPS-Statistics Indonesia, *Human development index 2019*. Jakarta, Indonesia: BPS-Statistics Indonesia, 2020.

[14] BPS-Statistics Indonesia, *Human development index 2020*. Jakarta, Indonesia: BPS-Statistics Indonesia, 2021.

[15] BPS-Statistics Indonesia, *Human development index 2021*. Jakarta, Indonesia: BPS-Statistics Indonesia, 2022.

[16] BPS-Statistics Indonesia, *Human development index 2022*. Jakarta, Indonesia: BPS-Statistics Indonesia, 2023.

[17] T. B. Pedersen, S. Lehtola, I. F. Galván, and R. Lindh, "The versatility of the Cholesky decomposition in electronic structure theory," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 14, no. 1, 2024, doi: 10.1002/wcms.1692.

[18] S. Ahmadi and M. Rezghi, "Generalized low-rank approximation of matrices based on multiple transformation pairs," *Pattern Recognition*, vol. 108, 2020, doi: 10.1016/j.patcog.2020.107545.

[19] J. Shi, W. Yang, and X. Zheng, "Robust generalized low rank approximations of matrices," *PLoS ONE*, vol. 10, no. 9, 2015, doi: 10.1371/journal.pone.0138028.

[20] Z. Li, Z. Hu, F. Nie, R. Wang, and X. Li, "Multi-view clustering based on generalized low rank approximation," *Neurocomputing*, vol. 471, pp. 251–259, 2022, doi: 10.1016/j.neucom.2020.08.049.

[21] R. Nariswari, T. S. Prakoso, N. Hafiz, and H. Pudjihastuti, "Biplot analysis: a study of the change of customer behaviour on e-commerce," *Procedia Computer Science*, vol. 216, pp. 524–530, 2022, doi: 10.1016/j.procs.2022.12.165.

[22] W. Yan and N. A. Tinker, "Biplot analysis of multi-environment trial data: principles and applications," *Canadian Journal of Plant Science*, vol. 86, no. 3, pp. 623–645, 2006, doi: 10.4141/P05-169.

[23] A. I. Engloner and J. Podani, "A new statistical method for the comparison of biplots with the same objects and variables," *Ecological Indicators*, vol. 154, 2023, doi: 10.1016/j.ecolind.2023.110802.

[24] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimedia Tools and Applications*, vol. 80, pp. 8091–8126, 2021, doi: 10.1007/s11042-020-10139-6.

[25] M. Marghany, "Principles of genetic algorithm," in *Synthetic Aperture Radar Imaging Mechanism for Oil Spills*, Amsterdam, Netherlands: Elsevier, 2020, pp. 169–185, doi: 10.1016/B978-0-12-818111-9.00010-0.

[26] J. C. F. García, D. Kalenatic, and C. A. L. Bello, "Missing data imputation in multivariate data by evolutionary algorithms," *Computers in Human Behavior*, vol. 27, no. 5, pp. 1468–1474, 2011, doi: 10.1016/j.chb.2010.06.026.

[27] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of performance of data imputation methods for numeric dataset," *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, 2019, doi: 10.1080/08839514.2019.1637138.

[28] S. J. Hadeed, M. K. O'Rourke, J. L. Burgess, R. B. Harris, and R. A. Canales, "Imputation methods for addressing missing data in short-term monitoring of air pollutants," *Science of The Total Environment*, vol. 730, no. 1, Aug. 2020, doi: 10.1016/j.scitotenv.2020.139140.

[29] T. Bakhtiar and Siswadi, "Orthogonal procrustes analysis: its transformation arrangement and minimal distance," *International Journal of Applied Mathematics and Statistics*, vol. 20, no. M11, pp. 16–24, 2011.

[30] J. C. Gower, "Procrustes analysis," in *International Encyclopedia of the Social & Behavioral Sciences*, vol. 50, no. 11, Amsterdam, Netherland: Elsevier, 2015, pp. 79–81, doi: 10.1016/B978-0-08-097086-8.43078-3.

[31] J. L. Kern, "On the correspondence between procrustes analysis and bidimensional regression," *Journal of Classification*, vol. 34, no. 1, pp. 35–48, 2017, doi: 10.1007/s00357-017-9224-z.

BIOGRAPHIES OF AUTHORS

Fahrezal Zubedi earned bachelor's degree from Gorontalo State University and master's degree in Mathematics from University of Indonesia. He is a doctoral student of Statistics and Data Science at IPB University. His research interests include mathematical computing, statistical computing and statistical modeling. He is a lecturer in the statistics study program at Gorontalo State University. He can be contacted at email: zubedifahrezal@apps.ipb.ac.id.

I Made Sumertajaya is a professor in the Statistics and Data Science Study Program at IPB University. He obtained bachelor's, master's, and doctoral degrees in Statistics from IPB University. His thesis title is “Recovery of inter block and interaction effects on multi locations and multi response”. He is actively engaged in conducting competitive research and provides supervision for both undergraduate and postgraduate students. His primary research interests cover several advanced statistical areas, specifically mixed models, experimental design, time series analysis, spatial models, and statistical modelling. Over the last ten years, the key subjects he has taught include multivariate analysis, experimental design, statistical methods, panel data analysis, and advanced statistical inference. He can be contacted at email: imsjaya@apps.ipb.ac.id.

Khairil Anwar Notodiputro is a professor in the Statistics and Data Science Study Program at IPB University. He obtained bachelor's and master's degrees from IPB University, and earned a Ph.D. in Statistics from Macquarie University, Australia. His thesis title is “Modified fisher scoring algorithms for image reconstruction from projection”. He is actively engaged in conducting competitive research and provides supervision for both undergraduate and postgraduate students. His primary research interests cover several advanced statistical areas, specifically mixed models, small area estimation, time series analysis, and statistical machine learning. Over the last ten years, the key subjects he has taught include time series analysis and forecasting, advanced data analysis, statistical analysis, statistical methods, and data science. He can be contacted at email: khairil@apps.ipb.ac.id.

Utami Dyah Syafitri earned bachelor's and master's degrees from IPB University, and a Ph.D. in Applied Economics from the Universiteit Antwerpen, Belgium. Her thesis title is “Optimal design of mixture experiments”. She is actively doing competitive research and supervising undergraduate and postgraduate students. Her research interests include experimental design, optimal design, and statistical modeling. Over the last ten years, she has taught the following subjects: statistical methods, experimental design, and probability theory. She can be contacted at email: utamids@apps.ipb.ac.id.