DOI: 10.22098/jhs.2024.15095.1015

Journal of Hyperstructures 13 (2) (2024), 297-304

Journal Homepage: https://jhs.uma.ac.ir/

Research Paper

A NEW HYBRID METHOD FOR DATA ANALYSIS WHEN A SIGNIFICANT PERCENTAGE OF DATA IS MISSING

Ahmad Nouraldin[1] and Behrouz Fathi Vajargah[2,*]

[1] Department of Applied Mathematics, University of Guilan Rasht, Iran, ahmadnouraldin5@gmail.com

[2] Department of Statistics, University of Guilan, Rasht, Iran, behrouz.fathi@gmail.com

ARTICLE INFO

Article history:
Received: 18 May 2024
Accepted: 16 September 2024
Communicated by Nasrin Eghbali

Keywords:
Missing Data
Imputation
Mean Square Error
Standard Deviation

MSC:
16D25
62D10

ABSTRACT

This article aims to compare the efficiency of different imputation methods with missing data. We use mean, median, Expected-Maximization (EM), regression imputation(RI) and multiple imputations (MI) to replace missing data. In fact, we employ three proposed combination methods, namely EM imputation with MI imputation (EMMI), EM imputation with regression imputation (EMR), and regression imputation with MI imputation (MI). We will compare these methods using an example study of Waterborne Container Trade by the US Customs Port (2000-2017) where the methods with different missing percentages. Several criteria, are used to compare estimations efficiency, such as mean, Standard Deviation (SD), and Mean Squared Error (MSE). The results show that the efficiency of composite imputation methods in almost all situations, in terms of MSE, RMI imputation method outperforms other methods. Nevertheless, when the missing percentage is small, the EMR imputation method performs better. In terms of the SD criterion, we find that the MI method is better than the other methods, where the RMI method is good when the missing percentage is large. When the missing percentage is in the range (40-50%), the EMR and RMI imputation methods give a better MSE.

*Address correspondence to B. Fathi Vajargah; Department of Statistics, University of Guilan, Rasht, Iran, behrouz.fathi@gmail.com.

1. INTRODUCTION

In practice, missing data is a common problem. When an observation has no value assigned to it, it is considered to be missing data. Any set of data may contain missing data if for any item, an input has not been entered or generated. Missing data can occur for a variety of reasons. There are a lot of reasons why variables might be counted as null or missing, such as data encoding, internet outages, a missing page of printed information, etc. The missing data mechanism introduced by Little and Rubin can be classified as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [7]. In MCAR, missing data is not dependent on observed or unobserved values. Because of the smaller sample size, the missingness increases standard errors, but it doesn't cause bias. It usually depends on observed values, not unobserved values. In this case, missingness is referred to as MAR. When the mechanism is based on missing data itself, it is classified as MNAR [5]. Missing data can be handled in several ways [9]. Popular methods such as the deletion method, overall mean imputation, and missing-indicator method produce biased estimates. Due to the reduction in sample size, statistical power is weakened as well as parameter estimates are biased, especially when the missingness mechanism is not MCAR. Recent years have made the imputation of missing data more popular [11]. A novel technique has been proposed to handle missing data by employing a partitioning approach on the dataset with missing values. The Expectation-Maximization (EM) method is then used to fill in the missing values in each partition [12].

A recent review of pharmacy literature shows that the proportion of studies reporting how missing data was handled is very low [8]. When findings are not reported, interpretations and validity of research may be biased. In this paper, we introduce the concept of missing data and discuss how missing data is classified, in addition to comparing different imputation methods when dealing with missing data [10].

Our article discusses imputation methods such as mean imputation, median imputation, regression imputation, EM imputation, and multiple imputations. Our next step is to introduce three composite imputation methods: EM with regression (EMR), EM with multiple imputation (EMMI), and regression imputation with multiple imputation (RMI). These three methods are compared with five methods of imputation: mean imputation, median imputation, regression imputation, EM imputation, and multiple imputations.

2. MISSING DATA MECHANISM

The missing data mechanism expresses the relationship between missing data and response values in the data matrix. Little and Rubin provide the following classification for the missing data mechanism:

2.1. Missing completely at random (MCAR).

MCAR is when a missing value is not related to any other value in the data set. Conceptually, data that are MCAR are not usually attributed to a question in the survey or other phenomenon, whether observable or unobservable.

2.2. Missing at random (MAR).

Data that are MAR are missing based on another observable instance, such as an underlying or confounding factor causing respondents to not answer questions. Certain groups may not respond to a question, as a result of an underlying reason.

2.3. **Missing not at random (MNAR).** Finally, MNAR, or data that contains non-ignorable missingness, are data that do not meet the criteria of either MCAR or MAR. Unlike MCAR and the use of an objective statistical test, subjective analysis is required to ascertain whether data are MNAR. In MAR, there may be a correlation between an observable phenomenon and why data are missing, but not a direct cause. Data that are MNAR, on the other hand, can be attributed to an unobservable factor that is directly affecting the reason that the data values are missing. This can be the question itself being the cause of the missing response, or underlying assumptions .

3. METHODS FOR HANDLING MISSING DATA

3.1. **Mean imputation.** In the mean imputation method, missing values of a variable are replaced by the mean of other observed values in the variable. Therefore, this method is limited to numerical data. Although by using this method, the sample size is maintained and the use is uncomplicated, the variance will be downwardly biased irrespective the underlying missing data mechanism [6]. The mean imputed value is given by

$$\overline{y}^* = \sum_{i=1}^{n-m} \left(\frac{y_i}{n-m}\right).$$

3.2. **Median imputation.** Median imputation consists of replacing all occurrences of missing values (NA) within a variable by the median.

3.3. **Classical regression method.** This method ignores all records with missing variables (inputs or outputs). This reduces the amount of information that needs to be analyzed. This approach is called complete case analysis. This method has two drawbacks [3].

3.4. **Expectationmaximization algorithm.** The EM algorithm was developed by [2]. The EM algorithm is a general iterative algorithm for calculating maximum likelihood estimates of parameters in the case that there are missing values in the data set. The EM algorithm consists of two steps: the mathematical expectation step (step E) and the maximization step (step M). In this algorithm, the missing value is replaced with another variable. It checks if this value is the most probable value, and if not, another value is substituted. This method continues until the most probable value is reached. This algorithm uses full data to calculate the mean, variance, and covariance.

3.5. **Multiple imputation.** Multiple imputation was proposed by [1]. In this method D imputed values for each of the missing observation is generated and hence we get D complete data set. From each of the complete data set an estimate of the parameter of interest q is obtained by using a standard technique, assuming no nonresponse is present. This process results in valid statistical inferences that properly reflect the uncertainty due to missing values.

4. Composite methods

In general, a composite method with equivalent weight is a combination of two or more methods which can be defined as follows.

$$\tilde{y}_i = W_M(\widehat{y}_{i1} + \widehat{y}_{i2} + \cdots + \widehat{y}_{iM}); \qquad i = 1, 2, \cdots, m,$$

where $\widehat{y}_{ij}$ is the imputation ith from method jth, and also the equivalent weight $W_M = \dfrac{1}{M}$ where M is the number of combination methods. Here, we develop three composite methods which is a combination of two frequently used methods in the literature as follows.

4.1. **EM imputation with MI imputation (EMMI).** EMMI is a combination of EM and MI imputations. The EMMI imputed value is given by

$$\tilde{y}_i = \frac{1}{2}(\widehat{y}_i + y_i^{'}); \qquad i = 1, 2, \cdots, m,$$

where $\widehat{y}_i$ and $y_i^{'}$ are the EM and MI imputation methods respectively.

4.2. **EM imputation with regression imputation (EMR).** EMR is a combination of EM and R imputations. The EMR imputed value is given by

$$\tilde{y}_i = \frac{1}{2}(\widehat{y}_i + y_i^{*}); \qquad i = 1, 2, \cdots, m,$$

where $\widehat{y}_i$ and y_i^{*} are the EM and R imputation methods respectively.

4.3. **Regression imputation with MI imputation (RMI).** RMI is a combination of R and MI imputations. The RMI imputed value is given by

$$\tilde{y}_i = \frac{1}{2}(y_i^{*} + y_i^{'}); \qquad i = 1, 2, \cdots, m,$$

where y_i^{*} and $y_i^{'}$ are the R and MI imputation methods respectively.

5. application to real data

In this section, we apply all the studied imputation methods to a real dataset (Exports and Imports dataset). The dataset was extracted from Waterborne Container Trade by the US Customs Port (2000-2017). This study examined the efficiency of the used imputation methods and compared them with the new composite methods, in which the exports and imports dataset was used to determine whether the data change the results for the methods used. The results are shown in Tables 1, 2, 3, 4, 5, 6 and Figures 1, 2, 3.

According to tables 1-6 and figures 1-3, composite methods perform better than mean and median imputation methods when the missing percentage is high. In addition to the combined methods, the EM and multiple imputation methods perform well in most situations when the missing ratio is less than 20%. Figures 1-3 illustrate two cases, case A represents the export data set, and case B represents the import data set.

For large missing percentage, the composite method has given better MSE than that of the previously mentioned methods.

However, when the missing percentage is large there is no exact method that performs well in all cases. So we suggest that we divide the data (2, 4, 10 parts) with the case of deletion and by using imputation methods inter of all parts we will get better results.

Table 1. Mean Squared Error (MSE) for exports with missing data (20%, 30%, 40%, 50%)

Method	M.S.E (20%)	M.S.E (30%)	M.S.E (40%)	M.S.E (50%)	Mean
Full data	96085.9	96085.9	96085.9	96085.9	96085.9
Missing data	104747.2	88189.9	140792.8	142604.2	119083.5
EM	95855.7	110297.5	86510.9	66615.7	89819.9
Regression	95424.5	75472.9	102679.4	98171	92936.9
MI	97406.4	118751.3	90526.4	84053.5	97684.4
Mean	83215.6	63137.3	79308.5	66615.7	73069.3
Median	84991.7	64854.5	83204.9	70116.9	75792
EMR	95640.1	92885.2	94595.2	82393.4	91378.5
EMMI	96631	11452.4	88518.7	75334.6	67984.2
RMI	96415.4	93612.1	96602	91112.3	94435.5
Mean for 8 methods	93197.6	78807.9	90243.3	79301.6	85387.9

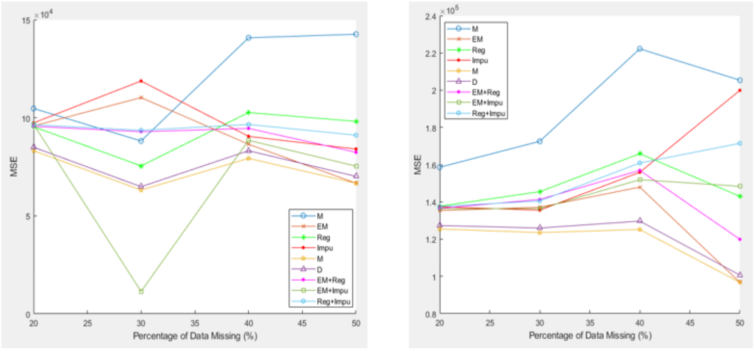
Table 2. Mean Squared Error (MSE) for imports with missing data (20%, 30%, 40%, 50%)

Method	M.S.E (20%)	M.S.E (30%)	M.S.E (40%)	M.S.E (50%)	Mean
Full data	139577.2	139577.2	139577.2	139577.2	139577.2
Missing data	158596.5	172535.9	222208.6	205333.5	189668.6
EM	135335.2	137146.5	147952.5	96824.6	129314.7
Regression	137665.3	145494.8	165878.9	142982.4	148005.4
MI	137375.7	135551.9	155904.2	199903.1	157183.7
Mean	125436	123522.6	125169.9	96824.6	117738.3
Median	127318.3	125974.4	129690.6	100681.9	120916.3
EMR	136500.3	141320.6	156915.7	119903	138659.9
EMMI	136355.5	136349.2	151928.4	148363.9	143249.3
RMI	137520.5	140523.4	160891.6	171442.8	152594.6
Mean for 8 methods	134188.4	135735.4	149291.5	134615.8	138457.8

6. conclusion

In this study, we examined performance and efficiency of imputation methods and compare them with new composite methods. This article compared five single imputation methods and three composite imputation methods for missing data for two variables (Exports and Imports). Our results indicated the efficiency of composite imputation methods in almost all situations in terms of MSES, where the RMI imputation method outperforms other methods. Nevertheless, when the missing percentage is small, the EMR imputation method performs better. In terms of the SD criterion, we find that the MI method is better than the other method, where the RMI method is good when the missing percentage is large.

On the other hand, when the missing percentage is high (40-50%), the estimator obtained by the EMR and RMI imputation methods caused smaller MSE. Moreover, we can see that



(*a*) (*b*)

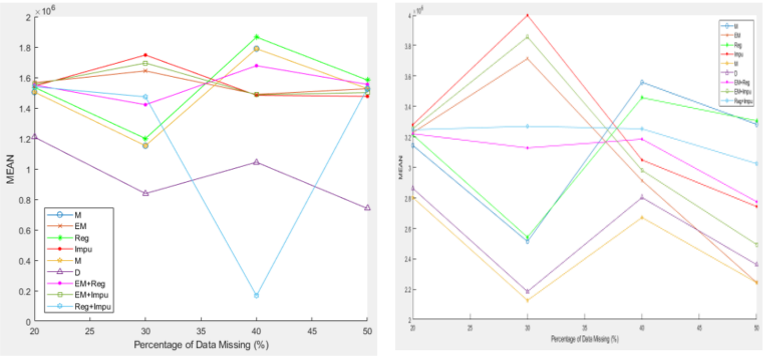
FIGURE 1. Mean Squared Error (MSE) for exports (a) and imports (b) with missing data (20%, 30%, 40%, 50%).

TABLE 3. Mean for exports with missing data (20%, 30%, 40%, 50%)

Method	M.S.E (20%)	M.S.E (30%)	M.S.E (40%)	M.S.E (50%)	Mean
Full data	1551963.9	1551963.9	1551963.9	1551963.9	1551964
Missing data	1504098.9	1151980.8	1789158.6	1525880.4	1492780
EM	1566029.8	1642813.1	1489563.2	1525880.4	1556072
Regression	1535320.9	1198979.8	1864847.7	1582776.1	1545481
MI	1543296.2	1747281.3	1483182.9	1476354.6	1562529
Mean	1504098.9	1151980.8	1789158.6	1525880.4	1525880.4
Median	1208236.2	837738.7	1043658.2	739664.5	957324.4
EMR	1550674.5	1420896.5	1677205.5	1554328.3	1550776
EMMI	1554663	1695047.2	1486373.1	1501117.5	1559300
RMI	1539308	1473130.6	167401.5	1529565.4	1177351
Mean for 8 methods	1500203	1395984	1375174	1429446	1425202

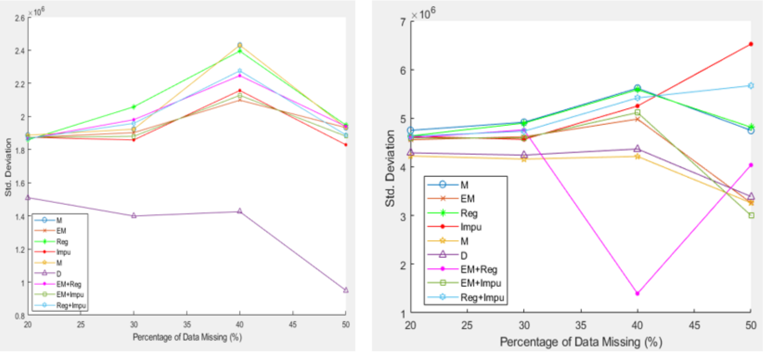
TABLE 4. Mean for imports with missing data (20%, 30%, 40%, 50%)

Method	M.S.E (20%)	M.S.E (30%)	M.S.E (40%)	M.S.E (50%)	Mean
Full data	1932310.1	1932310.1	1932310.1	1932310.1	1932310
Missing data	1887365.7	1923157.5	2430615.7	1932737.7	2043469
EM	1872832.6	1902433.1	2098522.4	1932737.7	1951631
Regression	1858496.2	2058464.9	2394987.8	1948038.1	2064997
MI	1874729.2	1858427.3	2157041.6	1829115.4	1929828
Mean	1887365.7	1923157.5	2430615.7	1932737.7	2043469
Median	1509985.2	1398905.1	1425035.1	949663.2	1320897
EMR	1865664.4	1980449	2246755.1	1940387.9	2008314
EMMI	1873780.9	1880430.2	2127782	1880926.6	1940730
RMI	1866612.7	1958446.1	2276014.7	1888576.8	1997413
Mean for 8 methods	1826183	1870089	2144594	1787773	1907160



(*a*) (*b*)

FIGURE 2. Mean for exports (a) and imports (b) with missing data (20%, 30%, 40%, 50%).



(*a*) (*b*)

FIGURE 3. Std. Deviation (SD(for exports (a) and imports (b) with missing data (20%, 30%, 40%, 50%).

TABLE 5. Mean for exports with missing data (20%, 30%, 40%, 50%)

Method	M.S.E (20%)	M.S.E (30%)	M.S.E (40%)	M.S.E (50%)	Mean
Full data	3235685.3	3235685.3	3235685.3	3235685.3	3235685
Missing data	3144161.5	2513025.7	3559023.1	3282995.1	3124801
EM	3227932.8	3714259.7	2913246.1	2243278.1	3024679
Regression	3213410.7	2541543.5	3457720.9	3305901.5	3129644
MI	3280151.2	3998940.5	3048469.4	2743029.9	3267648
Mean	2802277.5	2126142.6	2670706.7	2243278.1	2460601
Median	2862088.5	2183968.7	2801920.2	2361181.7	2552290
EMR	3220671.8	3127901.6	3185483.5	2774589.8	3077162
EMMI	3254042	3856600.1	2980857.8	2493154	3146163
RMI	3246780.9	3270242	3253095.2	3024465.7	3198646
Mean for 8 methods	3138419	3102450	3038937	2648610	2982104

the proposed composite methods perform better than the single methods: median method in terms of mean criterion. In terms of MSE, for small missing percentage, the EMR method always results in better MSE than that of mean, median, and MI methods.

Acknowledgments Authors are thankful to the referee(s) for their useful comments.

Table 6. Mean for imports with missing data (20%, 30%, 40%, 50%)

Method	M.S.E (20%)	M.S.E (30%)	M.S.E (40%)	M.S.E (50%)	Mean
Full data	4700251.9	4700251.9	4700251.9	4700251.9	4700252
Missing data	4749958.6	4916515.8	5617089.7	4749377.3	5008235
EM	4557403.2	4618395.9	4982288.1	3260561.1	4354662
Regression	4635869.1	4899525.8	5585960.9	4814920.3	4984069
MI	4626114.7	4564700.5	5250062.1	6523706.3	5241146
Mean	4224048.3	4159612.8	4215088.9	3260561.2	3964828
Median	4287432.4	4242177.1	4367320	3390456.9	4071847
EMR	4596636.2	4758960.9	1396490.2	4037740.7	3697457
EMMI	4591758.9	4591548.2	5116175.1	3001795.5	4325319
RMI	4630991.9	4732113.2	5418011.5	5669313.3	5112607
Mean for 8 methods	4518782	4570879	4541425	4244882	4468992

References

[1] D. B. Rubin, *Inference and missing data*, Biometrika, 63(3) (1976), 581–592.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the royal statistical society: series B (methodological), 39(1) (1977), 1–22.

[3] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*, Cambridge university press, (2006).

[4] J. P. Vandenbroucke, E. V. Elm, D. G. Altman, P. C. Gtzsche, C. D. Mulrow, S. J. Pocock, ... and Strobe Initiative, *Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration*, Annals of internal medicine, 147(8) (2007), W-163.

[5] J. C. Jakobsen,C. Gluud, J. Wetterslev, and P. Winkel, *When and how should multiple imputation be used for handling missing data in randomised clinical trialsa practical guide with flowcharts*, BMC medical research methodology, 17(1) (2017), 1–10.

[6] J. R. Dettori, D. C. Norvell, and J. R. Chapman, *The sin of missing data: is all forgiven by way of imputation?*, Global spine journal, 8(8) (2018), 892–894.

[7] R. J. Little and D. B. Rubin, *Statistical analysis with missing data (Vol. 793)*, John Wiley and Sons, (2019).

[8] S. W. Narayan, K. Yu Ho, J. Penm, B.Mintzes, A. Mirzaei, C. Schneider and A. E. Patanwala, *Missing data reporting in clinical pharmacy research*, American Journal of Health-System Pharmacy, 76(24) (2019), 2048–2052.

[9] C. K. Enders, *Applied missing data analysis*, Guilford Publications, (2022).

[10] A. Mirzaei, S. R. Carter, A. E. Patanwala and C. R. Schneider, *Missing data in surveys: Key concepts, approaches, and applications*, Research in Social and Administrative Pharmacy, 18(2) (2002), 2308–2316.

[11] M. Asif and K. Samarth, *Imputation methods for multiple regression with missing heteroscedastic data*, Thailand Statistician, 20(1) (1976), 1–15.

[12] A. Nouraldin, B. Fathi Vajargah, S. Baghar Mirashrafi, *A new approach for imputation missing data using partition with Expectation maximization method*, Computational Sciences and Engineering (CSE), (2023).