Vital Annex: International Journal of Novel Research in Advanced Sciences
ISSN: 2751-756X
Volume 05 Number 02 (2026)
https://innosci.org/IJNRAS

Article

Evaluating the Effectiveness of Multiple Imputation for Handling Missing at Random Data : An Applied Study

Hayder Sami Alwan[1]

1. Misan University, Faculty of Business Management and Economics, 2026
* Correspondence: haydersami.a87@uomisan.edu.iq

Abstract: This research aims to assess the effectiveness of the Multiple Imputation (MI) method for processing randomly missing data. These missing data consider an essential challenge in statistical analysis because they influence on accuracy of outcomes and reliability of conclusions. Hence, suitable methods are required to handle such missing observations compares to maintain the robust statistical inference. This study also compares effectiveness of the Multiple Imputation (MI) with Deletion method, which is considered one of the most commonly used traditional techniques for treating missing data. Both methods are applied to handle missing values and to test their impact on statistical results. The simulation strategy was utilized using MATLAB 24 to generate data sets consisting of specific percentages of missing entries. The study was applies containing the following four main economic factors: exchange rate, price of a barrel of oil, foreign exchange reserves, and inflation rate which represent performance of the Iraqi economy during 2005-2024. The outcomes of the statistical analysis showed that both of Multiple Imputation Method and the Deletion Method provided statistically major results at a significance level of less than 0.05. This refers to that both methods persisted statistically significance even with presence of missing values. The findings also explained MI method obviously outperforms the Deletion method in retaining the true variance of the data and upholding the original sample size. Also it plays a major role in reducing bias and addressing uncertainty, related to missing data. In particular if the percentage for missing observations is high. In addition, the results indicates that MI is more stable and realistic than the deletion method. In spite of the Deletion approach explained more than 86% of the model while the multiple imputation method explained less than 58%, nearly 28% of the explanations evaluated by the deletion method are misleading, primarily because of the loss some of original data. Thus, the results recommend that MI represents a suitable method than the traditional deletion method for handling missing data.

Keywords: Multiple Imputation, Missing Data, Missing at Random, Deletion Method, Missing Completely at Random, Within- Imputation Variance, Imputation Phase, Consolidation Phase, Between - Imputation Variance

Citation: Alwan H. S. Evaluating the Effectiveness of Multiple Imputation for Handling Missing at Random Data : An Applied Study. Vital Annex: International Journal of Novel Research in Advanced Sciences 2026, 5(2), 46-60.

Received: 23th Jan 2026
Revised: 30th Jan 2026
Accepted: 28th Feb 2026
Published: 19th Mar 2026

1. Introduction

The problem of missing data is one of most common issues faced by researchers in statistical and applied studies. The presence of missing values may lead to an increase in variance and standard deviation, a reduction in sample size, bias parameter estimation, decreased statistical power and instability in statistical models. Consequently, this may result in misleading outcomes, which negatively affect the accuracy of statistical conclusions [1]. Rubin classified the mechanisms of missing data into several types, among which Missing at Random (MAR) represents intermediate case between Missing Completely at Random (MCAR) and Missing Not at Random (MNAR). Among the most advanced method used to address the problem of missing data is Multiple Imputation method, which is considered one of the most advanced and effective technique for handling this issue. This method has demonstrated high efficiency in dealing with missing values, particularly when the missingness mechanism follows the MAR assumption [2]. The Multiple Imputation approach works by generating several complete datasets that include estimated values for the missing observations. Each dataset is then analyzed statistically and the results are subsequently combined to obtain more accurate and reliable statistical estimates [3]. Another commonly used technique is the Deletion method, which involves excluding cases (rows) containing missing values prior to performing the statistical analysis. Consequently, the analysis is conducted only on complete observations within the dataset [4].

Research problem

Randomly missing data mainly influences on the quality of statistical estimation and validity of statistical inferences, particularly in economic models, that are sensitive to changes in sample size and time series data structure. The missing data produce challenges related to bias and instability of parameters, that negatively affects the accuracy of the results. For that reason, an important question appears about the best appropriate method for handling missing values in a way that preserves the estimation features and prevents results distortion. This work tests the extent to which MI can decrease bias and enhance the stability and accuracy of estimating economic variables when comparing with deletion method under conditions of randomly missing data using a simulation environment applied in MATLAB 24.

Research objective

This study aims to analyze the effectiveness of the Multiple Imputation method in handling randomly missing data and to statistically compare its performance with the Deletion method. This is achieved through a simulation approach using MATLAB version 24 applied to a set of economic variables, namely exchange rate, crude oil price per barrel, exchange rate, foreign reserves and inflation rate, which represent indicators the performance of the Iraqi economy during the period (2005-2024).

Furthermore, the study seeks to measure the impact of each method on bias and variance , in addition to evaluating the reliability and accuracy of the resulting statistical estimates.

Missing data

Missing data refer to cases (rows) in which complete values are not available for some or all variables. In other words, it represents the absence of one or more values for a specific variable, making it impossible for the researcher to access those values during the analysis stage. This situation may arise due to several reasons, such as measurement errors, date entry errors, participants withdrawal or non- response, damaged records, od difficulties encountered by researchers during the data collection process [5]. Such a missing values represent a lack of information that requires particular attention, especially during the statistical analysis stage, as they may lead to a reduction in sample size, weakened predictive ability of statistical models, and biased parameter estimates. Therefore, the methods used to handle missing data largely depend on the statistical mechanism that caused the missingness [6].

Types of missing data

There are several classifications of missing data, and these classifications have a direct impact on selecting the appropriate estimation method for handling missing values. In 1987, Little and Rubin classified missing data into several main types based on the mechanism that causes the data to be missing [7].

1- **Missing Completely at Random (MCAR)**

This type of missingness occurs when data are missing completely at random, meaning that the probability of missing data is entirely independent of both the observed and unobserved values in the dataset. In other words, there is no specific cause or systematic pattern leading to the missing values. In such cases, a researcher may lose a particular value without it being related to any other variable in the study, indicating that the missingness occurs purely at random.

$$P(R\,|Y_{obs}, Y_{mis}) = P(R)$$

In this case, deletion methods can be used without bias [8].

2- Missing at Random (MAR)

In this case, the missingness depends only on the observed values, and not on the missing values themselves, that is:

$$P(R|\,Y_{obs}, Y_{mis}) = P(R\,|Y_{obs})$$

This expression means that the probability of a particular value being missing can be explained by variables that are already observed and recorded in the dataset [9].

3- Missing Non at Random (MNAR)

It occurs when the probability of a value being missing depends on the missing value itself, meaning that the cause of the missingness is unobserved or directly related to the missing value. This type requires specialized models and is considered the most complex, that is [10]'

$$P(R\,|Y_{obs}, Y_{mis}) = P(R\,|Y_{mis})$$

In general, based on the above, it can be concluded that :

R: Missingness Indicator .

$Yobs$: Observed Values (Observations) .

$Ymis$: Missing Values .

2. Materials and Methods

1- Multiple Imputation (MI)

It is a statistical approach that replaces missing values with multiple plausible values instead of a single one by generating multiple complete datasets, analyzing each dataset separately, and then combining the results according to specific rules to obtain a single final estimate.[11] Each dataset contains different estimates for the missing values, and the advantages of using multiple imputation include: [12]

- Addressing uncertainty associated with missing data through multiple imputations and statistical pooling.
- Providing unbiased estimates.
- Preserving the true variance of the data and the correlations among variables.

Key Components of Multiple Imputation

A-Imputation phase:

It is considered the first stage in implementing multiple imputation, in which the following [13].

Generating multiple complete copies of the dataset by replacing missing values with plausible estimates.

Constructing the imputation model by estimating model parameters using only the observed data Y_obs.

Next, the process involves

- Generating the missing values m multiple times.
- Imputing them using a probabilistic model linking the missing values with explanatory variables.

The basic imputation equation is as follows:

$$\boldsymbol{\varepsilon}^{\mathbf{i}} \qquad \boldsymbol{i = 1, 2, \ldots, m}\widehat{\mathbf{Y}}^{\mathbf{i}}_{\mathbf{mis}} = \mathbf{X}\widehat{\mathbf{B}}^{\mathbf{i}}$$

Where:

$\widehat{\mathbf{Y}}^{\mathbf{i}}_{\mathbf{mis}}$: :The missing values imputed in the $\boldsymbol{i}$ -th dataset of the multiple imputation .

$\widehat{\boldsymbol{B}}^{\boldsymbol{i}}$: Estimation of model parameter in the $\boldsymbol{i}$ -th imputation .

$\boldsymbol{X}$: Matrix of observed (measured) independent variables .

$\boldsymbol{\varepsilon}^{\boldsymbol{i}}$: Random error of introducing uncertainty.

$\boldsymbol{m}$: Number of imputation sets .

B-Analysis phase:

At this stage, each complete dataset is analyzed separately to obtain the estimates, where the statistical model is applied to these data, resulting in different estimates of the parameters [14][15].

The general mathematical formula for the analysis stage is:[3]

$$\theta^i = f(Y^i)$$

θ^i: Estimate resulting from the $\boldsymbol{i}$ -th imputation

$f(\ )$: Statistical model used in the analysis

Y^i: The $\boldsymbol{i}$ -th imputed dataset

C- Integration phase :

This stage is considered the final stage, in which the estimates are combined using Rubin's rules to obtain a final estimate that accounts for uncertainty, while considering both the within-group and between-group variances. If we have $\boldsymbol{m}$ imputed datasets, and the estimator $\hat{\theta}_i$ is obtained from each dataset, then the overall estimate s given as follows :[2]

A- The mean of the estimates

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$$

B- Within – imputation variance

$$\bar{W} = \frac{1}{m}\sum_{i=1}^{m}W_i$$

C- Between - imputation variance

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\theta})^2$$

D- Total variance

$$T = \bar{W} + \left(1 + \frac{1}{m}\right)B$$

E- The Confidence interval of the estimated parameter

$$\bar{\theta} \pm t_{v,1-\alpha/2}\sqrt{T}$$

Where:

$\boldsymbol{m}$: Number of imputation

θ_i : Represents the estimate of the parameter from the i-th imputed dataset

$\bar{\theta}$: Mean of the estimates

W_i: Within-imputation variance

$\bar{W}$: Average within-imputation variance

T: Total variance

B : Between-imputation variance compensations

v: Degrees of freedom according to Rubin's formula

2-Method of deletion

This method is considered one of the most traditional and widely used approaches for handling missing data due to its simplicity and the fact that it does not require complex models. The principle of this method is to exclude rows (cases) that contain missing values in any variable within the dataset and perform the analysis using only the remaining

data, known as complete cases. This method provides unbiased estimates when Missing Completely at Random (MCAR) holds.[8] However, it is criticized for reducing the sample size, which in turn decreases statistical power and can introduce bias if the assumption of random independence of missingness is not satisfied, since the probability of a value being missing does not depend on the variable itself or other variables.[11]
The basic equation to identify complete cases after deletion is:[9]

$$n_c = \sum_{i=1}^{n} R_i$$

Where:
n: Total number of rows
n_c: Number of complete cases after deletion
R_i: Row response indicator, which takes the value (1) if all row values in the row are complete, and (0) if the row contains missing values.

Procedurally, the deletion method can be represented by a function that indicates the status of the row.
Thus if,

$$D = \{X_1.X_2, \dots., X_n\}$$

Then

$$D = \{X_i \in D | R_i = 0\}$$

This means that the final data consists only of completed rows and contains no missing values, and the sample size after applying this method is

$$n = n - k$$

Where:
n: It represents the original number of rows in the dataset.
k: Represents the number of rows containing missing values.

The condition for obtaining unbiased estimates
For the estimates obtained using the deletion method to be unbiased, the missing data must be Missing Completely at Random (MCAR). This condition is considered the requirement for unbiasedness, such that the presence of missingness does not affect any of the variable values in the study [4].

$$P(Ri = 1 | Xi, Yi) = P(Ri = 1)$$

This means that the probability of missing data is not related to the values of the variables themselves. Under this condition, complete case deletion becomes acceptable and does not lead to biased estimates.
Where:
$P(R_i = 1 | X_i, Y_i)$: The probability that the i-th row is complete given the values of that row
$P(R_i = 1)$: The probability that a row is complete in general, independent of the values of row i .
X, Y : Study variables [8].

3. Results and Discussion

Practical Aspect:

First: Data collection:

The study relied on real data from the Iraqi context during the period 2005–2024. The importance of this period lies in representing the changes in exchange rate policies that existed before this period, and it also follows the issuance of the new Central Bank Law in 2004. The data included a set of economic variables, all obtained from a single source: the Central Bank of Iraq (2024), which are

1-The Iraqi Dinar exchange rate against the US Dollar during the study period, represented in the model as exchange rate.

2-The crude oil price per barrel in US Dollars during the study period, represented in the model as oil price.

3-Foreign currency reserves in US Dollars (in billions) during the study period, represented in the model as foreign reserves.

4-The inflation rate during the study period, represented in the model as inflation

As shown in the following table.

Table 1. Data of variables used in the model for the period (2005 – 2024)

Years	exchange_rate	oil_price	foreign_reserves	inflation
2005	1250	55	15	8.5
2006	1270	58	18	8.2
2007	1290	62	22	7.9
2008	1310	65	25	7.5
2009	1330	70	30	7.2
2010	1350	75	35	6.8
2011	1380	80	42	6.5
2012	1420	85	50	6.2
2013	1450	90	58	6.0
2014	1470	95	65	5.8
2015	1460	85	70	6.0
2016	1440	80	75	6.2
2017	1420	75	80	6.5
2018	1380	70	85	6.8
2019	1320	65	90	7.0
2020	1310	60	85	7.2
2021	1300	55	80	7.5
2022	1290	65	82	7.3
2023	1280	75	85	7.0
2024	1270	78	88	6.8

Resource: Central Bank of Iraq. (2024). Monetary Policy report 2023. Baghdad

Second: Data preparation using MATLAB 24, under the title "Starting the Comprehensive search for Multiple Imputation ," four economic variables were created to represent the performance of the Iraqi economy during the study period, based on the data presented in the previous table. After that, the final statistical framework (for the sample of the first five years) was constructed, and the descriptive statistics of the data, according to the study variables, appeared as follows:

Table 2. Descriptive statistics for model variables

variable	arithmetic mean	standard deviation
Exchange_Rate	1349.50	71.85
Oil_Price	71.85	11.60
Foreign Reserves	59.00	26.74
Inflation	6.95	0.74

Source: Prepared by the researcher based on the results MATLAB 24

The table above illustrates the organizational structure of the data, combining all model variables into a unified framework that allows for systematic statistical analysis. Only five years and two measures - one for central tendency and the other for dispersion - were used for illustrative purposes to avoid overloading the report within the program. The statistical analysis of the model, however, will include all data for the entire study period.

Third: Simulating the missing data problem:

The results from MATLAB 24 showed that the targeted missingness rate would be 30%, with 24 missing values out of 80. Accordingly, the problem was successfully simulated. The diagnostic analysis of the data appeared as follows:

Table 3. Simulation of the missing data problem

variable	Number of missing values	Percentage of missing values from total data
Exchange_Rate	4	20%
Oil_Price	7	35%
Foreign Reserves	8	40%
Inflation	3	15%

Source: Prepared by the researcher based on the results MATLAB 24

The temporal analysis of missingness showed that the year with the highest number of missing values was 2013, where three missing values were recorded. Diagnostic analysis is considered a very important step for understanding the nature of missing data and subsequently determining the appropriate method for handling it.

Fourth: Application the Deletion Method:

As mentioned previously in the theoretical framework, the deletion method handles missing data by removing rows that have lost some of their data. Through simulations conducted on MATLAB 24, three rows with missing data were removed, thus reducing the number of rows from 20 to 14, as shown in Table (4) below. After deleting the rows, the data becomes as shown in Table (5) below :

Table 4. Missing data according to the simulation on the program MATLAB 24

years	Exchange_Rate	Oil_Price	Foreign Reserves	Inflation
2005	1250	55	15	8.5
2006	1270	58	18	8.2
2007	Non	62	22	7.9
2008	1310	65	25	7.5
2009	1330	70	30	7.2
2010	1350	75	35	6.8
2011	1380	Non	42	6.5
2012	1420	85	Non	6.2
2013	1450	90	58	6
2014	1470	95	65	5.8
2015	1460	85	70	6
2016	Non	80	75	6.2
2017	1420	75	80	6.5
2018	1380	70	85	6.8
2019	1320	65	90	Non
2020	1310	60	85	7.2
2021	1300	55	80	7.5
2022	1290	65	Non	7.3
2023	1280	75	85	7
2024	1270	78	88	6.8

Source: Prepared by the researcher based on the results MATLAB 24

Table 5. Data prepared for statistical analysis by the elimination method

years	Exchange_Rate	Oil_Price	Foreign Reserves	Inflation
2005	1250	55	15	8.5
2006	1270	58	18	8.2
2008	1310	65	25	7.5
2009	1330	70	30	7.2
2010	1350	75	35	6.8
2013	1450	90	58	6
2014	1470	95	65	5.8
2015	1460	85	70	6
2017	1420	75	80	6.5
2018	1380	70	85	6.8
2020	1310	60	85	7.2
2021	1300	55	80	7.5
2023	1280	75	85	7
2024	1270	78	88	6.8

Source: Prepared by the researcher based on Table No. (4).

After applying the deletion method in MATLAB 24, the results appeared as follows:

Table 6. Results of the elimination method

Linear regression model:

Inflation ~ 1 + Exchange_Rate + Oil_Price + Foreign_Reserves

Estimated Coefficients:

	Estimate	SE	T_Stat	P_Value
(Intercept)	14.968	0.91783	16.308	1.5618e-08
Exchange_Rate	-0.0037649	0.00088622	-4.2483	0.001694
Oil_Price	-0.034045	0.005446	-6.2513	9.4892e-05
Foreign Reserves	-0.0079777	0.0015569	-5.1241	0.00044804

Number of observations: 14, Error degrees of freedom: 10

Root Mean Squared Error: 0.149

R-squared: 0.973, Adjusted R-Squared: 0.964

F-statistic vs. Constant model: 118, p-value = 4.12e-08

Source: Prepared by the researcher based on the results MATLAB 24

The previous results indicate that when the Iraqi dinar exchange rate increases by one unit, inflation decreases by 0.00376%. Although this relationship is highly statistically significant at a significance level exceeding 99%, and it is also economically logical since the relationship between the exchange rate and inflation is inverse, the magnitude of the effect is very small and does not reflect economic reality. This indicates a decline in the efficiency of the deletion method in estimation.

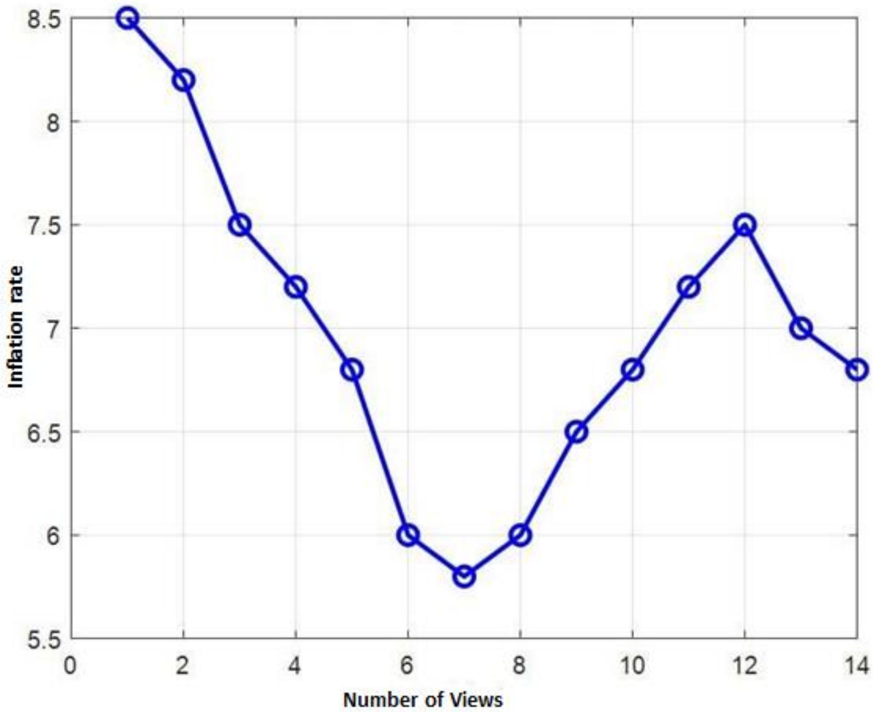
Regarding the price of a barrel of oil, a one-unit increase leads to an inverse change in inflation by 0.04%. This is economically reasonable because Iraq is considered a rentier economy that heavily depends on oil revenues.

Therefore, an increase in oil prices leads to an increase in the supply of foreign currency (US dollar), which contributes to reducing inflation. However, the statistical significance of this variable was not significant.

As for foreign reserves, an increase of one unit leads to a decrease in inflation by 0.008%, which is also economically logical for the same reasons mentioned above. Moreover, this relationship is statistically significant at a 99% confidence level.

The results also show that the model has high explanatory power, as the coefficient of determination $R - squared$ = 0.973, meaning that the model explains 97.3% of the variation in the inflation rate. Furthermore, after adjusting the coefficient of determination for the number of variables, the model still maintains high quality, explaining 96.4% of the variation in the inflation rate.

However, this quality may be illusory, since the deletion method may remove rows containing outlier values, which increases the variance among the data of each variable. This issue was also explained in the theoretical section. Consequently, the model may appear better than it actually is, while in reality it is not. Figure (1) illustrates the inflation rate after applying the deletion method for data treatment.



Source: Results MATLAB 24

Figure 1. Changes in inflation over time (after deletion)

Fifth:-Application the Multiple Imputation Method:

As mentioned in the theoretical section, this method is based on the principle of generating several complete versions of the real dataset, where the missing values are replaced with plausible values based on the probability distribution of the observed (available) real data. Therefore, the multiple imputation method is theoretically considered better than the deletion method, as it preserves the natural variance of the data and reduces the bias that may arise from missing values.

In this step, a simulation will be conducted using MATLAB 24 to perform multiple imputation for the missing values, and the results obtained are as follows:

1. The program performed the simulation using simple imputation method, as follows:
   Column 1: 1 value was imputed, mean = 1344.74
   Column 2: 2 values were imputed, mean = 71.39

Column 3: 2 values were imputed, mean = 58.22

Column 4: 1 value was imputed, mean = 6.9

After simple imputation: 20 complete row :

Column 1 represents the exchange rate, where one missing value was imputed using the mean of 1344.74

Column 2 represents oil prices, with two missing values replaced by the mean of 71.39

Column 3 represents foreign currency reserves, with two missing values imputed using

the mean of 58.22

Column 4 represents inflation, with one missing value replaced by the mean of 6.94

The simulation using the simple imputation method indicates that the pattern of missing values is not completely random, which supports the use of multiple imputation instead of simple deletion. Keeping all 20 rows helps to maintain the statistical power of the analysis.

2. The program performed the simulation using the simplified multiple imputation method, where five imputed datasets were created. This number is considered statistically optimal, as it is sufficient to achieve stable estimates.[2]

Each of the five datasets represents a possible scenario for the missing values, and each dataset exhibits a specific variance relative to the others. This between-dataset variance reflects the uncertainty inherent in the imputation process for the missing values.

… Imputed Dataset 1

… Imputed Dataset 2

... Imputed Dataset 3

... Imputed Dataset 4

... Imputed Dataset 5

Five imputed datasets were created

The multiple imputation method is based on the following equation:

$$\theta = \left(\frac{1}{m}\right) \boldsymbol{\Sigma}\, \theta_{\boldsymbol{i}}$$

$$Var\ (\theta) = W + \left(1 + \frac{1}{m}\right) B$$

Where:

θ: Pooled estimates

W: Within-Imputation Variance

B: Between-Imputation Variance

m: Number of Imputed Datasets

This method is characterized by preserving the variance through the addition of random noise that maintains the standard normal distribution. It also reduces estimation bias and accounts for the uncertainty that is inherent in the deletion method. The results showed that all imputed datasets were of high quality in explaining the model, and although Dataset 4 had the highest explanatory quality, the difference with the other datasets was not significant.

To obtain a single coefficient of determination representing the model's explanatory power, the mean across all datasets was calculated, which was 93% with a standard deviation of 0.0031 . This confirms that the model has high explanatory quality according to the multiple imputation method, and the low standard deviation indicates the stability of the estimated parameters.

Dataset 1 : $R^2 = 0.9302$

Dataset 2 : $R^2 = 0.9256$

Dataset 3 : $R^2 = 0.9299$

Dataset 4 : $R^2 = 0.9344$

Dataset 5 : $R^2 = 0.9298$

The statistical methodology of Rubin's results shows that the imputation method achieved an analysis of missing data with an efficiency of over 95%, and leads to the following conclusions:

A / The imputation method preserves the original sample size.

B / The estimation of parameters will be more accurate and stable.

C / The standard error of the estimates will decrease.

D / The method satisfies the assumptions of basic statistical models.

Sixth: Final Comparison between the Deletion and Multiple Imputation Methods

The results of the comparison between the two methods using MATLAB 24 showed the following:

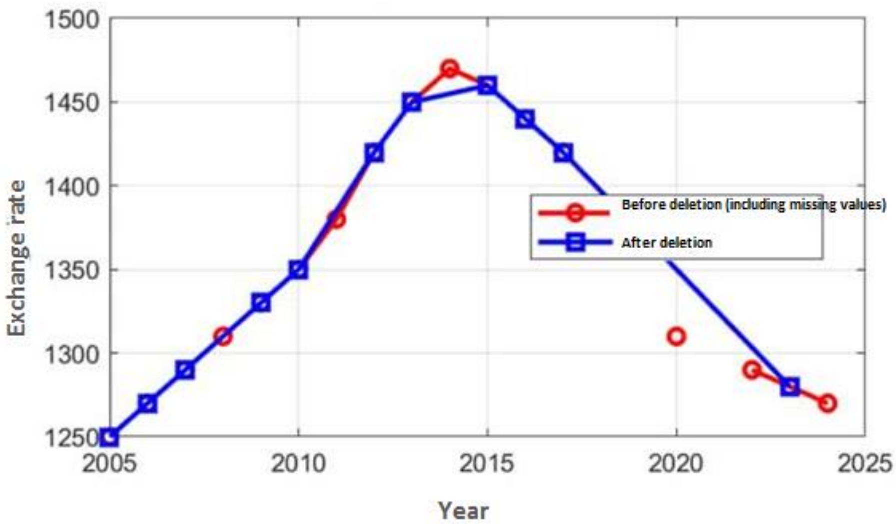
Standard	Delete method	Imputation method
Sample size in number of years	14	20
Coefficient of determination	0.973	0.93
Statistics F	29.25	27.16
Value (p) for the model	0.0015	0.001
The constant	905.032	1017
inclination	6.6181	4.6055

Seventh: Graphs

The following graphs illustrate the difference between the deletion and multiple imputation methods through visual observation.

1. Exchange Rates Before and After Deletion

Figure(2) illustrates how the deletion method reduced the exchange rate data and produced an unrealistic depiction of the time series of exchange rates over the period.

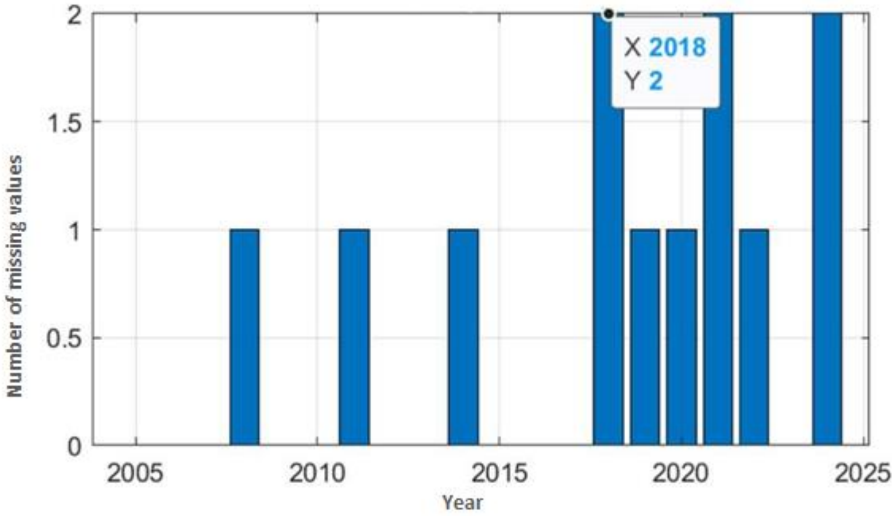


Source: Program MATLAB 24

Figure 2. Comparison of exchange rate before and after deletion

2. Distribution of Missing Data over the Study Period

Figure (3) shows that the missing data were concentrated in certain years, with 2018 and 2021 having the highest data loss, followed by 2024. The remaining missing data in other years were roughly equal, while some years had no missing data at all.

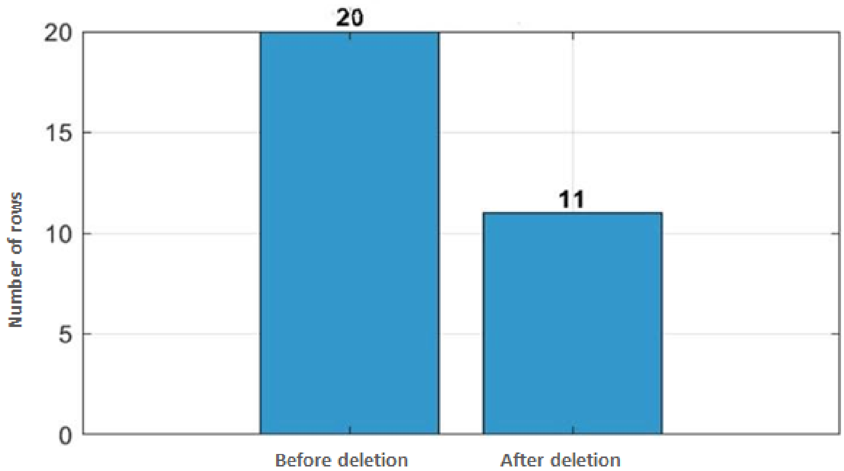


Source: Program MATLAB 24

Figure 3. Distribution of missing data over the years

3. Data Size Before and After Deletion:

Figure (4) shows the sample size before and after deletion.

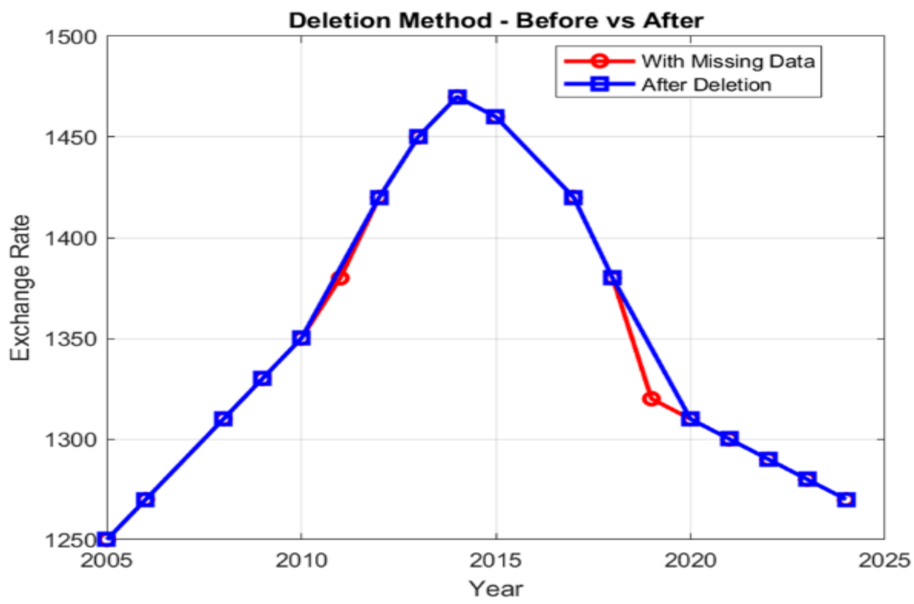


Source: Program MATLAB 24

Figure 4. Data size before and after deletion

4. Exchange Rates Using the Deletion Method and the Multiple Imputation Method:

Figure (5) illustrates how the multiple imputation method was able to address the shortcomings of the deletion method, as seen by comparing it with Figure (2).



Source: Program MATLAB 24

Figure 5. Deletion Method–Before and after

Eighth: Interpretation of results :

he model estimated using both the deletion method and the multiple imputation method showed statistical significance at a significance level of less than 0.05 . This indicates that both models were statistically significant despite the presence of some missing data.

Moreover, the deletion method appeared to provide better results than the multiple imputation method, as it explained more than 86% of the model, whereas the multiple imputation method explained less than 58% of the model. However, 28% of the explanatory power provided by the deletion method may be misleading and this can be explained by the following points:

1. When rows containing missing data are deleted, extreme values may also be removed, which increase the variability among the remaining data. As explained in the theoretical section, this may cause the model to appear better than it actually is, even though this improvement is not genuine.
2. The multiple imputation method preserves difficult or extreme observations, making the model more realistic.
3. The multiple imputation method introduces a higher level of uncertainty into the model by imputing missing values rather than deleting them, whereas the deletion method may produce a model that appears more precise than it actually is.
4. The variance decreases under the deletion method, whereas it increases under the multiple imputation method.

From an economic interpretation perspective, a change of one unit (one dollar) in oil prices leads to a positive change of 4.92 in exchange rates when using the deletion method. In contrast, under the multiple imputation method, a change of one unit (one dollar) in oil prices leads to a positive change of 4.6 in exchange rates.

This indicates that the multiple imputation method provides more accurate estimates, and its accuracy is expected to improve further as the model is expanded by including additional explanatory variables.

4. Conclusion

In light of the results obtained from the previous analysis, it has become evident that the multiple imputation method represents a methodological necessity to ensure the credibility of economic results in the presence of incomplete data, a problem faced by many developing economies, including the Iraqi economy. Therefore, this method should not be viewed merely as an advanced statistical technique; rather, it serves as an important tool that enhances the quality and reliability of applied research.

This fact is reflected in the following conclusions reached by the researcher.

1. The multiple imputation method demonstrated its ability to preserve the original sample size (20 observations) whereas the deletion method reduced the sample size by 30% .
2. The multiple imputation method preserved the natural variance of the data, whereas the deletion method distorted the data distribution by removing outliers, which are essential for understanding the true economic context.
3. The multiple imputation method provided a comprehensive solution to the uncertainty associated with missing data by generating multiple datasets and combining their results according to Rubin's rules.
4. The study showed that the coefficient of determination (R^2) for the multiple imputation method was 93%, while it was higher for the deletion method at 97.3%. However, the deletion method distorted the original statistical structure of the model, so R^2 alone does not fully reflect model quality. Other considerations must also be taken into account alongside the coefficient of determination.
5. The standard deviation in the multiple imputation method (0.0031) confirms the stability and reliability of the statistical results.
6. The multiple imputation method demonstrated higher efficiency in handling unstable data, as reflected by the data examined over the period 2005–2024 .

Recommendations

In light of the previous conclusions and based on the application of simulations on economic data, the researcher presents the following recommendations:

1. Adopt the multiple imputation method as a primary approach when dealing with missing data in economic studies, particularly in the Iraqi information environment, where the absence or poor quality of data for some economic indicators is often observed.
2. The deletion method should be avoided except in very rare cases where the missing data are completely random.
3. In studies with missing data, the researcher should clearly describe the method used to handle missing data in the research methodology section, rather than limiting it to the practical analysis only, to ensure statistical transparency.
4. Iraqi institutions should develop a project for handling missing data and make it accessible to researchers, so that researchers are not required to process missing data themselves. This approach would ensure accuracy, efficiency, and results that better reflect reality.

REFERENCES

[1] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.
[2] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley, 1987.
[3] C. K. Enders, *Applied Missing Data Analysis*. New York, NY: Guilford Press, 2010.
[4] P. D. Allison, *Missing Data*. Thousand Oaks, CA: Sage Publications, 2001.
[5] F. Abdel Hamid, "Modern methods for processing missing data," *Journal of Educational Research*, 2018.
[6] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annual Review of Psychology*, vol. 60, pp. 549–576, 2009.
[7] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 1st ed. New York, NY: Wiley, 1987.
[8] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Hoboken, NJ: Wiley, 2019.

[9] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. New York, NY: Chapman & Hall, 1997.
[10] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, pp. 147–177, 2002.
[11] A. Al-Ramahi, "Using multiple imputation in missing data," University of Baghdad, 2020.
[12] J. A. C. Sterne et al., "Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls," *BMJ*, vol. 338, 2009.
[13] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. Boca Raton, FL: CRC Press, 2018.
[14] G. King, J. Honaker, A. Joseph, and K. Scheve, "Analyzing incomplete political science data: An alternative algorithm for multiple imputation," *American Political Science Review*, vol. 95, no. 1, pp. 49–69, 2001.
[15] R. J. A. Little, "Modeling the drop-out mechanism in repeated-measures studies," *Journal of the American Statistical Association*, vol. 90, no. 431, pp. 1112–1121, 1995.