

Handling Missing Data in Clustering: A Comparative Evaluation of Six Methods

Xinrui Hu

Emory University

Jiaxi Geng

Emory University

Hanqi Luo

Emory University

Melissa Young

Emory University

Yi-An Ko

`yi-an.ko@emory.edu`

Emory University

Research Article

Keywords: k-means, clustering, missing data, incomplete data, imputation

Posted Date: June 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8371651/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Handling Missing Data in Clustering: A Comparative Evaluation of Six Methods

Xinrui Hu¹, Jiaxi Geng¹, Hanqi Luo², Melissa F Young², Yi-An Ko (corresponding author)^{1,*}

¹Department of Biostatistics and Informatics, Rollins School of Public Health, Emory University,
Atlanta, GA, USA.

²Hubert Department of Global Health, Rollins School of Public Health, Emory University,
Atlanta, GA, USA.

*Corresponding author. yi-an.ko@emory.edu

Abstract

Background

Most clustering methods require complete multivariate data as input; however, missing data are common in observational studies. Although several approaches have been developed to handle missing data in clustering analysis, comprehensive evaluations of their performance remain limited.

Methods

We compared six approaches for handling missing data in clustering: complete case analysis (CCA), k-POD, the framework for multiple imputation in cluster analysis (F.MICA), stacked multiple imputation (stackMI), cluster analysis with missing values via multiple imputation (clusterMI), and multiple imputation clustering analysis (MICA). Four imputation-based approaches were included due to their flexibility across diverse scenarios. Simulation studies assessed performance under varying proportions of missing data, numbers of imputations, and covariance structures, using percentage agreement (PA), adjusted Rand index (ARI), and normalized mutual information (NMI) as evaluation metrics. Finally, all six methods were applied to data from the JiVitA-3 trial to explore nutrient profiles among pregnant women.

Results

ClusterMI consistently identified the correct number of clusters across all scenarios. For CCA, MICA, and stackMI, the proportion of correctly identified clusters declined as missingness increased. F.MICA correctly identified clusters over 90% of the time under low-variance data structures but failed under high-variance conditions. In terms of clustering accuracy, all methods except CCA and stackMI maintained high accuracy (>90%) despite increasing missingness and varying covariance structures. The number of imputations had minimal impact on performance.

In the JiVitA-3 application, all methods identified two clusters corresponding to higher versus lower micronutrient concentrations.

Conclusions

While each method has limitations, clusterMI and F.MICA (when the correct number of clusters is specified) demonstrated relatively robust performance. k-POD offered computational advantages when the number of clusters was known. These findings provide practical guidance for researchers selecting clustering methods for epidemiological and clinical studies with missing data.

Keywords: k-means, clustering, missing data, incomplete data, imputation

Introduction

Clustering analysis is a useful method for identifying patterns within data across various domains by grouping data points based on their similarity or proximity(1–4). For instance, clustering analysis can be used to classify pregnant women with similar micronutrient profiles, facilitating the development of targeted intervention strategies for those with sub-optimal micronutrient biomarker concentrations. These identified patterns can further be analyzed to explore their associations with functional outcomes.

Most existing clustering methods require complete data as the input, which presents a significant challenge in observational studies. Missing data is prevalent in observational studies due to non-response, incomplete questionnaires, errors, and/or poor sample quality. Complete case analysis, which involves excluding observations with missing or incomplete data, is commonly used in clustering analysis. However, complete case analysis becomes problematic when a substantial proportion of the data is missing or when the clustering results are influenced by the missing values.

Numerous methods have been developed to address missing data in clustering analysis. These methods can be generally categorized into two strategies. One approach involves imputing missing data first, performing clustering on the imputed dataset, and then combining the results. The other strategy does not require imputation and directly handles missing values during the clustering process.

By imputing missing data, various algorithms have been recommended. Basagaña et al. (2013) proposed a structured framework to integrate multiple imputation into cluster analysis, which incorporated a variable reduction procedure and the selection of the final number of clusters(5). Plaehn et al. (2019) proposed to cluster data after stacking multiply-imputed datasets

together and assign the final cluster membership based on the most frequently assigned cluster among the imputed datasets(6). Audigier et al. (2023) used partition pooling to improve accuracy and calculated the instability to obtain the optimal cluster number(7). Recently, Lee et al. (2023) proposed a comprehensive two-stage framework for clustering incomplete data. The first stage involves clustering each of the imputed datasets, and in the second stage clustering analysis is applied to the initial cluster centers to identify the final cluster configuration(8). Storlie et al. (2019) used a Dirichlet Process Mixture model for clustering that allows for an unknown number of clusters. Under the modified split-merge Markov Chain Monte Carlo (MCMC) algorithm for sampling, missing data are sampled from the posterior distribution. Additionally, the framework is designed to handle mixed variable types and incorporate variable selection, ensuring that only meaningful variables are included in the clustering analysis(9). Lastly, Boluki et al. (2019) proposed an optimal clustering method that incorporates the generation of missing values within the framework of the original random labeled point process (RLPP). This approach creates a new RLPP that generates the observed points, including those with missing values. The optimal clustering solution is achieved by marginalizing out the missing features(10). Zhang et al. (2016) optimized the fuzzy c-means clustering, in which each data point is assigned a probabilistic membership across clusters, by iteratively updating cluster prototypes, membership probabilities, and imputing missing values using a maximum likelihood criterion(11).

Without imputation of missing data, several approaches have been developed to handle missing data directly in clustering algorithms. Afridi et al. (2018) proposed a game-theoretic rough set model to deal with uncertainty in clustering due to missing values. They addressed the challenge of determining thresholds for making three-way decisions in clustering, namely, whether to include, exclude, or defer an object's inclusion in a cluster. This model balances

accuracy and generality in cluster formation to improve how thresholds are set and uncertain objects are classified(12). Chai et al. (2016) proposed k-POD, an extension of k-means, that employed a majorization-minimization approach to update cluster assignments and centroids by focusing on the observed data(13).

Although various methods for clustering with incomplete data have been proposed, comprehensive comparisons of these approaches, along with user guidelines, remain scarce. Himmelspach et al. (2010) compared seven different approaches focusing on k-means and fuzzy c-means with imputation applied before clustering. They found that the missing mechanism impacts the accuracy of classification. However, none of the methods they examined included a built-in mechanism for determining the optimal number of clusters(14). Boluki et al. (2018) compared five clustering methods for handling missing values, including approaches with and without imputation. However, their comparison was based on a relatively small sample (20 to 70 data points), which could reduce classification accuracy(10). Storlie et al. (2018) compared three different approaches, each relying on different assumptions regarding data independence and distribution, which may not hold in real-world datasets. Moreover, the accuracy and complexity of these methods are likely influenced by data availability(9). Conducting such comparisons can offer researchers important insights when choosing the most suitable method.

In this project, six methods for handling missing data in clustering analysis were introduced and compared. While k-means clustering was employed to demonstrate their implementation, most methods are not restricted to a specific clustering algorithm. The selection of these methods was guided by their adaptability to heterogeneous datasets and their minimal data pre-processing requirements. Furthermore, these approaches are readily applicable across various research contexts. Specifically, these selected methods included complete case analysis,

k-POD(13), stacked multiple imputation(6), multiple imputation clustering analysis(8), cluster multiple imputation(7), and a framework for multiple imputation in cluster analysis(5). These methods were evaluated through simulation studies across varying proportions of missing data, numbers of imputations, and covariances. To demonstrate their applicability, the methods were applied to the JiVitA-3 trial micronutrient dataset for pregnant women(15,16).

The remainder of the paper is structured as follows. Section 2 presents an overview of each selected approach for handling missing data in clustering analysis. Section 3 details the simulation settings and presents the results for each approach, followed by a comparison of the results. Section 4 describes the data application. A brief conclusion and discussion are provided in Section 5. The R code implementing the clustering techniques is provided in the Appendix.

Methods

We summarized 6 selected methods for handling missing data in cluster analysis. Except for k-POD, other methods impose no restrictions on the choice of clustering algorithm.

Complete Case Analysis

This method includes only complete cases in the analysis. Individuals or records with missing data in any variables of interest are removed from the analysis.

K-POD

Chi et al. (2016) proposed *k-POD*, which is an extension of k-means clustering. It handles missing data by iteratively updating cluster assignments and centroids using only the observed data without the requirement of assumptions about the missing data patterns.

The projection operator ($P_{\Omega}(Y)$) is created first for the original data matrix (Y) onto a subset of the indices that correspond to the observed entries Ω , which is considered as complete. Through this step, a new data matrix that contains only observed data points is created. Next, k-

means is applied to this matrix. The missing values are imputed by the cluster centroids that they belong to. K-means is applied again to the new/updated dataset. This process of clustering followed by missing value updates is iteratively repeated under the Majorization-Minimization (MM) algorithm until the cluster assignments converge.

$$[P_{\Omega}(Y)]_{ij} = \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed} \\ 0, & \text{if } y_{ij} \text{ is missing} \end{cases}$$

where y_{ij} is the value from column i and row j in matrix Y .

Stacked Multiple Imputation (stackMI)

Plaehn (2019) proposed this method, which combines multiple imputed data and applies clustering analysis on it. (6). Specifically, multiple imputation is implemented to create M complete data sets ($X_{N \times J}^m$, where $m = 1, 2, \dots, M$). These imputed datasets are then stacked to form a single combined dataset ($\tilde{X}$),

$$\tilde{X} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(M)} \end{bmatrix}$$

to which the chosen clustering algorithm is applied. Because each individual show in M imputed datasets, they may be assigned to different cluster memberships across imputation. The cluster that has the highest frequency being assigned is chosen as the final cluster for each individual.

Multiple Imputation Clustering Analysis (MICA)

Lee et al. (2023) proposed a comprehensive framework for clustering incomplete data, consisting of two stages(17). In the first stage, multiple imputation is performed to generate M complete data sets, denoted as $X_{N \times J}^m$ for $m = 1, \dots, M$. Subsequently, clustering is conducted separately on each imputed dataset. The number of clusters for each dataset is determined using the silhouette width and is allowed to vary across datasets ($k_1, k_2, k_3, \dots, k_M$).

In the second stage, the cluster centers obtained from the first stage are aggregated into a single matrix $Y_{c \times J}$, where $c = \sum_{i=1}^M k_i$, and J represents the number of variables of interest. Clustering is then applied to matrix Y to determine the final cluster centers $K_{n \times J}$, where n is the final cluster number. Each individual is subsequently assigned to the nearest cluster in K . For individuals with incomplete data, initial assignment may vary across imputations. The final cluster assignment is determined based on the most frequently assigned cluster.

Cluster Analysis with Missing Values by Multiple Imputation (clusterMI)

Audigier and Niang (2022) proposed a method that leverages instability to determine the optimal number of clusters and integrates multiple clustering results into a consensus partition(7). The process begins with multiple imputation to handle missing values, generating M complete data sets. Clustering is then performed on each imputed dataset for various numbers of cluster (k). For each k , instability is calculated to measure variability of clustering results both across imputations and within each imputation using bootstrap resampling.

Within-Imputation Instability ($\bar{U}$)

$$\bar{U} = \frac{1}{M} \sum_{m=1}^M U_m^{boot}$$

Between-Imputation Instability (B)

$$B = \frac{1}{M^2} \sum_{m=1}^M \sum_{m'=1}^M \delta(\Psi_m, \Psi_{m'})$$

where $\delta(\Psi_m, \Psi_{m'})$ measures the agreement between clustering results by evaluating how consistently data points are assigned to clusters.

Total Instability (T)

$$T = \bar{U} + B$$

The instability values for each k are then compared, and the k with the lowest instability is selected as the optimal/final number of clusters (k_{fin}). Under the k_{fin} , clustering is applied to each imputed dataset to obtain individual clustering results. To produce the final clustering solution, non-negative matrix factorization (NMF) is used to exam the co-occurrence of data points within the same cluster across different partitions.

A Framework for Multiple Imputation in Cluster Analysis (F.MICA)

Basagaña et al. (2013) proposed a structured approach to integrate multiple imputation into cluster analysis(5). The method introduces a criterion for selecting the optimal number of clusters and incorporates a relabeling process to ensure consistency across datasets.

The procedure begins by generating M complete datasets through multiple imputation. The goodness-of-fit criterion ($CritCF$) is applied to determine the optimal number of clusters. For each imputed data set $X_i (i = 1, 2, 3, \dots, M)$, backward selection is used to identify the optimal cluster number k_{fin} for subsequent steps. Under the k_{fin} , clustering algorithm is applied to each imputed dataset.

To ensure consistency, a relabeling process is applied, where the labels from the first dataset are used as a reference. All possible label permutations for the remaining datasets are computed, and the agreement with the first dataset is evaluated using Cohen's kappa. The permutation with the highest kappa is selected, and all other datasets are relabeled accordingly. Finally, individuals are assigned to the cluster to which they are most frequently assigned across all imputations.

$$CritCF = [(\frac{2m}{2m+1})(\frac{1}{1+W/B})]^{\frac{\log_2(k+1)+1}{\log_2(m+1)+1}}$$

where m is the number of variables, k is the number of clusters, and W and B are the within- and between-cluster inertias (see below), respectively. We suppose that there are n subjects ($s_1, s_2, \dots, s_n$) and k clusters ($C_1, C_2, \dots, C_k$), with corresponding cluster centroids ($c_1, c_2, \dots, c_k$).

$$W = \sum_{i=1}^k \sum_{s \in C_i} \delta(c_i, s)$$

$$B = \sum_{i=1}^k |C_i| \delta(c_i, g)$$

where $\delta(c_i, s)$ is the distance between s_1 and s_2 , $|C_i|$ is the number of individuals in cluster C_i , and g is the centroid of the entire dataset.

Method Comparisons

The following criteria were used to evaluate and compare methods in the simulation study.

- a. *Percentage Agreement (PA)* measures the percentage of data points that are assigned to the same clusters in two clustering results, reported as 0% to 100%(6).

$$PA = \frac{\text{Number of agreeing pairs}}{\text{Total number of pair of points}}$$

- b. *Adjusted Rand Index (ARI)* assesses the similarity between two clustering results while accounting for chance(18). It compares how well two different clustering assignments (e.g., from different algorithms or methods) agree with each other, adjusting for the possibility of random assignments.

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

where

$$RI \text{ (Rand Index)} = \frac{TP + TN}{TP + TN + FP + FN}$$

TP (true positive) refers to the number of data point pairs that are correctly clustered together in both the true labels and the predicted clusters. True Negative (TN) refers to the number of data point pairs that are correctly assigned to different clusters in both the true labels and the predicted clusters. False Positive (FP) refers to the number of data point pairs that are incorrectly clustered together in the predicted clusters but are placed in different clusters in the true labels. False Negative (FN) refers to the number of data point pairs that are incorrectly placed in different clusters in the predicted clusters but are in the same cluster in the true labels.

The ARI ranges from -1 to 1, higher values indicate better agreement. ARI is more reliable than PA, especially when the clusters sizes differ substantially.

Normalized Mutual Information (NMI) is a normalization of the Mutual Information (*MI*) score, scaling the results between 0 (no mutual information) and 1 (perfect correlation)(19,20). It measures the similarity between two clustering results. *MI*, derived from information theory, quantifies the amount of shared information between two clusters and can detect nonlinear correlations between them.

$$NMI(U, V) = \frac{2 \times MI(U, V)}{H(U) + H(V)},$$

$$where MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i, j) \log \left(\frac{P(i, j)}{P(i)P(j)} \right),$$

$$H(U) = - \sum_{i=1}^{|U|} P(i) \log P(i)$$

U is the true cluster, V is the predicted cluster. $P(i, j)$ is the joint probability that a data point is in cluster i in U and cluster j in V . $H(U)$ and $H(V)$ quantifies the uncertainty or randomness of the cluster assignments.

Simulation Study

A simulation study was conducted to evaluate the performance of the above six approaches for handling missing data in clustering analysis under various scenarios. The number of clusters ($K = 4$) and sample size ($n = 1000$) were fixed, while the percentage of missingness was varied (20% and 50%). For methods that require missing data imputation, $m = 20$ and 50 imputations were implemented using the R **mice** package, assuming data are missing at random (MAR). Predictive mean matching was used as imputation model.

The simulation process consisted of two stages. In the first stage, we assigned imbalanced cluster sizes for the ease of labeling cluster memberships, using the ratio $c_1 : c_2 : c_3 : c_4 = 2 : 9 : 6 : 5$. In the second stage, for each cluster k , seven variables were selected from a multivariate normal distribution with mean vector μ_k and covariance matrix Σ_k . We designed two covariance matrices for comparisons, $\Sigma_{1,k}$ (larger variances) and $\Sigma_{2,k}$ (smaller variances).

Missingness was generated under MAR using the R **mice** package. One of seven variables was randomly selected to contain missing values, while the remaining six served as weighted predictors ($w_1 = 0.5$, others = 0). A weighted sum score (wss) was calculated to determine the missing data pattern,

$$wss_i = w_1 \cdot y_{1,i} + w_2 \cdot y_{2,i} + \dots + w_j \cdot y_{j,i},$$

where $y_{j,i}$ is the value of j th variable for the i th individual, w_j is the weight value for the j th variable. A total of 100 simulations were performed under each scenario.

Figure 1 shows the percentage of correct cluster number (PCCN) identified by each method across 100 simulations under two covariance scenarios (scenario 1: larger variances, scenario 2: smaller variances) when 20% and 50% of the data are missing. As missingness increases from 20% to 50%, the capability of CCA to identify the correct cluster number decreases from 70-80% to around 50% (Supplementary Table 1). F.MICA identifies the correct cluster number more than 90% of the time under Scenario 2 (not affected by the proportion of missing data) but fails when data variances increase under Scenario 1. In contrast, the PCCN for MICA and stackMI are substantially higher under Scenario 1 compared to Scenario 2.

Figure 2 presents the distribution of ARI values (shown as boxplots) for each method among the simulated datasets in which the correct number of clusters was identified. The PA for CCA decreases from 95% to 70% as the proportion of missing data increases from 20% to 50%. In contrast, other methods maintain accuracy above 90% despite the increased missingness, demonstrating robustness to covariance structures. The performance of stackMI is comparatively poor. Overall, the number of imputations does not appear to influence performance. Supplementary Table 1 displays the summary of clustering performance metrics, including ARI, NMI, and PA across all methods.

Application

We applied each of the methods to the micronutrient data from the JiVitA-3 trial in pregnant women(15,16). The JiVitA-3 trial was a randomized controlled study that evaluate the efficacy of a daily multiple micronutrient supplement versus iron-folic acid, provided from early gestation through 12 weeks postpartum, for improving birth outcomes and reducing neonatal mortality. Incident pregnancies were ascertained by 5-weekly community surveillance activities at the JiVitA project site in the Gaibandha and Rangpur Districts of northwestern Bangladesh,

with ~44,000 women enrolled and ~28,000 live births (16). More intensive data collection was done in a limited geographic area, including blood collection for micronutrient biomarker assessments, conducted at a median of 10.0 (IQR: 8.1-12.9) weeks of gestation and prior to supplementation, as described (15).

In this analysis, we included data from 1,696 participants with micronutrient biomarker data. A total of 6 micronutrient biomarkers (plasma ferritin, folate, retinol, vitamin B12, 25-hydroxyvitamin D, and zinc) and 1 inflammatory biomarker (alpha1-acid glycoprotein [AGP]) measured at the enrollment visit were included in this clustering analysis. Among the included participants, 1324 women had complete data. Given that some micronutrient biomarkers are affected by the acute-phase response(21–23), plasma ferritin was adjusted for inflammation by AGP using the BRINDA adjustment method(24). The percentage of missing data across 7 clustering variables ranged from 1.9% to 13.9% (Table 1). Maternal age, maternal education level, household socioeconomic status (SES), ferritin, folate, retinol, vitamin B12, 25-hydroxyvitamin D, zinc and AGP were used for Multiple Imputation by Chained Equations (MICE)(25) under the MAR assumption.

Figure 3 displays the k-means cluster centers of the seven input variables using the six methods. CCA, MICA, clusterMI and F.MICA separated the women into two clusters, reflecting similar patterns of higher versus lower micronutrient levels. For k-POD, the number of clusters must be pre-specified. In the two-cluster model, the separation between groups was clear, and as the number of clusters increased, the distinctions became more refined, particularly for vitamin B12, vitamin D, and folate. StackMI also identified two clusters, but the cluster centers were substantially different from others with smaller differences. Specifically, the cluster separation for zinc, vitamin D, vitamin B12, and retinol was minimal. (Supplementary Table 4 and 5).

Discussion

We compared six methods to handle missing values in clustering analysis. Under the MAR assumption, k-POD, MICA, F.MICA, and clusterMI demonstrated high ARI, NMI, and PA values across scenarios with up to 50% missing data, given that the correct number of clusters was identified. Although stackMI identified the correct number of clusters more than 70% of the time under our simulation settings, its overall performance remained poor. When the proportion of missing data increased to 50%, only F.MICA identified the correct number of clusters more than 95% of the time, and this occurred only under one covariance scenario.

Table 2 summarizes the advantages and disadvantages of each method. CCA and k-POD are both computationally efficient since there is no imputation involved. CCA performs well when missing data are low but loses accuracy as the proportion of missingness increases. When clusters are more separated, k-POD is more likely to detect the correct clusters. Similar findings have been reported by others, indicating that the performance of k-POD is affected by the underlying data structure(10). However, k-POD lacks a built-in process of data-driven mechanism for selecting the optimal number of clusters and does not provide flexibility to use clustering algorithms other than k-means.

In contrast, methods that involve imputing missing data (F.MICA, clusterMI, MICA and stackMI) are considerably more computationally intensive. All of these methods, except clusterMI, allow the use of clustering algorithms other than k-means. F.MICA appears to be the most comprehensive approach, incorporating features such as variable selection while consistently demonstrating strong performance even with higher proportions of missing data. Unlike other methods, F.MICA uses CritCF as the criterion to select the number of clusters, which yields better results compared to using the silhouette width. However, F.MICA fails to

identify the true number of clusters when input variables are highly correlated, which may also be influenced by the choice of criterion.

StackMI may not adequately account for the heterogeneity among imputed datasets, leading to a poor performance across all scenarios(7). It is possible that, as the number of imputations increases, the combined imputed dataset may become more variable, resulting in imprecise clustering outcomes. Furthermore, this algorithm requires stacking all imputed datasets into a single, large dataset, substantially increasing computational demands and complicating the application of standard cluster selection criteria. MICA can perform well when the correct number of clusters is known, however, reliance on silhouette width may reduce its effectiveness when missingness is high. Moreover, it operates within the traditional probability framework and directly partitions observations into clusters without accounting for the uncertainty introduced imputation, which may result in misclassification. In contrast, clusterMI employs NMF to pool clustering results across multiple imputations. This strategy enhances robustness against clustering instability and performs better with small sample sizes or highly variable data(27).

There are several limitations. Additional factors, such as alternative clustering algorithms and missing data mechanisms, were not considered. To facilitate a more feasible comparison across multiple methods, we focused primarily on k-means. K-means has limitations including sensitivity to random initialization of centroids, which can lead to unstable convergence, and the requirement to pre-specify the number of clusters. Future studies should evaluate these methods under additional simulation scenarios, such as varying the number of clusters, to provide a more comprehensive understanding of their performance.

This study provides a comprehensive evaluation of six methods for handling missing data in clustering analysis. By systematically comparing their performance across various scenarios,

we found that clusterMI and F.MICA (when the correct number of clusters is specified) are the most robust and versatile approaches, particularly when the proportion of missing data is high. K-POD is considered as a computationally efficient alternative when the number of clusters is known; however, it is restricted to the k-means clustering framework. In our data application with up to 14% missing data, using different approaches did not result in appreciable differences in micronutrient profiles. The findings from our study offer practical guidance for researchers, helping them make informed decisions when selecting appropriate and effective clustering methods in the presence of missing data.

408 Table 1. Summary statistics and missingness of variables in the JiVitA-3 data

Variable	No. of Observations	Percentage or median (lower quartile, upper quartile)	Missingness (%)
Maternal Age (years)	1696	20.00 (18.00, 25.00)	0%
Maternal Education Level	1691		0.29%
None/Elementary School	522	30.78%	
Middle School	1069	63.03%	
High School	60	3.54%	
University/Trade School	40	2.36%	
Socioeconomic Status	1694		0.12%
Low	669	39.44%	
Median	657	38.74%	
High	368	21.70%	
Ferritin ($\mu\text{g/L}$)	1451	62.45 (38.79, 93.42)	14.4%
Adjusted Ferritin ($\mu\text{g/L}$)	1451	56.39 (35.05, 83.45)	14.4%
Folate (nmol/L)	1503	16.50 (12.27, 21.52)	11.4%
Retinol ($\mu\text{mol/L}$)	1588	1.06 (0.88, 1.26)	6.4%
Vitamin B12 (pmol/L)	1461	175.61 (133.54, 243.45)	13.9%
Vitamin D (nmol/L)	1495	45.27 (38.29, 53.93)	1.9%
Zinc ($\mu\text{g/L}$)	1474	714.5 (631.9, 825.1)	13.1%
Alpha(1)-Acid Glycoprotein (g/L)	1504	0.72 (0.56, 0.90)	11.3%

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424 Table 2. Summary of advantages and disadvantages of 6 methods

Methods	Advantages	Disadvantages
CCA ^a	Easy to implement.	Ignores incomplete data, which may lead to biased results.
k-POD	Fastest with low misclassification rate.	Need to provide the number of clusters, which may be incorrect because of using complete case data.
MICA ^b	Comprehensive framework supporting various clustering methods and cluster number selection.	Computationally intensive.
StackMI ^c	Straightforward to implement.	Increasing the number of imputations or the sample size can rapidly expand the data matrix, prohibiting subsequent clustering. Performs poorly.
ClusterMI ^d	Pools all clustering results together to refine the final cluster membership, leading to more stable and robust clusters.	Computationally heavy for large datasets or many imputations, complex to implement, and sensitive to imputation quality.
F.MICA ^e	Introduces CritCF to guide optimal cluster number selection and incorporates a relabeling process to ensure consistent cluster assignments across runs.	Computationally intensive for large datasets or many imputations, complex to implement, accuracy drops in high-variance scenarios.

425 ^a CCA, Complete Case Analysis; ^b MICA, Multiple Imputation Clustering Analysis; ^c StackMI,
426 Stacked Multiple Imputation; ^d ClusterMI, Cluster Analysis with Missing Values by Imputation;
427 ^e F.MICA, a Framework for Multiple Imputation in Cluster Analysis
428

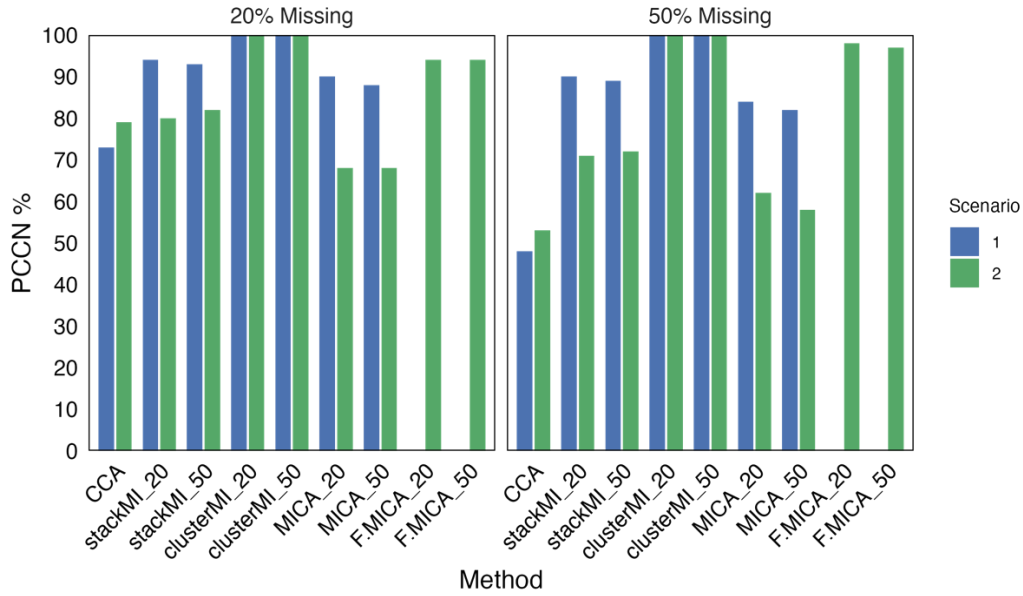


Figure 1. Percentage of correct cluster numbers (PCCN) obtained by different methods under two covariance scenarios (scenario 1: larger variances, scenario 2: smaller variances) with 20% and 50% missingness across 100 simulations. Methods with the suffix ‘_20’ and ‘_50’ indicate that multiple imputation were performed 20 and 50 times, respectively.

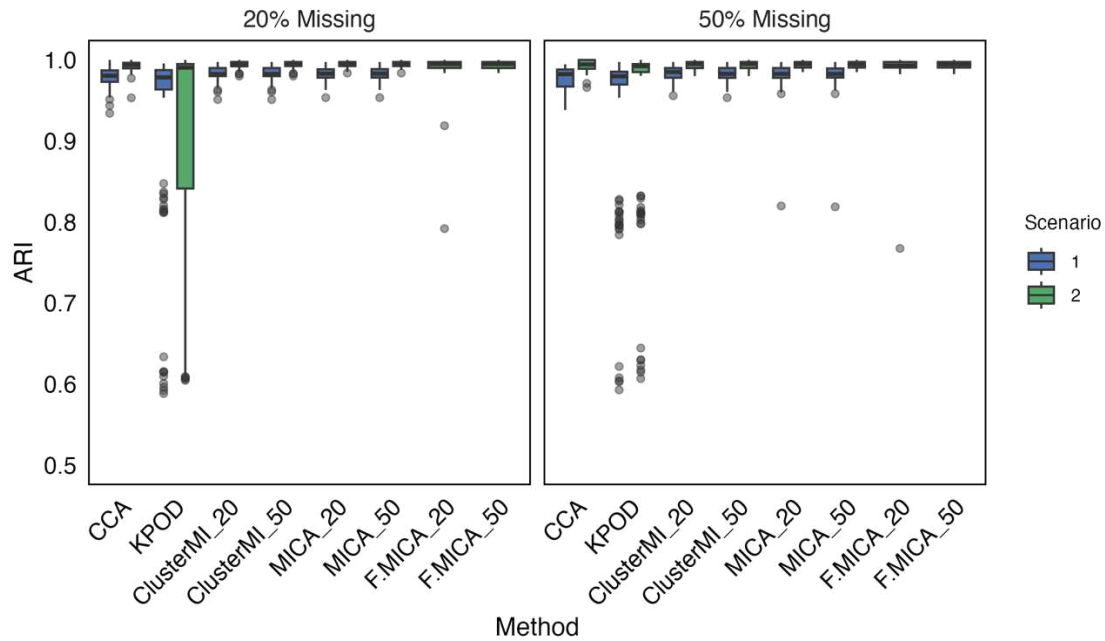


Figure 2. Boxplot of Adjusted Rand Index (ARI) using each method among the simulated datasets with correct number of clusters identified. Data are simulated under two covariance scenarios (scenario 1: larger variances, scenario 2: smaller variances) with 20% and 50% missingness. Methods with the suffix ‘_20’ and ‘_50’ indicate that multiple imputation were performed 20 and 50 times, respectively.

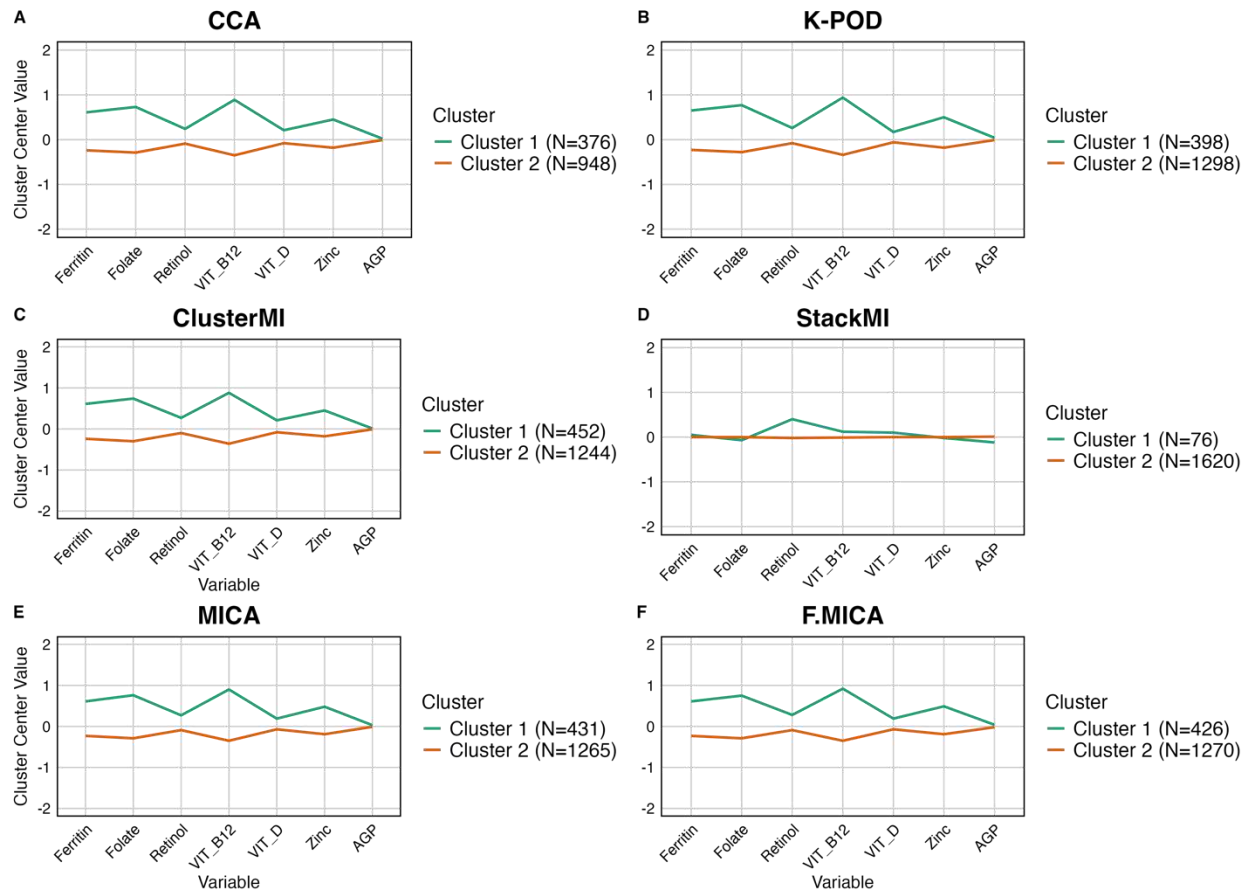


Figure 3. Cluster centers identified by six methods using the maternal micronutrition data measured at baseline enrollment in the JiVitA-3 study

Declarations

Ethics approval and consent to participate

This study involved secondary analysis using de-identified data from the JiVitA-3 randomized trial. The original JiVitA-3 trial received ethical approval from the Johns Hopkins Bloomberg School of Public Health institutional review board, Baltimore, Maryland, and the Bangladesh Medical Research Council, Dhaka, Bangladesh. Oral informed consent was provided by all study participants. The present analysis was determined to be exempt from additional IRB review.

Consent for publication

Not applicable

Availability of data and materials

The JiVitA-3 data used in the current study are not publicly available due to data use agreement restrictions. Access may be granted upon reasonable request to the study principal investigator.

Competing interests

The authors declare that they have no competing interests.

Funding

The JiVitA-3 trial and associated biomarker assessments were funded by the Bill & Melinda Gates Foundation (GH614). This work was supported by the Bill & Melinda Gates Foundation (INV-010744), Centers for Disease Control and Prevention, Eunice Kennedy Shriver National Institute of Child Health and Human Development, HarvestPlus, and the United States Agency for International Development. The funders were not involved in the study design, conduct, analysis, or reporting of results.

Authors' contributions

XH: analyzed and wrote the original manuscript. GJ, HL, MFY: reviewed and edited manuscript. YK: supervised the project, managed project administration, and reviewed and edited the manuscript.

Acknowledgements

We gratefully acknowledge Kerry Schulze, Parul Christian, and Keith West for providing access to the JiVitA-3 trial dataset and for their invaluable assistance in reviewing and guiding the interpretation of the biomarker data.

References

1. Gan G, Ma C, Wu J. Data clustering: theory, algorithms, and applications. SIAM; 2020.
2. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons; 2009.
3. McLachlan GJ, Peel D. Finite mixture models. John Wiley & Sons; 2000.
4. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv CSUR. 1999;31(3):264–323.
5. Basagaña X, Barrera-Gómez J, Benet M, Antó JM, Garcia-Aymerich J. A Framework for Multiple Imputation in Cluster Analysis. Am J Epidemiol. 2013 Apr 1;177(7):718–25.
6. Plaehn D. stackMI_Revisiting French tomato data: Cluster analysis with incomplete data. Food Qual Prefer. 2019 Sept;76:146–59.
7. Audigier V, Niang N. Clustering with missing data: which equivalent for Rubin’s rules? Adv Data Anal Classif. 2023 Sept 1;17(3):623–57.
8. Lee JW, Harel O. Incomplete clustering analysis via multiple imputation. J Appl Stat. 2023 July 4;50(9):1962–79.
9. Storlie CB, Myers SM, Katusic SK, Weaver AL, Voigt RG, Croarkin PE, et al. Clustering and variable selection in the presence of mixed variable types and missing data. Stat Med. 2018 Aug 30;37(19):2884–99.
10. Boluki S, Zamani Dadaneh S, Qian X, Dougherty ER. Optimal clustering with missing values. BMC Bioinformatics. 2019 June;20(S12):321.

- 503 11. Zhang L, Lu W, Liu X, Pedrycz W, Zhong C. Fuzzy C-Means clustering of incomplete data
504 based on probabilistic information granules of missing values. *Knowl-Based Syst.* 2016
505 May;99:51–70.
- 506 12. Afridi MK, Azam N, Yao J, Alanazi E. A three-way clustering approach for handling
507 missing data using GTRS. *Int J Approx Reason.* 2018 July;98:11–24.
- 508 13. Chi JT, Chi EC, Baraniuk RG. k-POD: A Method for k -Means Clustering of Missing
509 Data. *Am Stat.* 2016 Jan 2;70(1):91–9.
- 510 14. Himmelspach L, Conrad S. Clustering approaches for data with missing values: Comparison
511 and evaluation. In: 2010 Fifth International Conference on Digital Information Management
512 (ICDIM). IEEE; 2010. p. 19–28.
- 513 15. Schulze KJ, Mehra S, Shaikh S, Ali H, Shamim AA, Wu LSF, et al. Antenatal Multiple
514 Micronutrient Supplementation Compared to Iron-Folic Acid Affects Micronutrient Status
515 but Does Not Eliminate Deficiencies in a Randomized Controlled Trial Among Pregnant
516 Women of Rural Bangladesh. *J Nutr.* 2019 July 1;149(7):1260–70.
- 517 16. West KP, Shamim AA, Mehra S, Labrique AB, Ali H, Shaikh S, et al. Effect of maternal
518 multiple micronutrient vs iron-folic acid supplementation on infant mortality and adverse
519 birth outcomes in rural Bangladesh: the JiVitA-3 randomized trial. *JAMA.* 2014 Dec
520 24;312(24):2649–58.
- 521 17. Lee JW, Harel O. MICA_Incomplete clustering analysis via multiple imputation. *J Appl Stat.*
522 2023 July 4;50(9):1962–79.

- 523 18. Yeung KY, Ruzzo WL. An empirical study on principal component analysis for clustering
524 gene expression data. *Bioinformatics*. 2001;17(9):763–74.
- 525 19. Xu D, Tian Y. A Comprehensive Survey of Clustering Algorithms. *Ann Data Sci*. 2015
526 June;2(2):165–93.
- 527 20. Hong C, Qin X. A Label Propagation Algorithm Based on Neighbor Label Influence. In:
528 2024 9th International Conference on Intelligent Informatics and Biomedical Sciences
529 (ICIIBMS) [Internet]. Okinawa, Japan: IEEE; 2024 [cited 2025 Mar 13]. p. 105–11.
530 Available from: <https://ieeexplore.ieee.org/document/10792774/>
- 531 21. Thurnham DI. Interactions between nutrition and immune function: using inflammation
532 biomarkers to interpret micronutrient status. *Proc Nutr Soc*. 2014 Feb;73(1):1–8.
- 533 22. Suchdev PS, Namaste SM, Aaron GJ, Raiten DJ, Brown KH, Flores-Ayala R. Overview of
534 the Biomarkers Reflecting Inflammation and Nutritional Determinants of Anemia
535 (BRINDA) Project. *Adv Nutr*. 2016 Mar;7(2):349–56.
- 536 23. Thurnham DI, Northrop-Clewes CA. Inflammation and biomarkers of micronutrient status.
537 *Curr Opin Clin Nutr Metab Care*. 2016;19(6):458–63.
- 538 24. Luo H, Geng J, Zeiler M, Nieckula E, Sandalinas F, Williams A, et al. A Practical Guide to
539 Adjust Micronutrient Biomarkers for Inflammation Using the BRINDA Method. *J Nutr*.
540 2023 Apr;153(4):1265–72.

25. Buuren SV, Groothuis-Oudshoorn K. **mice**: Multivariate Imputation by Chained Equations
in *R*. *J Stat Softw* [Internet]. 2011 [cited 2025 Mar 13];45(3). Available from:
<http://www.jstatsoft.org/v45/i03/>

564 **Additional File**

565 File name: Additional file 1

566 File format: .docx

567 Title: Supplementary Table 1 to Supplementary Table 5

568 Description:

569 Supplementary Table 1: It contains the evaluation results for six clustering methods with
570 20% and 50% missing data under two covariance scenarios across 100 simulations. The
571 evaluation criteria include the percentage of correct cluster numbers (PCCN), Adjusted Rand
572 Index (ARI), Normalized Mutual Information (NMI), and Purity Agreement (PA).

573 Supplementary Table 2: It contains the characteristics of women in each cluster centers
574 derived by six methods when applied on the JiVitA-3 trail micronutrition data. The sample size
575 for each cluster was reported in the table. The characteristics including maternal age, maternal
576 educational level and socioeconomic status.

577 Supplementary Table 3: It contains the prevalence of micronutrient deficiencies and
578 inflammation within each cluster generated by six clustering methods. Including percentage of
579 women with iron deficiency, folate deficiency, vitamin A deficiency, vitamin B12 deficiency,
580 vitamin D deficiency, zinc deficiency, and elevated inflammation ($AGP > 1g/L$).

581 Supplementary Table 4 and 5: They contain the detailed result of the micronutrient
582 biomarker concentrations in each cluster derived by six clustering methods using the JiVitA-3
583 trail micronutrition data. The table Supplementary Table 4 contains the result from complete case
584 analysis (CCA), k-POD, multiple imputation cluster analysis (MICA). The Supplementary Table 5
585 contains the result from cluster multiple imputation (clusterMI), stacked multiple imputation (stackMI),
586 and a framework for multiple imputation in cluster analysis (F.MICA).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AdditionalFile1.docx](#)