Computational Sciences and Engineering 3(1) (2023) 143-152

University of Guilan

Computational Sciences and Engineering

journal homepage: https://cse.guilan.ac.ir/

A new approach for imputation missing data using partition with Expectation – maximization method

Ahmad Nouraldin[a], Behrouz Fathi Vajargah [b,*], Seyed Bagher Mirashrafi[c].

[a] Department of Applied Mathematics, University of Guilan, Rasht, Iran
[b] Department of Statistics, University of Guilan, Rasht, Iran
[c] Department of Statistics, University of Mazandaran, Babolsar, Iran

ARTICLE INFO

Article history:
Received 14 November 2023
Received in revised form 8 December 2023
Accepted 11 December 2023
Available online 11 December 2023

Keywords:
Imputation
Missing data
Partitioned data
Expectation–maximization

ABSTRACT

The issue of missing data is a pervasive challenge in research, posing a significant obstacle to the reliability and validity of study findings. To address this issue, researchers have developed numerous approaches for replacing missing values. In this study, we focus on one such method for imputing missing data. Specifically, our paper introduces a novel technique for addressing missing data (latent variables) by implementing a partitioning strategy for the data that contains these missing values. Subsequently, we utilize the Expectation-Maximization (EM) method to compensate for the missing values within each resulting partition. Our findings demonstrate the efficacy of segmenting data that includes missing values, revealing that employing a higher degree of segmentation leads to improved estimation accuracy. To evaluate the performance of our approach, we compared the results using two key indices, namely Mean Squared Error (MSE) and Standard Deviation (S.D), across complete data, missing data, and partitioned data scenarios. Notably, our analysis focused on situations where data loss completely at random within real-world datasets. In summary, this research contributes a new and effective method for addressing the challenge of missing data through data segmentation and the application of Expectation-Maximization techniques. Our results highlight the potential of this approach to enhance the accuracy and reliability of data analysis in the presence of missing values.

* Corresponding author.

E-mail addresses: behrouz.fathi@gmail.com (B. Fathi Vajargah)

https://doi.org/10.22124/cse.2023.26024.1066

1. Introduction

In today's world missing data is a major problem when dealing with real-world situations [1]. In many real-life situations, we encounter data with missing values. Thus, effective missing data handling techniques have many applications [2].

There are different methods for replacing missing data [3-5].The deletion technique is very easy to apply and does not require much knowledge about statistics. Various imputation methods have been proposed to replace missing values [6]. The purpose of imputation is to replace all missing values to obtain better estimators. Imputation missing data in statistical analysis is important for several reasons:

1. *Accurate results:* Missing data can lead to biased and inaccurate results. By appropriately handling missing data, we can ensure that the statistical analysis produces reliable and trustworthy results [7].
2. *Avoiding bias:* If missing data is not handled properly, it can introduce bias into the analysis, leading to incorrect conclusions and decisions. Proper handling of missing data helps to minimize bias and ensure the validity of the analysis [8].
3. *Maximizing the use of available information:* By handling missing data, we can make the most of the available information and utilize all the data points to improve the accuracy and precision of the statistical analysis [9].
4. *Maintaining statistical power:* Missing data can reduce the statistical power of the analysis, making it difficult to detect true effects and relationships. Proper handling of missing data helps to maintain the statistical power and improve the ability to detect meaningful patterns and associations in the data [7].
5. *Meeting research standards:* Many research and scientific journals require researchers to address missing data in their statistical analyses. Proper, imputation of missing data ensures that the analysis meets the standards and requirements of the research community [10].

Overall, imputation missing data in statistical analysis is crucial for producing reliable, unbiased, and accurate results that can be used to make informed decisions and draw valid conclusions.

Handling missing data is crucial in real-world situations where data is collected from various sources, such as surveys, medical records, or financial reports. Failure to handle missing data appropriately can have significant consequences, including: Firstly, incorrect decisions: If missing data is not handled properly, it can lead to incorrect decisions, such as misdiagnosis of a medical condition or incorrect financial forecasting. Secondly, loss of valuable information: Missing data can reduce the amount of information available for analysis, leading to a loss of valuable insights and opportunities for improvement. Thirdly, wasted resources: Improper, imputation of missing data can lead to wasted resources, such as time and money spent on collecting and analyzing incomplete data. Finally, legal and ethical issues: In some cases, failure to handle missing data appropriately can lead to legal and ethical issues, such as violating privacy laws or misinforming stakeholders [7].

For example, in medical research, missing data can lead to incorrect conclusions about the effectiveness of a treatment or medication, potentially putting patients at risk. In financial analysis, missing data can lead to inaccurate forecasting and investment decisions, resulting in financial losses for individuals and organizations.

Overall, imputation missing data appropriately is essential for making informed decisions, improving outcomes, and avoiding negative consequences in real-world situations.

This article discusses the imputation method for incomplete data by performing partitioning of the data containing the missing values, and then we use the EM algorithm to compensate for the missing values for each resulting part. We take into consideration the missing data percentages of 20%, 30%, 40%, and 50%. EM algorithm is essentially a variant of maximum likelihood estimation and is capable of imputation missing data, its applications in fault detection, signal detection and filtering in presence of missing data are interesting directions to explore [11].

The rest of the article addresses these topics: The first section discusses the missing data mechanism. The next section deals expectation–maximization Method of missing data. It also briefly explains the method used in this article. In the following, the data set used in this research is examined. In the next section, the experimental results of data assignment are presented by EM method. In the last section, the results of this research are presented.

2. Missing data mechanism

One of the most crucial issues to consider in the study of data with missing values is the missing data mechanism. The missing data mechanism expresses the relationship between missing data and response values in the data matrix. To understand the data missingness mechanism, the most commonly available models are outlined as follows [3],[12]:

1. Missing completely at random (MCAR): In MCAR, the missing value mechanism is independent of variable values, whether observed or missing. According to the MCAR mechanism, the observed data are random samples.
2. Missing at random (MAR): MAR requires that the cause of the missing values be unrelated to the missing values, but may be related to the observed values of other variables.
3. Missing not at random (MNAR): This mode of missing data is dependent on both observed and unobserved responses.

3. Expectation–Maximization Algorithm

3.1. What is MLE?

Suppose we have a data set and we supposed that it follows the distribution $f(x|\theta)$, θ is the parameter of this (given) distribution. If we want to estimate θ, we use MLE method. But, we cannot use MLE method in all situations, such as the case we have some data missing (Latent variables). So, in this type of problems we use Expectation-Maximization Algorithm.

3.2. EM Algorithm

The Expectation-Maximization (EM) algorithm is a method to find MLE of the parameters of a statistical model in case where the equations cannot be solved directly.

Gaussian mixture is a kind of statistical model which involves latent variables and hence can not be solved directly using MLE method.

Latent variables refer to some case such as personal identification e.g., the measure of intelligent, the measure of depuration (where we cannot measure them). We call these of types data as latent variables.

In machine learning, clustering is an example for missing data problems. Here the missing data are the cluster labels.

Missing data problems means that problem contains some latent variables. We can assume that clustering problem follow Gaussian mixture model.

Gaussian mixture models can be used to clusters unlabeled data points. That is, we do not know what samples came from which class, our goal is to use Gaussian mixture models to assign the data points to the appropriate cluster.

Since Gaussian mixture model contains latent variables, we apply EM algorithm to solve the problem.

We can not use the MLE method to estimate parameters, we have to use EM method to estimate the parameters of our probabilistic model.

3.3. Outline of EM algorithm

*Step 1***:** Initialize the parameters θ to be estimated.

*Step 2***:** Expectation step (E-step) - using the observed variable data of the set, estimate (guess) the values of the missing data.

Step 3: Maximization step (M-step) - complete data generated after the expectation step is used the parameters, by maximizing likelihood function.

Step 4: Repeat steps 2 and 3 until converge.

Now, how EM algorithm can be used for Gaussian mixture parameters?

Problem: Suppose we are given a set of N observations $\{x_1, x_2, \dots, x_N\}$ of a numeric variable X.

Let X be a mix of k normal distributions and the probability density are $f_1(x), f_2(x), \dots, f_N(x)$.

Let $X = \{x_1, x_2, \dots, x_N\}$ making k=3 clusters

Cluster1 Cluster2 Cluster3

$f_1(x)$ $f_2(x)$ $f_3(x)$

It means that each cluster has a distribution function. We can say the $X = \{x_1, x_2, \dots, x_N\}$ is the mixture of k normal distributions. We do not know the cluster labels. In the case that Gaussian mixture of k probability distributions.

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \cdots + \pi_k f_k(x) \quad (1)$$

$$\pi_1 \geq 0, \qquad i = 1,2, \dots, k$$

$$\pi_1 + \pi_2 + \cdots + \pi_k = 1,$$

and $X \sim N(\mu_i, \sigma_i^2) \Rightarrow f_i(x) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}}$, $i = 1,2, \dots, k$, where μ_i and σ_i^2 are mean and variance respectively.

Then $\begin{cases} \mu_1, \ \mu_2, \dots, \mu_k \\ \sigma_1, \ \sigma_2, \dots, \sigma_k \\ \pi_1, \ \pi_2, \dots, \pi_k \end{cases}$ are the all parameters of Gaussian mixture distribution and must be estimated.

Let θ denotes the set of parameters μ_i, σ_i^2 and π_i $(i = 1,2, \dots, k)$.

$$L(\theta) = \prod_{i=1}^{N} f(x_i) \quad (2)$$

$$\begin{aligned} \ell(\theta) = ln\, L(\theta) &= ln\left(\prod_{i=1}^{N} f(x_i)\right) \\ &= lnf(x_1) + lnf(x_2) + \cdots + lnf(x_N) \\ &= \sum_{i=1}^{N} ln\, f(x_i) \\ &= \sum_{i=1}^{N} ln[\pi_1 f_1(x_i) + \pi_2 f_2(x_i) + \cdots + \pi_N f_N(x_i)] \\ &= \sum_{i=1}^{N} ln\left[\frac{\pi_1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} + \frac{\pi_2}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}} + \cdots + \frac{\pi_N}{\sigma_N\sqrt{2\pi}} e^{-\frac{(x_i-\mu_N)^2}{2\sigma_N^2}}\right] \end{aligned}$$

This is the likelihood function of Gaussian mixture. Now, we have the algorithm (below).

Step 1: Initialize the means μ_i's, the variances σ_i^2's and the mixture coefficients π_i's.

Step 2: Calculate the following, (for $n = 1,2, \dots, N$, and $i = 1,2, \dots, k$),

$$\gamma_{in} = \frac{\pi_i f_i(x_n)}{\sum_{i=1}^{n} \pi_i f_i(x_n)}, \quad (3)$$
$$N_i = \gamma_{i1} + \gamma_{i2} + \cdots + \gamma_{iN},$$

Step 3: Recalculate the parameters using μ_i's, σ_i^2's, π_i's as following:

$$\mu_i = \frac{1}{N_i}(\gamma_{i1}x_1 + \gamma_{i2}x_2 + \cdots + \gamma_{iN}x_N), \quad i = 1,2, \dots, k, \quad (4)$$
$$\sigma_i^2 = \frac{1}{N_i}[\gamma_{i1}(x_1 - \mu_i)^2 + \gamma_{i2}(x_2 - \mu_i)^2 + \cdots + \gamma_{iN}(x_N - \mu_i)^2],$$
$$\pi_i = \frac{N_i}{N}.$$

Step 4: Evaluate the log-likelihood function and check the for convergence of either parameter as log-likelihood function. If converge, then stop; *Else* go to step 2.

4. Data partition (Proposed method)

In this section, we introduce the studied data partition mechanism. We divide the data twice, four, eight, and twenty times, respectively. For example, when the data is divided into two parts, the methods of imputation for the deleted data previously mentioned are applied to each new part, and then we get complete data. When we divide the data into four parts, we also apply the same method of imputation to the four new parts, so we get complete data. In the same manner also when dividing the data into eight parts and then into twenty parts. Note that method of imputation for the deleted data used and applied to each output part depend on the values of the same output part after partition. Therefore, increasing the partition gives a better result for the estimators, and this is confirmed by the results that we obtained.

To illustrate the case of partitioning the data into two parts, we assume we have 100 observations as follows:

$$x_1, x_2, x_3, \dots x_{100.} \tag{5}$$

Then by dividing this data into two parts, we will get two sets of data as follows:

$$x_1, x_2, x_3, \dots x_{50} \quad \& \quad x_{51}, x_{52}, x_{53}, \dots x_{100.} \tag{6}$$

After dividing we delete some data from each new part in different proportions (20%, 30%, 40%, 50%), and the size of the deleted data will be 10, 15, 20, and 25, respectively. Based on the remaining data, they replaced missing data (40, 35, 30, 25) with the imputation method and compared them in the case of complete data before deletion and before data division. Also, this matter shows us, after using the imputation method, the effectiveness of the proposed method (Data partition) and the method used.

Based on the proposed method, we can write an algorithm for a sample of observations that follows a specific distribution in the following form

1. Generating n number from any given distribution.
2. Do partition on the generated data.
3. Delete some data in each new part (The missing percentage).
4. Use an EM method for the imputation of the missing values in each part.
5. Calculating estimates and comparing them in the case of the full data before deletion.

5. Application

This section includes the results of the study of the proposed partition method on real data, through which we compare the studied estimations of the imputation method (EM). We apply the imputation method to a real dataset (Exports dataset). The dataset was extracted from Waterborne Container Trade by the US Customs Port (2000-2017). The assessment of the imputation method was based

on the mean squared error (MSE) and (S.D) standard deviation of the estimators with various missing percentages.

The missingness was by the MCAR mechanism. The proportions of missingness were 20%, 30%, 40 and 50%. Then, we used an EM algorithm to compensate for the missing values in the case for each part we got after partition. The standard deviation and MSE of the estimators of the used imputation method were then computed. all simulations were accomplished by using Spss and Matlab software.

Tables 1-4 show the standard deviations and MSEs of the estimators obtained by EM imputation method and when the proportions of missingness were 20%, 30%, 40% and 50%, respectively. ***Figures 1,2*** display the Mean Squared Error (MSE) and Std. Deviation (S.D) for estimators for exports with missing percentage (20%, 30%, 40%, 50%) by partition (1, 2, 4, 8, 16), respectively.

Figures 3,4 illustrate Mean Squared Error (MSE) Std. Deviation (S.D) for exports by Partition with missing data (20%, 30%, 40%, 50%), respectively.

Table 1. Standard deviations and MSEs of the estimators obtained by EM imputation method with missing percentage 20%

Partition / Indicator	Exports Full data	Exports Missing data	Without partition	Two partition	Four partition	Eight partition	sixteen partition
N Valid	1134	901	1134	1134	1134	1134	1134
Missing	0	233	0	0	0	0	0
MSE	96085.9	104747.2	95855.7	83365.2	88721.7	90480.1	91622.5
Std. Deviation	3235685.3	3144161.5	3227932.8	2807316.2	2987697.1	3046908.4	3085380.6

Table 2. Standard deviations and MSEs of the estimators obtained by EM imputation method with missing percentage 30%

Partition / Indicator	Exports Full data	Exports Missing data	Without partition	Two partition	Four partition	Eight partition	sixteen partition
N Valid	1134	812	1134	1134	1134	1134	1134
Missing	0	322	0	0	0	0	0
MSE	96085.9	88189.9	110297.5	72627	73962.1	80336.7	92154.9
Std. Deviation	3235685.3	2513025.7	3714259.7	2445709	2490668.8	2705333.1	3199808.9

Table 3. Standard deviations and MSEs of the estimators obtained by EM imputation method with missing percentage 40%

Partition / Indicator	Exports Full data	Exports Missing data	Without partition	Two partition	Four partition	Eight partition	sixteen partition
N Valid	1134	639	1134	1134	1134	1134	1134
Missing	0	495	0	0	0	0	0
MSE	96085.9	140792.8	86510.9	91450.4	91984.2	93884.6	98243.5
Std. Deviation	3235685.3	3559023.1	2913246.1	3079583.3	3090356.9	3161556.7	3308341.8

Table 4. Standard deviations and MSEs of the estimators obtained by EM imputation method with missing percentage 50%

Partition / Indicator	Exports Full data	Exports Missing data	Without partition	Two partition	Four partition	Eight partition	sixteen partition
N Valid	1134	530	1134	1134	1134	1134	1134
Missing	0	604	0	0	0	0	0
MSE	96085.9	142604.2	66615.7	66884.5	67313.9	69810.4	95211.5
Std. Deviation	3235685.3	3282995.1	2243278.1	2252328.7	2266791.7	2350858.4	3206237.9

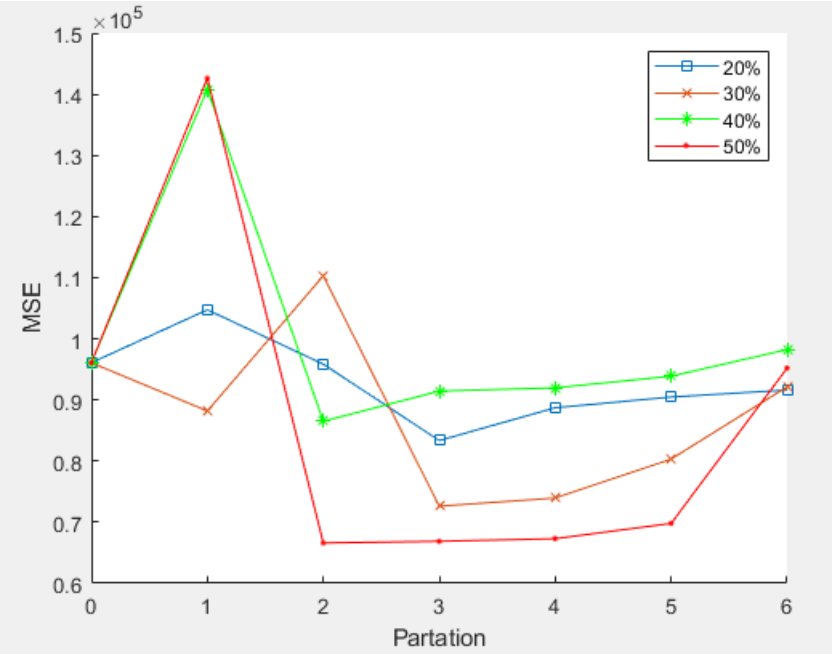

Figure 1. Mean Squared Error (MSE) for exports with missing percentage (20%, 30%, 40%, 50%) by partition (1, 2, 4, 8, 16)

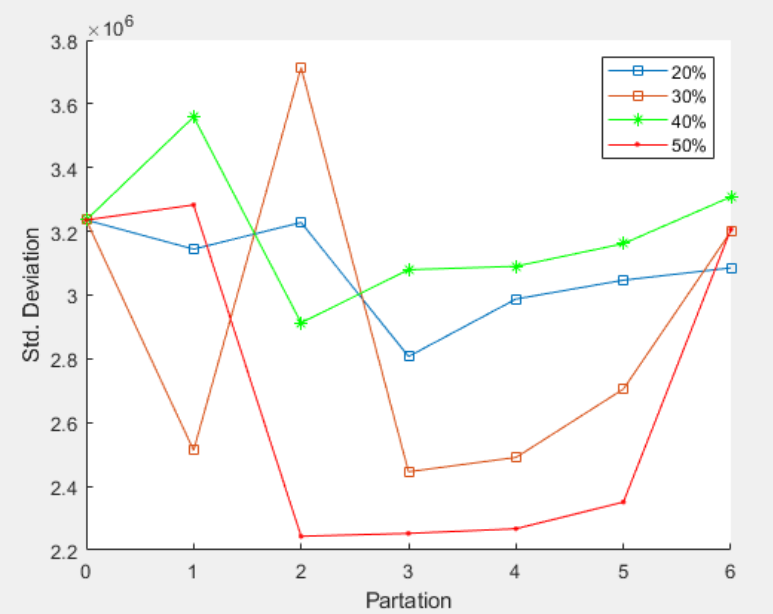

Figure 2. Std. Deviation (S.D) for exports with missing percentage (20%, 30%, 40%, 50%) by partition (1, 2, 4, 8,16)

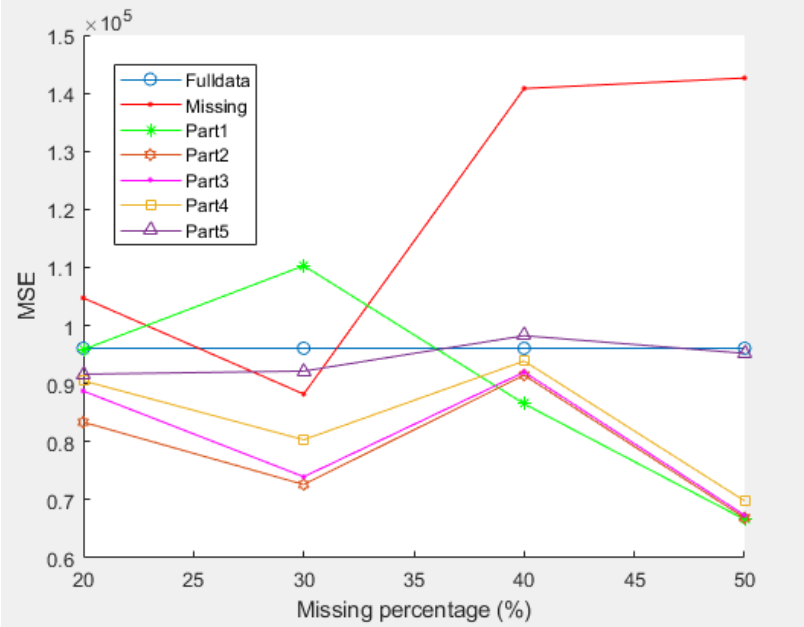

Figure 3. Mean Squared Error (MSE) for exports by Partition with missing percentage (20%, 30%, 40%, 50%)

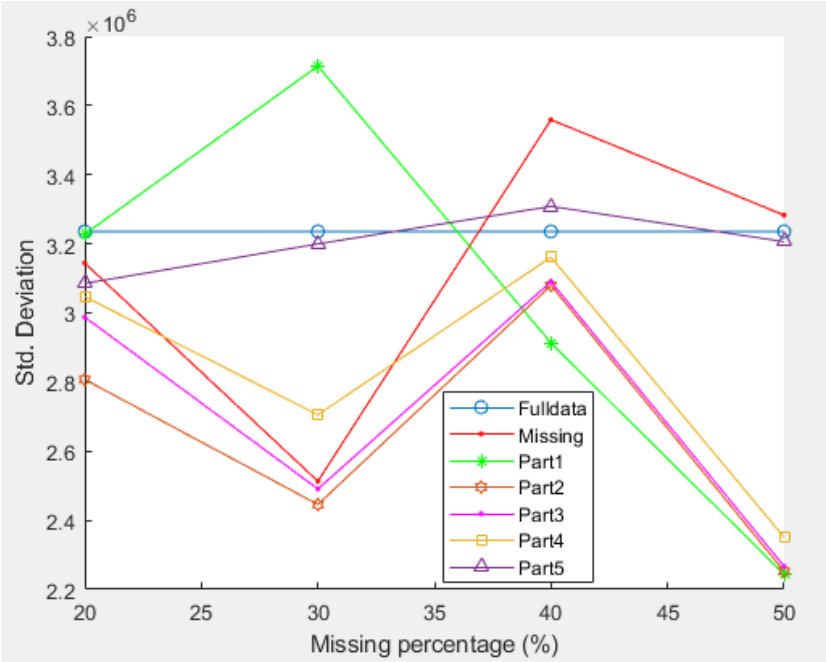

Figure 4. Std. Deviation (S.D) for exports by Partition with missing data (20%, 30%, 40%, 50%)

6. Discussion and Conclusion

In this study, we examined the performance and efficiency of the EM imputation method and compare results after data partition. This article compared seven states of partition (Exports with full data, Exports with Missing data, without partition, 2 partition, 4 partition, 8 partition, and 16 partitions) with missing data percentages (20%, 30%, 40%, 50%) for one variable (Exports). Our results indicated the efficiency of the proposed method algorithm. We note by using this method, the MSES (for partition states) approaches the MSES in the case of the complete data before the deletion and begins to give a better result than the previous one as we increase the division

respectively. Where we find that the MSE index in the case of complete data is equal to 96085.9, and in the case of 16 sections it is 91622.5 (20% missing data). The MSE index of the remaining missing cases (30%, 40%, 50%) is 92154.9, 98243.5, and 95211.5 respectively. On the other hand, when the missing percentage is high (40-50%), the proposed method is efficient. This is also applicable for the second index (MSE), where the MSE value improves and approaches of MSE value in the case of full data.

In order to continue working in the future, we recommend that you generalize the partition case and also apply other methods in order to impute the missing values using the method suggested in this article.

Acknowledgements

Authors are grateful to there anonymous referees and editor for their constructive comments.

References

[1] García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications, 19*, 263-282.

[2] Choudhury, S. J., & Pal, N. R. (2019). Imputation of missing data with neural networks for classification. *Knowledge-Based Systems, 182*.

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

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

[5] Asif, M., & Samart, K. (2022). Imputation Methods for Multiple Regression with Missing Heteroscedastic Data. *Thailand statistician, 20(1)*, 1-15.

[6] Lamjaisue, R., Thongteeraparp, A., & Sinsomboonthong, J. (2017). Comparison of missing data estimation methods for the multiple regression analysis with missing at random dependent variable. *Thammasat International Journal of Science and Technology*, *25*(5), 676-777.

[7] Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology, 64(5)*, 402-406.

[8] Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation, 7(4)*, 353-383.

[9] Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate behavioral research, 31(2)*, 197-218.

[10] Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods, 17(4)*, 372-411.

[11] Sammaknejad, N., Zhao, Y., & Huang, B. (2019). A review of the expectation maximization algorithm in data-driven process identification. *Journal of process control, 73*, 123-136.

[12] KA, N. D., Tahir, N. M., Abd Latiff, Z. I., Jusoh, M. H., & Akimasa, Y. (2022). Missing data imputation of MAGDAS-9's ground electromagnetism with supervised machine learning and conventional statistical analysis models. *Alexandria Engineering Journal, 61(1)*, 937-947.