Journal of Quality Measurement and Analysis *JQMA* 21(3) 2025, 1-19
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*
https://doi.org/10.17576/jqma.2103.2025.01

MISSING VALUES TREATMENT IN AGRONOMY DATASET USING PCA-BASED MULTIPLE IMPUTATION (BOOTSTRAP VERSUS BAYESIAN)

(*Rawatan Nilai Lenyap dalam Set Data Agronomi Menggunakan Imputasi Berganda Berasaskan PCA (Bootstap Lawan Bayesian)*)

RAHIMAH SALLEHUDDIN* & NORSHAHIDA SHAADAN

ABSTRACT

Missing values are prevalent in agronomy datasets and need consideration to ensure the applicability of statistical methods and avoid bias in treating them. Previous studies indicate that multiple imputation is more effective than single imputation, with Principal Component Analysis (PCA)-based methods effectively handling multicollinearity in multivariate data. However, such approaches are rarely applied to agronomy data, hence there is a need to assess their performance to add knowledge in the area. This study evaluates the performance of two PCA-based multiple imputation approaches on missing multivariate agronomy data: multiple imputation using regularised PCA through bootstrap procedure (BootMI-REM-PCA) and multiple imputation using regularised PCA through Bayesian procedure (BayesMI-REM-PCA). The data were obtained from the Department of Agriculture Sarawak. A simulation study was conducted using 500 simulated datasets at 5%, 10%, and 20% missingness. Results showed comparable performance between BootMI-REM-PCA and BayesMI-REM-PCA at 5% missingness, with equal coefficient of determination (R^2) values of 0.998, while BootMI-REM-PCA exhibited slightly lower root mean squared error (RMSE) of 1.527 and mean absolute error (MAE) of 0.160. However, BayesMI-REM-PCA outperformed at higher missing rates, achieving the lowest RMSE (2.238 at 10% and 3.051 at 20%) and MAE (0.315 at 10% and 0.601 at 20%), along with the highest R^2 values of 0.996 and 0.993, respectively. While imputation accuracy declines as missing data increases, BayesMI-REM-PCA preserves the characteristics of real data. The findings are expected to help agricultural scientists and researchers prepare high-quality data for accurate analysis.

Keywords: agronomy data, missing data imputation, multiple imputation, PCA-based imputation, multicollinearity

ABSTRAK

Nilai lenyap merupakan isu lazim dalam bidang agronomi yang perlu diambilkira bagi memastikan kepenggunaan kaedah statistik dan mengelakkan bias. Kajian terdahulu menunjukkan kaedah imputasi berganda lebih berkesan berbanding imputasi tunggal, dengan kaedah berasaskan analisis komponen utama (PCA) mampu menangani masalah multikolineariti dalam data multivariat. Namun, kaedah ini jarang diaplikasikan pada data agronomi. Prestasi kaedah ini perlu dinilai untuk menambah pengetahuan dalam bidang ini. Kajian ini menilai prestasi dua kaedah imputasi berganda berasaskan PCA terhadap data agronomi multivariat yang lenyap: imputasi berganda berasaskan PCA menggunakan prosedur bootstrap (BootMI-REM-PCA) dan imputasi berganda berasaskan PCA menggunakan prosedur Bayesian (BayesMI-REM-PCA). Data untuk kajian ini diperoleh daripada Jabatan Pertanian Sarawak. Kajian simulasi dijalankan menggunakan 500 set data yang dijana pada kadar kelenyapan 5%, 10%, dan 20%. Keputusan menunjukkan prestasi setanding antara BootMI-REM-PCA dan BayesMI-REM-PCA pada kadar kelenyapan 5%, dengan nilai pekali penentuan (R^2) yang sama (0.998) dan BootMI-REM-PCA menunjukkan ralat min punca kuasa (RMSE) 1.527 dan ralat min mutlak (MAE) 0.160 yang sedikit lebih rendah. Namun, prestasi BayesMI-

REM-PCA adalah lebih baik pada kadar kelenyapan lebih tinggi dengan RMSE terendah (2.238 pada 10% dan 3.051 pada 20%) dan MAE terendah (0.315 pada 10% dan 0.601 pada 20%), serta nilai R^2 tertinggi (0.996 dan 0.993). Walaupun ketepatan imputasi menurun apabila kadar kelenyapan data meningkat, BayesMI-REM-PCA berkesan mengekalkan ciri-ciri data sebenar. Penemuan ini dijangka dapat membantu penyelidik menyediakan data agronomi yang berkualiti untuk analisis yang tepat.

Kata kunci: data agronomi, imputasi data lenyap, imputasi berganda, imputasi berasaskan PCA, multikolineariti

1. Introduction

Agriculture is essential for sustainable food production. One key discipline is agronomy which focuses on crop and soil management (Maliwal & Mundra 2011). It often involves collecting multivariate data (Jinubala & Lawrance 2016; Roney *et al.* 2023; Stochero *et al*. 2024) leading to high correlations between variables (Fioroni *et al.* 2023; Wiangsamut & Koolpluksee 2020). Like many fields, agronomy faces data quality issues, particularly missing values (Arciniegas-Alarcón *et al.* 2020). It can result from equipment malfunctions, pest infestations, or disease outbreaks (Abbasi *et al.* 2019). Missing value limits statistical analysis as some requires complete data (Faisal & Tutz 2021) such as some multivariate analyses (Arciniegas-Alarcón *et al.* 2023), regression analysis (Mensching *et al.* 2020), and time series analysis (Fang *et al.* 2023). It can increase bias, reduce the precision of parameter estimates (Ayilara *et al.* 2019; Hughes *et al.* 2019), cause information loss (Chow *et al.* 2019; Kim *et al.* 2019), mislead conclusions and predictions (Chaudhry *et al.* 2019; Faisal & Tutz 2021), and impair the performance of Internet of Things device (Abbasi *et al.* 2019). These negative consequences worsen as the percentage of missing data increases (You *et al.* 2023).

Addressing missing values is necessary. A straightforward approach is discarding observation (Austin *et al.* 2021; Kim *et al.* 2019) but this reduces data size leading to information loss and bias (Chow *et al.* 2019; Sanju *et al.* 2023). Alternatively, the missing point can be filled with an estimated value, a process known as imputation. This produces clean data while reducing bias (Lohr 2010) and the information is preserved. Imputation can be single or multiple with methods ranging from statistical techniques like mean imputation or multiple imputation (MI) to machine learning methods such as random forest (RF) and k-nearest neighbours (k-NN) (Fu *et al.* 2021). The choice depends on the missing data pattern and mechanism whether Missing Not at Random (MNAR), Missing at Random (MAR), or Missing Completely at Random (MCAR) (Jose *et al.* 2021; Soley-Bori 2013). Under the MAR assumption, several imputation methods can provide unbiased estimates and accurate inferences (Li & Stuart 2019). Machine learning techniques like random forest (Fu *et al.* 2021; Kim *et al.* 2019; You *et al.* 2023), missForest (Sanju *et al.* 2023), and k-means (Dubey & Rasool 2020) are widely used though they are generally computationally costly (Alwateer *et al.* 2024) and require large datasets. They are said to possibly result in overfitting in high-dimensional data without proper tuning (Wani 2024) and typically produce single imputed values that do not explicitly account for uncertainty unless adapted for multiple imputation (Bertsimas *et al.* 2018). Statistical techniques are preferable for smaller datasets as in the case of experimental agriculture data. Mean imputation was popular but it falsely reduces data variability and fails to capture relationships between variables (Austin *et al.* 2021). Multiple imputation is recommended for capturing variability in imputed values and has shown strong performance across different multivariate datasets (Alwateer *et al.* 2024; Arciniegas-Alarcón *et al.* 2020; Fu

et al. 2021; Jose *et al.* 2021; Solfanelli *et al.* 2019). Stochero *et al*. (2024) applied distribution-free multiple imputation on experimental agronomic data and noted a decrease in imputation precision when data variability increased. To address correlations among variables, researchers applied regression-based imputation methods like multivariate imputation by chained equations (MICE) (Zhong *et al.* 2018). However, it can produce unstable imputations in the presence of multicollinearity (Sanju *et al.* 2023). Additionally, the strength of the relationship between variables may be overestimated when using regression-based imputation (Alwateer *et al.* 2024).

A promising alternative is imputation using Principal Component Analysis (PCA) which estimates missing values by capturing the underlying data structure (Josse *et al.* 2011). By reducing the dimensionality to uncorrelated components, it lowers computational demands, especially with many variables. Missing values are iteratively imputed using an expectation-maximisation (EM) algorithm through repeated PCA until convergence. However, in case of weak relationships, extensive missing data, or noise, iterative PCA (EM-PCA) may suffer from overfitting. Regularised EM-PCA counters this by adding a regularisation term to stabilise predictions and mitigate overfitting (Josse & Husson 2016). Additionally, multiple imputation is integrated (Josse *et al.* 2011) to handle uncertainty from missing data, making it well-suited for multivariate data as it handles relationships between variables, multicollinearity, and captures uncertainty from missing values (Gwelo 2019; Josse & Husson 2016). Multiple Imputation using PCA (MIPCA) in its regularised form has been widely adopted. Mensching *et al.* (2020) used bootstrap MIPCA to impute 19% missing data in a 23-variable dataset predicting dairy cows' ruminal pH. It performed well as verified by distribution comparisons and regression coefficients. This success was repeated on a smaller dataset with 7.6% missing data (Mensching *et al.* 2021). Compared to k-NN and mean-based imputation, MIPCA better captured variability in imputed values and provided more accurate joint distributions, especially with highly correlated variables (Chow *et al.* 2019). Bootstrap MIPCA also achieved 95% classification consistency in cardiometabolic health analysis (Wimalasena *et al.* 2023). It has also been applied to impute missing values in medical datasets (Gad *et al.* 2024). Similarly, Bayesian MIPCA has demonstrated reliability (Gomez-Bougie *et al.* 2018; Liu *et al.* 2018). It outperformed Amelia, MICE, and listwise deletion in simulations and provides stable imputations with narrower confidence intervals across various conditions (Audigier *et al.* 2016). MIPCA's ability to avoid overfitting makes it a competitive alternative to other multiple imputation methods (Josse & Husson 2012). Both MIPCA approaches were applied alongside chained equation procedures to high-dimensional datasets and demonstrating lower computational demand and faster imputation completion (Brini & van den Heuvel, 2024).

An effective imputation method for multivariate agronomy data is needed. PCA-based multiple imputation shows promise for capturing variables' relationships while mitigating multicollinearity, yet MIPCA remains underutilised in agronomy, with limited empirical evaluation in this field. Furthermore, in-depth studies comparing different MIPCA methods are scarce. This research examines the application of regularised MIPCA for multivariate agronomy data using two approaches, the bootstrap procedure and Bayesian treatment. They are empirically evaluated alongside the common mean method. The findings aim to provide insights into the effectiveness of MIPCA approaches and offer guidance for future applications in agronomy and beyond.

2. Methodology

This study investigates the performance of PCA-based multiple imputation methods using a regularised iterative PCA algorithm (MI-REM-PCA) for imputing missing values in multivariate agronomy data. Two approaches are examined, the bootstrap procedure (BootMI-

REM-PCA) and the Bayesian approach (BayesMI-REM-PCA). Secondary data is acquired from the Department of Agriculture, Sarawak, consisting of a multivariate dataset on a leafy green vegetable collected through a field experiment conducted in 2019. It includes the growth performance and yield of sweet Caixin cultivated using a bio-compost made from agricultural wastes. The dataset encompasses five variables measured across 172 Caixin plants, namely plant weight in grams (g), plant height in centimetres (cm), number of leaves, leaf width in centimetres (cm), and leaf length in centimetres (cm). All variables are measured on a ratio scale. The dataset contains 1.86% missing values. A subset of the data consisting of 161 complete observations (rows) is extracted as a reference dataset. The reference data is approximately multivariate normally distributed and consists of highly positively correlated variables, with high Pearson correlation coefficients of 0.92 (plant weight and plant height), 0.88 (plant weight and number of leaves), and 0.87 (plant height and number of leaves). The highest Variance Inflation Factor (VIF) and Condition Index (CI) observed are: plant weight (VIF=4.4, CI=41.2), plant height (VIF=4.9, CI=41.3), number of leaves (VIF=6.9, CI=42.5), leaf width (VIF=8.1, CI=42.3) and leaf length (VIF=8.3, CI=42.5). VIF values greater than 5 and CI values exceeding 30 indicate multicollinearity concerns (Kim 2019). This issue is addressed in the imputation procedure using PCA-based imputation methods where PCA transforms the correlated variables into uncorrelated components to mitigate multicollinearity.

The analysis involves imputation and validation processes. The imputation analysis is conducted through a simulation study to evaluate the performance of imputation methods where 500 datasets (Noghrehchi *et al.* 2021) are generated for three levels of missingness: 5%, 10%, and 20% (Loisel & Takane 2019; Sanju *et al.* 2023; Wimalasena *et al.* 2023). Two PCA-based multiple imputation approaches, bootstrap (BootMI-REM-PCA) and Bayesian (BayesMI-REM-PCA), are applied, alongside mean imputation as a baseline for comparison. The performance of BootMI-REM-PCA and BayesMI-REM-PCA is assessed relative to mean imputation across the varying missingness levels. The validation analysis examines whether the imputed data preserves the original characteristics of the real data. All analyses are performed using R software.

2.1. *Simulation study*

A simulation study is conducted to empirically assess the performance of imputation methods. For this purpose, datasets with missing values are generated by amputating the reference dataset using MAR mechanism, following the real data's missing patterns. For each missingness level (5%, 10%, 20%), 500 datasets are generated. The procedures of the simulation study are:

1. Extract a dataset with no missing values (reference dataset) from the original dataset.
2. Generate datasets with missing values following these steps:
    i. Examine the missing pattern of the original dataset that contains missing values.
    ii. Remove 5% of the data from the reference dataset following MAR mechanism and missing pattern of the original dataset. Generate 500 datasets with 5% missing data.
    iii. Repeat step ii for 10% and 20% missing levels.
3. Impute missing values using mean imputation, BootMI-REM-PCA, and BayesMI-REM-PCA.
4. Compute the performance metrics: root mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2).
5. Compute the average RMSE, MAE, and R^2 across 500 simulations.

6. Evaluate the performance of imputation methods.

2.2. *Mean imputation*

Mean imputation assigns a single value to substitute a missing data point. The mean of the observed values in each variable is computed and used to substitute any missing observation within that variable (Ben Aissia *et al.* 2017). In Figure 1, mean imputation is applied to a multivariate dataset containing two variables, x_1 and x_2.

Variable x_1	Variable x_2
x_{11}	x_{21}
x_{12}	NA
NA	x_{23}
x_{14}	x_{24}
x_{15}	NA

$$\bar{x}_1 = \frac{x_{11} + x_{12} + x_{14} + x_{15}}{4}$$

$$\bar{x}_2 = \frac{x_{21} + x_{23} + x_{24}}{3}$$

Variable x_1	Variable x_2
x_{11}	x_{21}
x_{12}	$\bar{x}_2$
$\bar{x}_1$	x_{23}
x_{14}	x_{24}
x_{15}	$\bar{x}_2$

Figure 1: Mean imputation

2.3. *Multiple imputation using regularised iterative PCA (MI-REM-PCA)*

PCA is a dimensionality reduction technique that identifies principal components (PC) capturing most data variability. Imputing missing values within a PCA framework uses these PCs to estimate missing values in all variables by considering relationships between variables and similarities among observations (Josse *et al.* 2011). When applying PCA to a dataset X, the reconstructed matrix, $\hat{X}$ is expressed as the first S terms of the singular value decomposition (SVD) of X (Josse & Husson 2016):

$$X = U_{nxS} \Lambda^{\frac{1}{2}}_{SxS} V^T_{pxS} \tag{1}$$

where U is the left singular vectors' matrix, V^T is the transpose of the right singular vectors' matrix (loadings matrix), Λ is the diagonal matrix of the eigenvalues, $U\Lambda$ is the principal components (scores) matrix, n is the number of observations, p is the number of variables, and S is the number of dimensions. Missing values are iteratively imputed until convergence using the EM algorithm to minimise the least square criterion $\left\| X - \hat{X} \right\|^2$ (Josse & Husson 2016). The elements of matrix X are expressed as (Josse & Husson 2016):

$$x_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} + \varepsilon_{ij} \ , \ \varepsilon_{ij} \sim N(0, \sigma^2) \tag{2}$$

The REM-PCA imputation is performed iteratively (Josse & Husson 2016) (Figure 2). In the initialisation step ($l = 0$), the mean matrix, M^0 of the incomplete matrix X is computed. The data matrix is centred by subtracting the mean of each variable. As the algorithm progresses ($l \geq 1$), the data is re-centred at each iteration in response to changes in the mean matrix M^l. PCA is performed on the re-centred imputed data and a PCA line is estimated. Missing points are substituted with the regularised PCA-fitted values while the observed values are retained. The

fitted matrix, X^l is computed by introducing the regularisation term, $(\hat{\sigma}^2)^l / \sqrt{\lambda_s^l}$, such that (Josse & Husson 2016):

$$\hat{x}_{ij}^l = \sum_{s=1}^{S} \left(\sqrt{\lambda_s^l} - \frac{(\hat{\sigma}^2)^l}{\sqrt{\lambda_s^l}} \right) u_{is}^l v_{js}^l \tag{3}$$

where σ^2 is the noise variance estimated as:

$$(\hat{\sigma}^2)^l = \frac{\left\| X^{l-1} - U^l (\Lambda^l)^{\frac{1}{2}} (V^l)^T \right\|^2}{np - nS - pS + S^2} \tag{4}$$

Step (3) to (6) are iteratively performed until convergence, where changes in the imputed matrix is lower than a predefined threshold, $\varepsilon = 10^{-6}$ (Josse & Husson 2016):

$$\sum_{ij} (\hat{x}_{ij}^{l-1} - \hat{x}_{ij}^l)^2 \leq \varepsilon \tag{5}$$

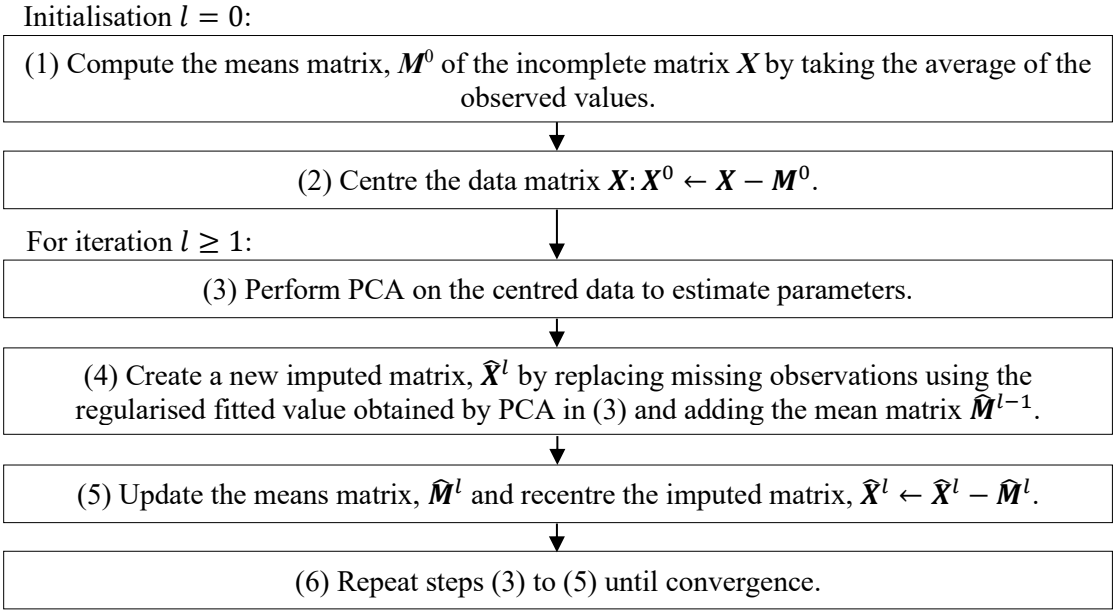

Figure 2: Regularised iterative PCA (REM-PCA) algorithm

The dimension, *S*, is predefined to minimise the mean squared error of prediction (MSEP). A generalised cross-validation (GCV) technique is utilised as it is less computationally costly than leave-one-out and k-fold techniques (Josse & Husson 2012). The *S* that minimises the GCV criterion is kept. The GCV value is expressed as (Josse & Husson 2016):

$$GCV(S) = \frac{np \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - (\hat{x}_{ij})^S \right)^2}{np - p - nS - pS + S^2 + S} \tag{6}$$

Uncertainties in estimations encompass variations in parameter estimates and variability due to noise. To address these variations in imputed values, multiple imputation is incorporated into the PCA framework, resulting in several imputed datasets. In this case, parameters are estimated by calculating estimates for each imputed dataset and then pooling the results to derive a final set of estimates (Soley-Bori 2013). The performance of multiple imputation using regularised iterative PCA (MI-REM-PCA) depends on the underlying missing mechanism and the nature of the variables. This method assumes that missing values occur under either MCAR or MAR mechanism and is limited to continuous data as PCA operates on numeric variables (Josse & Husson 2012). In this study, two approaches of MI-REM-PCA are employed and compared, the residuals bootstrap by Josse *et al.* (2011) and the Bayesian treatment introduced by Audigier *et al.* (2016).

Performing multiple imputations requires deciding the number of imputed datasets to generate, typically between 20 to 100 (Austin *et al.* 2021). Graham *et al.* (2007) highlighted that statistical power improves with an increase in the number of imputations. Through a simulation study testing 3 to 100 imputations, they observed that as the number of imputations decreased from 100 to 3, error measures increased while statistical power dropped. While increasing the number of imputations improve relative efficiency and statistical power, White *et al.* (2011) emphasised the importance of also considering Monte Carlo error, which is the standard deviation of the estimates obtained from repeated imputations. To control this error and ensure reproducibility, Royston and White (2011) suggested selecting a number of imputations greater than the missing percentage or between 100 to 1000, particularly for studies that compare statistical methods. In this study, 100 imputed datasets are generated to ensure high statistical power and minimise Monte Carlo error. This number is applied to both bootstrap and Bayesian approaches to ensure methodological consistency.

2.3.1. *MI-REM-PCA using bootstrap procedure (BootMI-REM-PCA)*

Following the methodology in Josse *et al.* (2011), a residual bootstrap method is employed to address the variations in parameter estimates during the PCA procedure (Figure 3). The procedure started with applying REM-PCA to create the imputed matrix X, followed by deriving the residuals matrix, $\hat{\varepsilon} = X - X$, which represents the noise or uncertainty in missing data. Assuming a Gaussian distribution with mean 0 and variance equal to the variance of the residuals, bootstrapping is performed to obtain 100 bootstrap matrices of residuals, ε^b where b = 1,2,3,…,B = 100. Each bootstrap replicate of the incomplete dataset X is formed by adding a residuals matrix, ε^b to the initial PCA estimator, $X^b = X + \varepsilon^b$. The REM-PCA algorithm is then applied to each bootstrap matrix X^b to estimate the PCA parameters, resulting in 100 estimators $X^1, X^2, \ldots, X^{100}$. This yielded a predictive distribution that represents the distribution of possible values for the missing data. Missing values are imputed using plausible values drawn from the predictive distribution with added Gaussian noise. This procedure resulted in 100 imputed datasets.

2.3.2. *MI-REM-PCA using Bayesian treatment (BayesMI-REM-PCA)*

This method estimates PCA parameters using a Bayesian approach which requires an understanding of the posterior distribution in relation to its prior distribution. This Bayesian method highlights the uncertainty associated with parameter estimates. This study adopted a Bayesian procedure suggested by Audigier *et al.* (2016) with a prior distribution of:

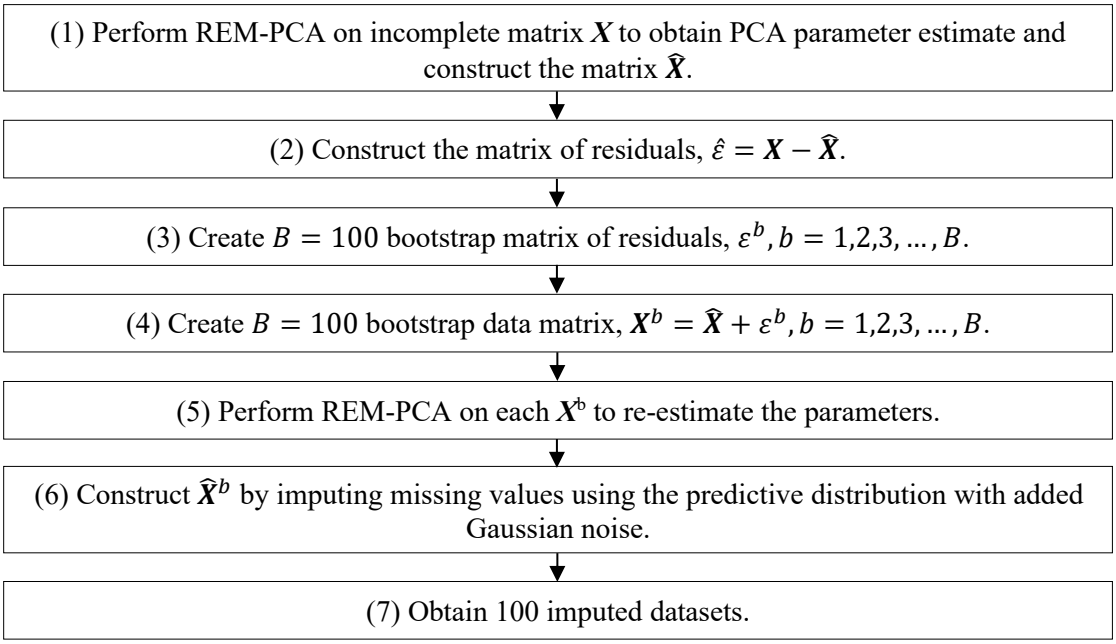

Figure 3: Multiple imputation using REM-PCA (bootstrap procedure)

$$\tilde{x}_{ij}^{(s)} \sim N(0, \tau_s^2),\ 1 \le s \le S \tag{7}$$

The posterior distribution is expressed as (Audigier *et al.* 2016):

$$\tilde{x}_{ij}^{(s)} \mid x_{ij}^{(s)} \sim N\left(\hat{x}_{ij}^{rPCA}, \frac{\hat{\sigma}^2 \sum_s \hat{\phi}_s}{\min(n-1, p)} \right) \tag{8}$$

where,

$$\hat{\sigma}^2 = \frac{\sum_{ij} (x_{ij} - \hat{x}_{ij})^2}{np - (p + S(n-1+p-S))}, \phi_s = \frac{\tau_s^2}{\tau_s^2 + \frac{\sigma^2}{\min(n-1, p)}} \text{ and } \hat{\tau}_s^2 = \frac{1}{np}\lambda_s - \frac{\hat{\sigma}^2}{\min(n-1, p)}$$

The algorithm is illustrated in Figure 4. Parameters are initially estimated using the REM-PCA. During the burn-in period ($l = 1$ to $Lstart = 1000$), the matrix X is imputed based on the estimated parameters. The posterior distribution is determined and used to impute missing values by drawing from it. Steps (2) to (5) are iteratively repeated for 1000 iterations aiming to achieve convergence. Imputed values are refined at each iteration based on the updated posterior distribution. Following the burn-in period, an imputed dataset is retained at regular intervals of L = 100 iterations until a total of 100 imputed datasets are obtained, specifically at *Lstart* + *L*, *Lstart* + 2*L*, … , *Lstart* + 100*L*. Keeping the imputed dataset at certain *L*s ensures that the retained datasets are more independent and representative of the underlying distribution. Finally, 100 imputed datasets are obtained.

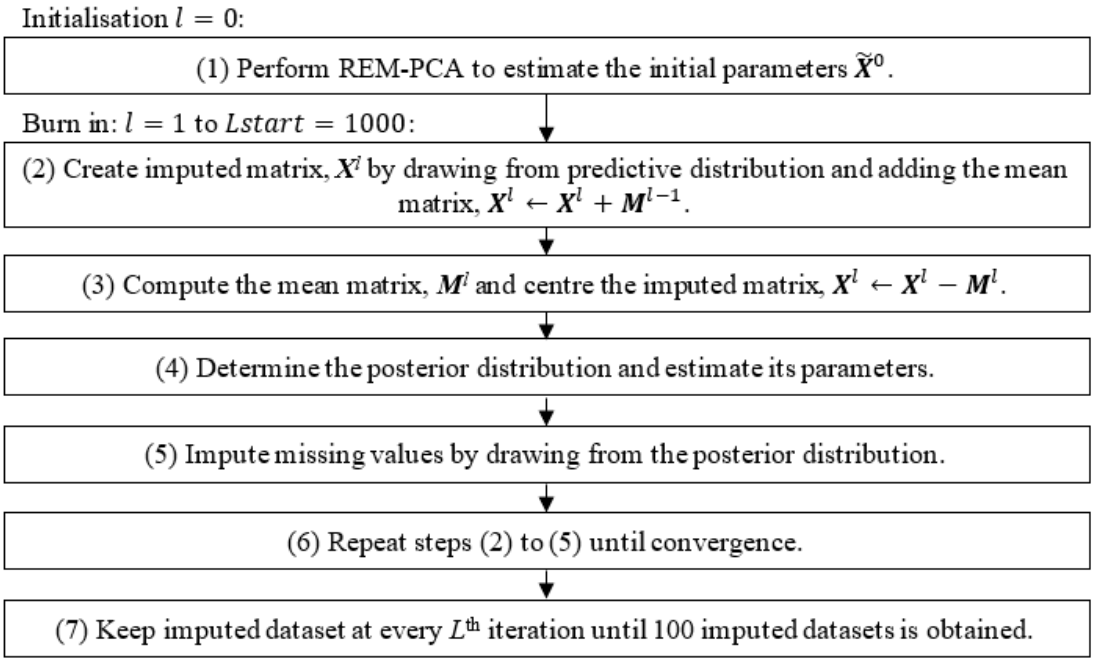

Figure 4: Multiple imputation using REM-PCA (Bayesian treatment)

2.4. *Performance indicators*

The assessment is based on three performance indicators, the root mean squared error (RMSE), mean absolute error (MAE) and coefficient of determination (R^2). The averages of these metrics are calculated for each imputation method for all levels of missingness.

2.4.1. Root mean squared error (RMSE)

Root mean squared error (RMSE) is a common metric used to quantify the difference between observed and predicted values. It measures the average deviation of imputed values from the observed values, calculated as:

$$RMSE = \sqrt{\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \hat{x}_{ij})^2} \tag{9}$$

where n is the number of observations, p is the number of variables, x_{ij} is the observed value in the i-th row and j-th column of the reference dataset, and x_{ij} is the imputed value in the i-th row and j-th column. Lower RMSE values signify better performance (Chicco *et al.* 2021).

2.4.2. Mean absolute error (MAE)

Mean absolute error (MAE) assesses the average absolute difference between the imputed and observed values. It is calculated as:

$$MAE = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left|x_{ij} - \hat{x}_{ij}\right| \tag{10}$$

where n is the number of observations, p is the number of variables, x_{ij} is the observed value in the i-th row and j-th column of the reference dataset, and x_{ij} is the imputed value in the i-th row and j-th column. A lower MAE value indicates better performance (Chicco *et al.* 2021).

2.4.3. Coefficient of determination (R^2)

The coefficient of determination, R^2 assesses how well the imputation method captures the variability in the data. It indicates the proportion of the total variance in the observed data that is explained by the imputed values. It is expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \hat{x}_{ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \bar{x})^2} \tag{11}$$

where n is the number of observations, p is the number of variables, x_{ij} is the observed value in the i-th row and j-th column of the reference dataset, x_{ij} is the imputed value in the i-th row and j-th column, and $\bar{x}$ is the mean of the observed values across all variables. The R^2 value ranges from 0 to 1, where a value closer to 1 indicates better performance (Saeipourdizaj *et al.* 2021).

2.5. *Validation analysis*

The validation analysis assesses the validity of the best imputation method obtained in the imputation analysis by comparing the summary statistics and PCA output of the reference and imputed data. The minimum, maximum, mean, median, standard deviation, range, skewness, and kurtosis are computed. These statistics are calculated for each of the 100 imputed datasets and averaged. Consistency in these summary statistics is evaluated between reference and imputed data. PCA results are compared in terms of the eigenvalues and the number of PCs contributing to more than 70% of the total variation (Jollife & Cadima 2016). Eigenvalues and cumulative percentages of total variation for 100 imputed datasets are averaged. Scree plots are used to visually compare the eigenvalues.

3. Results and Discussion

Missing values in the simulated datasets are imputed using mean imputation, BootMI-REM-PCA, and BayesMI-REM-PCA. For the MI-REM-PCA approaches, the number of dimensions, S, is predefined using the GCV technique ($S = 2$) and100 multiple imputations are performed. Table 1 illustrates examples of the imputed values obtained from the 250th simulation with 5% missingness. For both BootMI-REM-PCA and BayesMI-REM-PCA, the minimum, maximum, and average of the imputed values across 100 imputations are presented.

The error metrics, RMSE and MAE, along with the R^2 values are averaged across 500 simulations. Lower error measures indicate that the imputed values are closer to the true values while a higher R^2 indicates that the imputed values capture a larger proportion of the variation in the true values. The PCA-based multiple imputation methods, BootMI-REM-PCA and BayesMI-REM-PCA demonstrates superior performance compared to the common mean method, exhibiting lower errors and higher R^2 values, while both MI-REM-PCA approaches shows comparable performance (Table 2). At a low level of missingness (5%), BootMI-REM-PCA slightly outperformed BayesMI-REM-PCA with the lowest RMSE and MAE values of

1.5273 and 0.1595, respectively. BayesMI-REM-PCA obtained slightly higher errors with an RMSE of 1.5349 and an MAE of 0.1607. However, both approaches yielded an equal R^2 of 0.9980. As the proportion of missingness increased to 10% and 20%, the BayesMI-REM-PCA became superior. At 10% missingness, it achieved the lowest RMSE (2.2376) and MAE (0.3153) along with the highest R^2 (0.9961). With the missing percentage rising to 20%, the RMSE escalated to 3.0511, the MAE to 0.6007, and the R^2 dropped to 0.9932, yet still outperforming the BootMI-REM-PCA. These findings suggest that both BootMI-REM-PCA and BayesMI-REM-PCA offer advantages over the common mean approach when handling missing data. While BootMI-REM-PCA slightly outperformed BayesMI-REM-PCA at a lower level of missingness (5%), BayesMI-REM-PCA demonstrated superior performance as the proportion of missing data increased to 10% and 20%. This suggests that BayesMI-REM-PCA is able to produce an accurate estimation of missing values across different levels of missing data. This observation aligns with Audigier *et al.* (2016) which stated that BayesMI-REM-PCA is efficient in handling missing values across various sample sizes, missing rates, number of variables, and strengths of relationships between variables. On average, depending on the level of missingness, the BayesMI-REM-PCA imputed values deviate from the true values by approximately 1.535 to 3.051 units and the absolute difference between the imputed values and the true values ranges from approximately 0.161 to 0.601 units. Additionally, the R^2 values indicate that approximately 99.32% to 99.80% of the variance in the true values is explained by the imputed values across different levels of missingness. It is also observed that the performance of the imputation methods declined as the proportion of missing data increased as indicated by increasing RMSE and MAE values and decreasing R^2 values. This observation is also highlighted in (Sanju *et al.* 2023; Waljee *et al.* 2013).

Table 1: Examples of imputed values (250th simulation, 5% missing)

Variables	Observed (Reference) Values	Imputed Values						
		Mean Imputation	BootMI-REM-PCA			BayesMI-REM-PCA		
			Min	Max	Mean	Min	Max	Mean
Weight (g)	118.0	100.48	70.69	169.98	120.05	75.93	156.74	120.91
Height (cm)	25.5	29.99	19.03	33.50	25.08	18.50	35.53	25.85
Leaves (no.)	13.0	12.78	10.74	16.32	13.70	11.61	16.44	14.08
Width (cm)	13.0	11.79	11.36	14.90	13.06	10.93	14.51	12.90
Length (cm)	22.0	21.11	19.26	26.27	22.18	19.78	24.82	22.29

Table 2: Performance of the imputation methods

Missing Percentages	Imputation Methods	Performance Metrics		
		RMSE	MAE	R^2
5	Mean Imputation	2.5502	0.2381	0.9946
	BootMI-REM-PCA	1.5273	0.1595	0.9980
	BayesMI-REM-PCA	1.5349	0.1607	0.9980
10	Mean Imputation	3.6208	0.4607	0.9901
	BootMI-REM-PCA	2.2559	0.3162	0.9960
	BayesMI-REM-PCA	2.2376	0.3153	0.9961
20	Mean Imputation	4.9344	0.8903	0.9823
	BootMI-REM-PCA	3.1944	0.6213	0.9925
	BayesMI-REM-PCA	3.0511	0.6007	0.9932

The distributions of RMSE, MAE, and R^2 values across 500 simulations (Figure 5) reveal less variability for multiple imputations using PCA (BootMI-REM-PCA and BayesMI-REM-

PCA), compared to the more dispersed values from mean imputation. This indicates that PCA-based multiple imputation (MI-REM-PCA) produced a more stable and reliable imputation compared to the common mean method. BootMI-REM-PCA and BayesMI-REM-PCA yielded similar distributions of RMSE, MAE, and R^2, suggesting comparable performance. The BootMI-REM-PCA method provided the most accurate estimates of missing values at low levels of missingness (5%) while the BayesMI-REM-PCA performed better at higher missing rates (10% and 20%). Overall, the BayesMI-REM-PCA is superior as it demonstrates the ability to accurately estimate missing values across various levels of missingness. The results also indicate that the challenges in missing values imputation increase with higher levels of missingness.

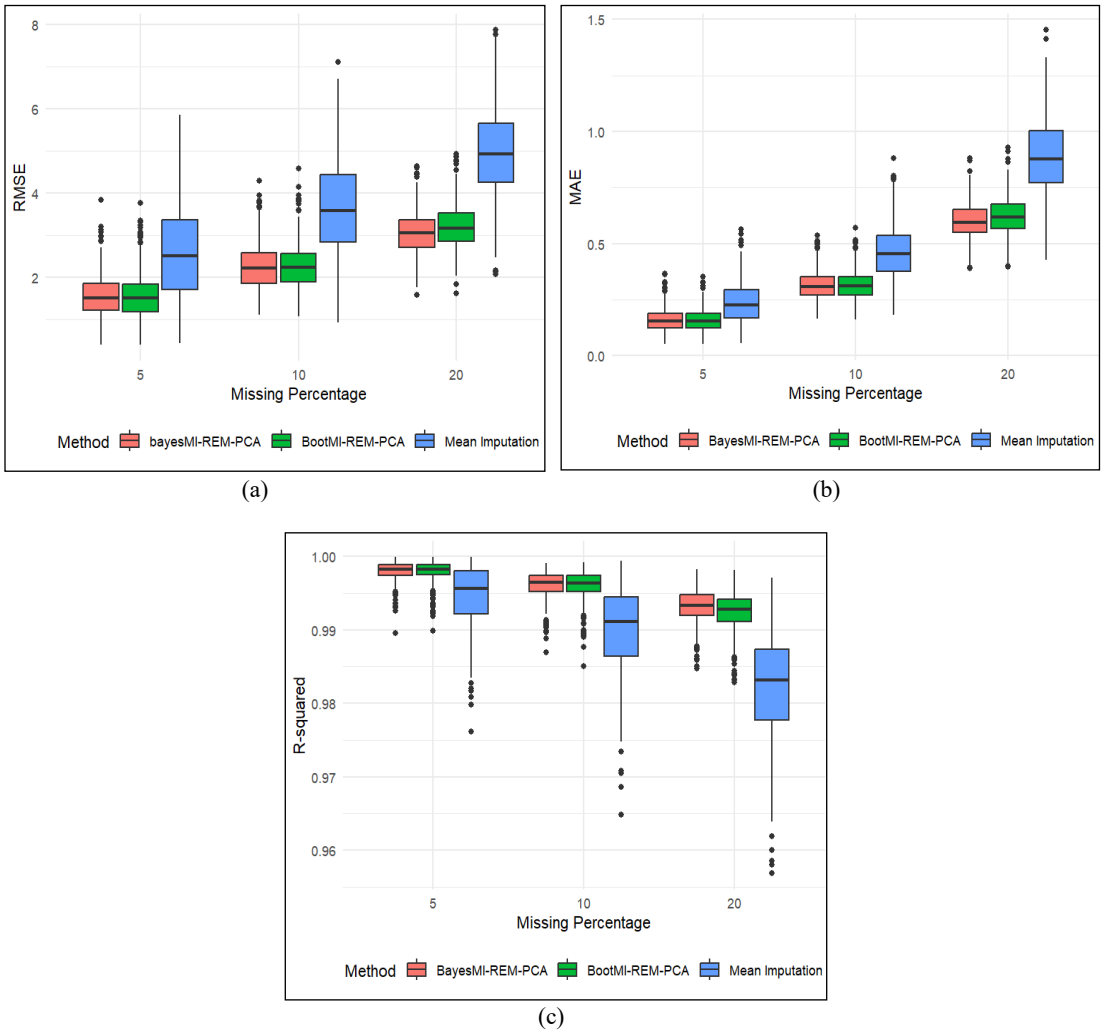

Figure 5: Distribution of the (a) RMSE, (b) MAE and (c) R^2 by imputation method and missing percentage

BayesMI-REM-PCA is identified as the best method considering various proportions of missing data. Validation is conducted through an examination of summary statistics and PCA performed on the datasets imputed using BayesMI-REM-PCA, compared to the reference data. Three imputed datasets are extracted from the first, 250th, and 500th simulations for each of the

5%, 10%, and 20% missingness. Subsequently, summary statistics are computed and PCA is conducted on these datasets along with the reference data. Table 3 presents that the minimum, maximum, mean, median, standard deviation, range, skewness, and kurtosis values exhibit close resemblance between the reference and imputed datasets. For example, the summary statistics of the 5% imputed dataset obtained from the 500th simulation show that the plant weight has a mean of 100.74, standard deviation of 39.77, median of 95, minimum of 35, maximum of 197.52, range of 162.52, skewness of 0.44, and kurtosis of -0.77. These values exhibit minimal differences with the observed values of 100.78 (mean), 39.80 (standard deviation), 95.00 (median), 35.00 (minimum), 197.00 (maximum), 162.00 (range), 0.44 (skewness), and -0.79 (kurtosis). Overall, minimal discrepancies are observed between the reference and imputed datasets in terms of these summary statistics. The imputed values maintain the original data's statistical properties thus indicating that the imputation using BayesMI-REM-PCA preserves the characteristics of the data.

Table 3: Descriptive statistics for reference and BayesMI-REM-PCA imputed data at various levels of missingness

Missing Rates	Var.	Simulation No.	Summary Statistics							
			Mean	SD	Median	Min	Max	Range	Skewness	Kurtosis
Reference (Complete) Dataset	Weight (g)	-	100.78	39.80	95.00	35.00	197.00	162.00	0.44	-0.79
	Height (cm)	-	30.00	5.39	30.00	21.00	45.00	24.00	0.18	-0.86
	Leaves (no.)	-	12.88	2.37	13.00	7.00	18.00	11.00	0.17	-0.47
	Width (cm)	-	11.81	1.23	12.00	8.50	15.00	6.50	-0.10	-0.05
	Length (cm)	-	21.17	1.96	21.00	17.50	27.50	10.00	0.63	0.13
5%	Weight (g)	1	100.82	39.73	95.30	34.84	197.62	162.79	0.43	-0.80
		250	100.82	39.89	94.86	35.00	197.00	162.00	0.44	-0.80
		500	100.74	39.77	95.00	35.00	197.52	162.52	0.44	-0.77
	Height (cm)	1	30.00	5.38	30.00	21.00	45.00	24.00	0.18	-0.84
		250	30.01	5.38	30.00	20.64	45.00	24.36	0.20	-0.83
		500	30.07	5.45	30.00	21.00	45.02	24.02	0.19	-0.86
	Leaves (no.)	1	12.88	2.37	13.00	7.00	18.01	11.01	0.17	-0.47
		250	12.86	2.35	13.00	7.00	18.06	11.06	0.16	-0.46
		500	12.84	2.35	13.00	7.00	18.08	11.08	0.16	-0.47
	Width (cm)	1	11.81	1.22	11.96	8.50	15.00	6.50	-0.09	-0.02
		250	11.85	1.25	12.00	8.50	15.04	6.54	-0.11	-0.09
		500	11.80	1.24	11.96	8.50	15.00	6.50	-0.06	-0.06
	Length (cm)	1	21.11	1.93	20.99	17.46	27.50	10.04	0.58	0.04
		250	21.21	2.01	21.00	17.44	27.51	10.07	0.59	0.01
		500	21.17	1.97	21.00	17.50	27.50	10.00	0.61	0.10
10%	Weight (g)	1	100.63	39.61	94.78	34.56	197.68	163.11	0.43	-0.79
		250	100.45	39.52	94.75	34.79	197.55	162.76	0.43	-0.79
		500	100.69	39.74	95.19	35.00	197.42	162.42	0.44	-0.76
	Height (cm)	1	30.07	5.44	30.08	20.61	45.02	24.41	0.17	-0.84
		250	30.02	5.43	30.00	20.66	45.12	24.45	0.21	-0.81
		500	30.07	5.44	30.03	20.96	45.03	24.07	0.18	-0.86
	Leaves (no.)	1	12.89	2.40	13.00	7.00	18.08	11.08	0.20	-0.50
		250	12.89	2.43	13.00	7.00	18.35	11.35	0.23	-0.49
		500	12.88	2.39	13.00	7.00	18.15	11.15	0.17	-0.54
	Width (cm)	1	11.82	1.22	11.95	8.50	14.89	6.39	-0.14	-0.08
		250	11.86	1.31	12.00	8.50	15.38	6.89	0.07	-0.01
		500	11.77	1.21	11.96	8.49	14.84	6.35	-0.13	-0.07
	Length (cm)	1	21.10	1.90	20.99	17.45	27.5	10.05	0.55	0.12
		250	21.19	1.95	20.99	17.41	27.00	9.59	0.49	-0.27
		500	21.21	2.02	20.99	17.48	27.50	10.02	0.60	0.03

Table 3 (Continued)

		1	100.65	39.89	95.47	25.51	201.41	175.9	0.41	-0.72
	Weight (g)	250	100.54	39.87	95.26	24.78	197.25	172.46	0.40	-0.78
		500	100.70	39.83	96.71	33.66	197.59	163.93	0.43	-0.76
		1	30.04	5.42	30.22	19.55	45.00	25.46	0.16	-0.76
	Height (cm)	250	29.87	5.43	29.99	20.27	45.06	24.78	0.25	-0.75
		500	30.06	5.45	30.05	20.04	45.00	24.96	0.17	-0.82
		1	12.87	2.38	13.00	7.89	18.08	10.19	0.20	-0.62
20%	Leaves (no.)	250	12.81	2.41	13.00	6.97	18.33	11.35	0.19	-0.48
		500	12.85	2.39	13.00	6.99	18.25	11.26	0.13	-0.48
		1	11.81	1.28	11.96	8.49	14.91	6.42	-0.14	-0.26
	Width (cm)	250	11.93	1.32	11.98	8.58	15.54	6.95	0.15	-0.10
		500	11.77	1.26	11.85	8.48	15.11	6.63	-0.05	-0.18
		1	21.24	1.94	21.04	16.68	27.51	10.82	0.42	0.16
	Length (cm)	250	21.14	1.96	20.96	17.04	27.01	9.96	0.48	-0.12
		500	21.19	2.02	21.00	17.08	27.50	10.42	0.53	0.10

Table 4 presents PCA results for reference and imputed datasets. Eigenvalues are comparable across datasets. In all datasets, PC1 captures at least 70% of the total variation. Scree plots (Figure 6) consistently show that PC1 has eigenvalues greater than one, indicating a similar elbow point across all datasets. This consistency suggests that the imputation preserves the underlying data structure.

Table 4: Eigenvalues (λ) and cumulative percentage of variation by PC for reference and BayesMI-REM-PCA imputed datasets

PC	Measure	Reference	5% Imputed			10% Imputed			20% Imputed		
			Simulation No.			Simulation No.			Simulation No.		
			1	250	500	1	250	500	1	250	500
1	λ	3.718	3.767	3.755	3.692	3.787	3.767	3.739	3.808	3.767	3.720
	Cum. %	74.36	75.34	75.10	73.84	75.74	75.34	74.78	76.16	75.34	74.40
2	λ	0.729	0.667	0.708	0.776	0.683	0.724	0.752	0.784	0.782	0.791
	Cum. %	88.94	88.68	89.26	89.37	89.40	89.82	89.82	91.84	90.98	90.21
3	λ	0.337	0.341	0.317	0.312	0.304	0.289	0.276	0.191	0.224	0.248
	Cum. %	95.68	95.49	95.61	95.62	95.48	95.61	95.33	95.66	95.45	95.18
4	λ	0.139	0.140	0.137	0.142	0.144	0.131	0.144	0.140	0.138	0.145
	Cum. %	98.46	98.29	98.34	98.46	98.36	98.24	98.22	98.46	98.20	98.09
5	λ	0.077	0.086	0.083	0.077	0.082	0.088	0.089	0.077	0.090	0.095
	Cum. %	100	100	100	100	100	100	100	100	100	100

Multiple imputation using PCA-based approaches proved effective for handling missing values in a highly correlated agronomy dataset. Both bootstrap and Bayesian approaches show more stable imputations with lower error and higher R^2 values than common mean imputation. These PCA-based approaches overcome the limitation of mean imputation which cannot capture and utilise the relationships between variables during imputation. Previous research has also established that PCA-based imputation approaches outperform methods such as MICE, k-NN, and Amelia in handling missing values in datasets with highly correlated variables (Chow *et al.* 2019; Audigier *et al.* 2016). Of the two PCA-based approaches studied, BayesMI-REM-PCA stands out for its superior performance. Its performance is comparable with BootMI-REM-PCA at 5% missingness and becomes superior when the missing rates increase to 10% and 20%. Overall, the results highlight BayesMI-REM-PCA's promising performance in providing accurate and precise estimations of missing values in multivariate datasets while preserving the data's original characteristics and structure.

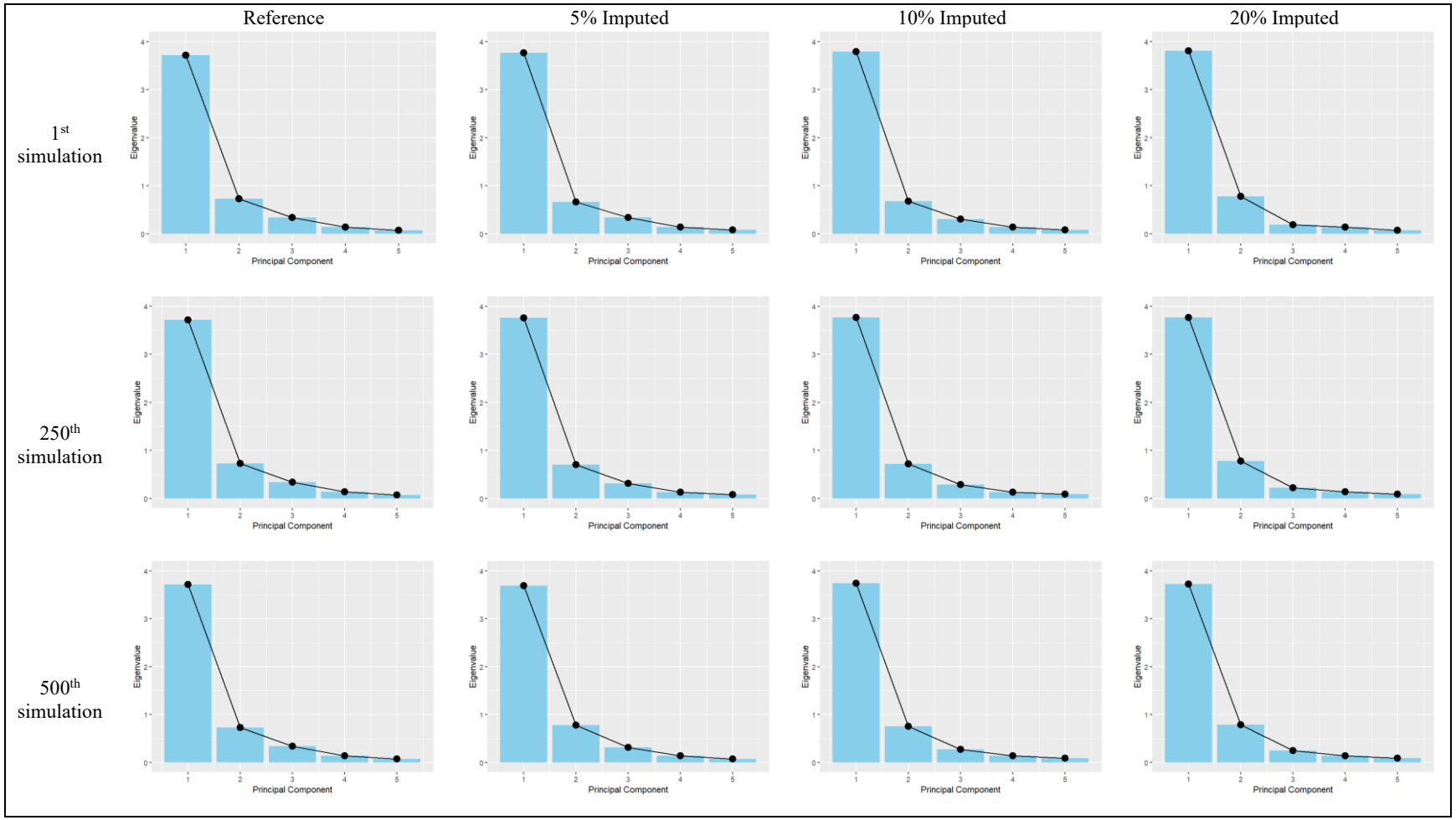

Figure 6: Comparison of scree plots between reference and BayesMI-REM-PCA imputed datasets

4. Conclusion

This study evaluates two multiple imputation approaches in regularised iterative PCA, bootstrap (BootMI-REM-PCA) and Bayesian (BayesMI-REM-PCA) along with mean imputation for imputing missing values in multivariate agronomy data. BootMI-REM-PCA resamples the residuals from the PCA model and adds variability based on the observed data distribution, whereas BayesMI-REM-PCA incorporates prior distributions and generates imputations based on posterior parameter distributions. The results indicate that the PCA-based methods outperformed mean imputation by capturing variable relationships and producing more stable imputations. These methods capture variability between imputations and handle multicollinearity. While BootMI-REM-PCA performed best at 5% missingness, its performance remains comparable to that of BayesMI-REM-PCA at this level. However, BayesMI-REM-PCA excelled at 10% and 20% missingness thus demonstrates a more stable performance across different missingness levels. As the proportion of missing values increases, so does the uncertainty in missing data. The superior performance of BayesMI-REM-PCA at 10% and 20% levels suggests the advantage of a probabilistic approach where incorporating prior distributions allows for more effective handling of uncertainty compared to the resampling-based BootMI-REM-PCA. This is further supported by the increasing gap in RMSE, MAE, and R^2 between both methods as missingness rises from 10% to 20%. Based on these findings, BayesMI-REM-PCA is recommended for imputing multivariate agronomy data as it provides accurate imputations particularly as missingness levels increase. In terms of computational time, both approaches do not require high computational demand, although BootMI-REM-PCA is faster. For one dataset in this study (number of observations = 161, number of variables = 5, number of multiple imputations = 100), BootMI-REM-PCA completed imputation in 0.78, 1.05, and 1.52 seconds for 5%, 10%, and 20% missingness, respectively. BayesMI-REM-PCA required 20.68, 20.98, and 21.39 seconds for the same levels of missingness. The computational time is expected to increase with larger datasets and higher levels of missingness. Overall, BayesMI-REM-PCA is recommended. At low missing rates (5%), BootMI-REM-PCA may still be considered due to its higher precision and shorter computational time. Beyond this study, multiple imputation using PCA can be applied to diverse agronomy datasets including longitudinal or temporal data, or datasets with long-gap missingness to assess broader applicability. Future research should explore factors such as the number of multiple imputations, sample sizes, the number of variables, and the strength of relationships between variables.

Acknowledgments

The authors would like to thank the Department of Agriculture Sarawak, particularly RO Siren Linggang of the Agriculture Research Centre Semongok for providing the data used in this study.

References

Abbasi M., Yaghmaee M.H. & Rahnama F. 2019. Internet of Things in agriculture: A survey. *Proceedings of the 3rd International Conference on Internet of Things and Applications*, pp. 1-12.

Alwateer M., Atlam E.S., Abd El-Raouf M.M., Ghoneim O.A. & Gad I. 2024. Missing data imputation: A comprehensive review. *Journal of Computer and Communications* **12**(11): 53-75.

Arciniegas-Alarcón S., García-Peña M. & Rodrigues P.C. 2020. New multiple imputation methods for genotype-by-environment data that combine singular value decomposition and Jackknife resampling or weighting schemes. *Computers and Electronics in Agriculture* **176**: 105617.

Arciniegas-Alarcón S., Garcia-Peña M., Krzanowski W.J. & Rengifo C. 2023. Missing value imputation in a data matrix using the regularisuied singular value decomposition. *MethodsX* **11**: 102289.

Audigier V., Husson F. & Josse J. 2016. Multiple imputation for continuous variables using a Bayesian principal component analysis. *Journal of Statistical Computation and Simulation* **86**(11): 2140-2156.

Austin P.C., White I.R., Lee D.S. & van Buuren S. 2021. Missing data in clinical research: A tutorial on multiple imputation. *Canadian Journal of Cardiology* **37**(9): 1322-1331.

Ayilara O.F., Zhang L., Sajobi T.T., Sawatzky R., Bohm E. & Lix L.M. 2019. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes* **17**: 106.

Ben Aissia M.A., Chebana F. & Ouarda T.B.M.J. 2017. Multivariate missing data in hydrology – Review and applications. *Advances in Water Resources* **110**: 299-309.

Bertsimas D., Pawlowski C. & Zhuo Y.D. 2018. From predictive methods to missing data imputation: An optimisation approach. *Journal of Machine Learning Research* **18**(196): 1-39.

Brini A. & van den Heuvel E.R. 2024. Missing data imputation with high-dimensional data. *The American Statistician* **78**(2): 240–252.

Chaudhry A., Li W., Basri A. & Patenaude F. 2019. A method for improving imputation and prediction accuracy of highly seasonal univariate data with large periods of missingness. *Wireless Communications and Mobile Computing* **2019**(1): 4039758.

Chicco D., Warrens M.J. & Jurman G. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **7**: e623.

Chow C., Andrášik R., Fischer B. & Keiler M. 2019. Application of statistical techniques to proportional loss data: Evaluating the predictive accuracy of physical vulnerability to hazardous hydro-meteorological events. *Journal of Environmental Management* **246**: 85-100.

Dubey A. & Rasool A. 2020. Clustering-based hybrid approach for multivariate missing data imputation. *International Journal of Advanced Computer Science and Applications* **11**(11): 710-714.

Faisal S. & Tutz G. 2021. Multiple imputation using nearest neighbour methods. *Information Sciences* **570**: 500-516.

Fang L., Xiang W., Zhou Y., Fang J., Chi L. & Ge Z. 2023. Dual-branch cross-dimensional self-attention-based imputation model for multivariate time series. *Knowledge-Based Systems* **279**: 110896.

Fioroni N., Mouquet-Rivier C., Meudec E., Cheynier V., Boudard F., Hemery Y. & Laurent-Babot C. 2023. Antioxidant capacity of polar and non-polar extracts of four African green leafy vegetables and correlation with polyphenol and carotenoid contents. *Antioxidants* **12**(9): 1726.

Fu Y., Liao H. & Lv L. 2021. A comparative study of various methods of handling missing data in UNSODA. *Agriculture* **11**(8): 727.

Gad A., Malouche D., Chhabra M., Hoang D., Suk D., Ron N., Dygulska B., Gudavalli M.B., Nadroo A.M., Narula P. & Elmakaty I. 2024. Impact of birth weight to placental weight ratio and other perinatal risk factors on left ventricular dimensions in newborns: A prospective cohort analysis. *Journal of Perinatal Medicine* **52**(4): 433-444.

Gomez-Bougie P., Maiga S., Tessoulin B., Bourcier J., Bonnet A., Rodriguez M.S., Le Gouill S., Touzeau C., Moreau P., Pellat-Deceunynck C. & Amiot M. 2018. BH3-mimetic toolkit guides the respective use of BCL2 and MCL1 BH3-mimetics in myeloma treatment. *Blood, The Journal of the American Society of Hematology* **132**(25): 2656-2669.

Graham J.W., Olchowski A.E. & Gilreath T.D. 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention science* **8**: 206-213.

Gwelo A.S. 2019. Principal components to overcome multicollinearity problem. *Oradea Journal of Business and Economics* **4**(1): 79-91.

Hughes R.A., Heron J., Sterne J.A.C. & Tilling K. 2019. Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology* **48**(4): 1294-1304.

Jinubala V. & Lawrance R. 2016. Analysis of missing data and imputation on agriculture data with predictive mean matching method. *International Journal of Science and Applied Information Technology* **5**(1): 01-04.

Jollife I.T. & Cadima J. 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**(2065): 20150202.

Jose J., Vishwakarma G.K. & Bhattacharjee A. 2021. Illustration of missing data handling technique generated from hepatitis C induced hepatocellular carcinoma cohort study. *Journal of King Saud University - Science* **33**(4): 101403.

Josse J. & Husson F. 2012. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique* **153**(2): 79-99.

Josse J. & Husson F. 2016. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software* **70**(1): 1-31.

Josse J., Pagès J. & Husson F. 2011. Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification* **5**: 231-246.

Kim J.H. 2019. Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology* **72**(6): 558–569.

Kim W., Cho W., Choi J., Kim J., Park C. & Choo J. 2019. A comparison of the effects of data imputation methods on model performance. *Proceedings of the 21st International Conference on Advanced Communication Technology (ICACT)*, pp. 592-599.

Li P. & Stuart E. A. 2019. Best (but oft-forgotten) practices: Missing data methods in randomised controlled nutrition trials. *The American Journal of Clinical Nutrition* **109**(3): 504-508.

Liu G., Yang Y. & Li B. 2018. Fuzzy rule-based oversampling technique for imbalanced and incomplete data learning. *Knowledge-Based Systems* **158**: 154-174.

Lohr S.L. 2010. *Sampling: Design and Analysis*. 2nd Ed. Belmont, CA: Brooks/Cole.

Loisel S. & Takane Y. 2019. Comparisons among several methods for handling missing data in principal component analysis (PCA). *Advances in Data Analysis and Classification* **13**: 495-518.

Maliwal P.L. & Mundra S.L. 2011. *Agronomy at a Glance: Vol-2 Objective Fundamentals*. 3rd Ed. Udaipur, Rajasthan, India: Agrotech Publishing Academy.

Mensching A., Hummel J. & Sharifi A.R. 2020. Statistical modeling of ruminal pH parameters from dairy cows based on a meta-analysis. *Journal of Dairy Science* **103**(1): 750-767.

Mensching A., Zschiesche M., Hummel J., Grelet C., Gengler N., Dänicke S. & Sharifi A.R. 2021. Development of a subacute ruminal acidosis risk score and its prediction using milk mid-infrared spectra in early-lactation cows. *Journal of Dairy Science* **104**(4): 4615-4634.

Noghrehchi F., Stoklosa J., Penev S. & Warton D.I. 2021. Selecting the model for multiple imputation of missing data: Just use an IC!. *Statistics in Medicine* **40**(10): 2467-2497.

Roney P.R., Lai L.S., Hamsein N.N. & Sallehudin R. 2023. Broccoli (Brassica oleraceae var. italica) growth performance in lowland using fertigation under protected rain shelter: The influence of spacings and accessions. *Malaysian Applied Biology* **52**(5): 205-211.

Royston P. & White I.R. 2011. Multiple imputation by chained equations (MICE): Implementation in Stata. *Journal of statistical software* **45**(4): 1-20.

Saeipourdizaj P., Sarbakhsh P. & Gholampour A. 2021. Application of imputation methods for missing values of PM10 and O3 data: Interpolation, moving average and K-nearest neighbour methods. *Environmental Health Engineering and Management Journal* **8**(3): 215-226.

Sanju, Kumar V. & Deepender. 2023. Evaluation of imputation techniques for genotypic data of soybean crop under missing completely at random mechanism. *Indian Journal of Agricultural Research* **57**(5): 701-705.

Soley-Bori M. 2013. Dealing with missing data: Key assumptions and methods for applied analysis. Technical Report. School of Public Health, Boston University.

Solfanelli F., Gambelli D., Vairo D. & Zanoli R. 2019. Estimating missing data for organic farming by multiple imputation: The case of organic fruit yields in Italy. *Organic Agriculture* **9**: 295-303.

Stochero E.L.M., Lúcio A.D.C. & Jacobi L.F. 2024. Data variability in the imputation quality of missing data. *Acta Scientiarum. Agronomy* **46**: e66185.

Waljee A.K., Mukherjee A., Singal A.G., Zhang Y., Warren J., Balis U., Marrero J., Zhu J. & Higgins P.D. 2013. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open* **3**(8): e002847.

Wani A.A. 2024. A review of challenges and solutions for using machine learning approaches for missing data. *International Journal of Engineering Applied Sciences and Technology* **9**(5): 36-50.

White I.R., Royston P. & Wood A.M. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* **30**(4): 377-399.

Wiangsamut B. & Koolpluksee M. 2020. Yield and growth of Pak Choi and Green Oak vegetables grown in substrate plots and hydroponic systems with different plant spacing. *International Journal of Agricultural Technology* **16**(4): 1063-1076.

Wimalasena S.T., Ramirez Silva C.I., Sun Y.V., Stein A.D., Rivera J.A. & Ramakrishnan U. 2023. Clustering of cardiometabolic risk factors in Mexican pre-adolescents. *Diabetes Research and Clinical Practice* **202**: 110818.

You J., Ellis J.L., Adams S., Sahar M., Jacobs M. & Tulpan D. 2023. Comparison of imputation methods for missing production data of dairy cattle. *animal* **17**: 100921.

Zhong H., Hu W. & Penn J.M. 2018. Application of multiple imputation in dealing with missing data in agricultural surveys: The case of BMP adoption. *Journal of Agricultural and Resource Economics* **43**(1): 78-102.

Agriculture Research Centre Semongok
Department of Agriculture Sarawak
P.O.Box 977, 93720 Kuching,
Sarawak, MALAYSIA
E-mail: rahimahs@sarawak.gov.my[*]

College of Computing, Informatic, and Mathematics
University Technology MARA
40450 Shah Alam,
Selangor, MALAYSIA
E-mail: shahida@tmsk.uitm.edu.my

Received: 16 December 2024
Accepted: 4 April 2025

[*]Corresponding author