Turkish Journal of Science & Technology
16(1), 39-46, 2021

Correlation Based Regression Imputation (CBRI) Method for Missing Data Imputation

Uğur ÜRESİN[1*]
[1] Industrial Engineering, Faculty of Management, Istanbul Technical University, Istanbul, Turkey
[*1] uresin15@itu.edu.tr

(Geliş/Received: 12/11/2020; Kabul/Accepted: 31/01/2021)

Abstract: To complete missing values in a dataset is crucial for data mining and machine learning applications. If a parameter of a dataset has missing values, the values of the other parameters corresponding to those missing values should not be excluded from the dataset to prevent information in the dataset. Missing values should be handled carefully to avoid their affecting analyses and to prevent loss of information. There are many methods and practical applications to predict missing values (imputation) in a dataset. Most of the methods that consider other parameters to predict missing values generally require lots of observations. These methods do not give successful results for datasets with a low number of observations according to the number of parameters. In this study, an algorithm is proposed for small datasets to predict missing values. The performance of the proposed method is compared with methods that are generally used for small datasets. The proposed method (CBRI) was tested with real data on how much the dimensions of vehicle bodies produced by an automotive manufacturer deviated from design specifications. Some of the values deliberately and randomly removed from the real data were predicted by the proposed method and the standard error of the predictions was calculated. The results were compared with the arithmetic mean assignment and median value assignment methods using the same data, and much more successful results were obtained with the proposed method.

Keywords: Missing data, imputation, data manipulation

Eksik Veriler için Korelasyona Dayalı Regresyon Yöntemi İle Değer Atama

Öz: Veri madenciliği ve makine öğrenmesi uygulamalarında eksik verilerin tamamlanması oldukça önemlidir. Bir veri setinin herhangi bir parametresinde eksik değerler varsa, o eksik değerlere karşılık gelen diğer parametrelere ait değerlerin ölçümden çıkarılmaması gerekir. Aksi takdirde bilgi kaybı meydana gelecektir. Söz konusu bilgi kaybının oluşmaması için eksik veriler uygun bir şekilde tamamlanmalıdır. Bir veri kümesindeki eksik değerleri tahmin etmek (değer atamak) için birçok yöntem ve pratik uygulama vardır. Eksik değerleri tahmin etmek için diğer parametreleri dikkate alan yöntemlerin çoğu genellikle çok sayıda gözlem gerektirir. Bu yöntemler, parametre sayısına göre az sayıda gözlem içeren veri setlerinde başarılı sonuçlar vermemektedir. Bu çalışmada, eksik değerleri tahmin etmek için küçük veriler için bir algoritma önerilmiştir. Önerilen yöntemin performansı, genellikle küçük veri kümeleri için kullanılan yöntemlerle karşılaştırılmıştır. Önerilen yöntem, bir otomotiv üreticisinin ürettiği taşıt gövdelerindeki boyutların, dizayn spesifikasyonlarından ne kadar saptığını içeren gerçek bir veri ile test edilmiştir. Gerçek veriden kasıtlı ve rasgele olarak silinen değerler, önerilen yöntem ile tahmin edilmiş ve tahminlerin standard hatası hesaplanmıştır. Sonuçlar, aynı veri kullanılarak yapılan aritmetik ortalama atama ve medyan değer atama yöntemleri ile kıyaslanmış ve önerilen yöntem ile çok daha başarılı sonuçlar elde edilmiştir.

Anahtar kelimeler: Eksik veri, eksik değer atama, veri işleme

1. Introduction

The first process of data science and machine learning applications is the data wrangling process which consists of data gathering, assessment and cleaning. One of the tasks in this process is called imputation. The imputation method involves replacing the missing value with a value predicted based on data mining from existing information in the data set [1]. Missing data is a serious problem for researchers in many disciplines. Because traditional statistical methods and software assume that all variables are measured for all observations [2]. Besides missing data reduces the accuracy of the statistical models and produces biased predictions of a model, leading to invalid results [3]. In particular, missing observations are frequently encountered in raw data generated by collecting data from different sources. Missing observations are also encountered in the raw data generated intermittently due to physical constraints. Removing an observation with missing values in one or more parameters from the raw data may result in the loss of existing information drastically. Therefore, it is crucial to complete the missing data rationally.

* Corresponding author: uresin15@itu.edu.tr. ORCID Number of the author: [1] 0000-0002-9100-9697

There are different methods in the literature to predict missing observations. The selection of the appropriate method depends on incomplete observation mechanisms. These mechanisms are examined in three categories in the literature. These categories are 'missing completely at random' (MCAR), missing at random '(MAR) and 'not negligible' (NI) [4,5]. It is assumed that missing observations depend entirely on chance, and an X variable is not dependent on any other variable and X variable itself [5]. Incomplete observation due to chance is the incomplete observation of an X variable, it is dependent on other variables in the data set but not itself. That is, the absence of observation is that it depends on other variables except itself. The non-negligible missing case is that the missing value in an X variable depends on all variables in the data set, including the X variable. In other words, missing data is not due to chance.

The method of assigning the mean value is to replace the missing observations of a variable with the arithmetic mean of the complete observations for the variable. As a result of this operation, the arithmetic mean obtained for complete and assigned observations will be equal to the arithmetic mean obtained with complete observations. The positive aspect of this method is that it is easy to apply and takes into account all observations with complete information. The disadvantage of this method is that since all observations in the variables are replaced by a single value, the actual distribution of the variables changes and the variance in the data set decreases artificially. Correlation between this variable and other variables depending on the decrease in variance. In this approach, the data should be incomplete (MCAR) depending on the chance [6].

Another method frequently used for incomplete observation is the minimum assignment method where the value in the variable is used instead of the missing observation for the variable with missing observations. Another method is the maximum assignment method. In the maximum assignment method, for the variable with missing observations, the largest value in that variable is used [7].

In the median (median value) assignment method, it is used instead of the missing observation left in the middle after the observations in the variable containing missing observations are listed in order from small to large. If the size of the data set is odd, the median is the middle observation; If the number of observations is double, the median is the average of the two values in the middle [8].

Another method is the regression value assignment method. This method is used to predict missing values by establishing regression equations on variables that do not contain missing values. This method calculates multiple linear regression estimation and has options to increase the estimation with random components. In the first step of the regression value assignment method, regression equations that predict the missing value variables are obtained from the variables without missing values. In the second step, the values of the variables with missing values are predicted. These predictive values are used in place of missing values. Thus, a data set without missing values is created. In this method, the regression method is used as a tool to predict missing observations. If the missing observations are random (MCAR) and depend on other independent variables that do not contain missing observations, the least-squares coefficients are consistent [9]. Besides, this method gives almost unbiased results in large samples. This method is based on the assumption that there is a high correlation relationship between the variable with missing data and other variables. If this relationship is not sufficient, it is recommended to use other methods such as substituting the mean value instead of this method. However, it is very difficult for each parameter to have a high correlation relationship with each other for datasets with many parameters. If the parameters are not correlated with each other, the error rate of the created multiple regression models will be very high and the error of the predicted missing value will also be very high. Therefore, it is extremely critical to establish regression models correctly.

In this study, for the prediction of the missing observations, a simple regression model was established among the parameter that has the highest correlation and the parameter that has the missing observations. To test this method, dimensional data generated in a quality control process in the automotive industry were used. The dimensional data is a numerical data that includes how much certain points on a manufactured vehicle deviate from design specifications. Thus, measurement points on the vehicle are parameters of the data. Therefore, points that are physically close to each other are expected to be parameters that correlate with each other. This data was divided into two as training and test data to create regression models and to test the created model, and some values in the test data were randomly removed from the data, and the success of the regression models created with training data was tested. While creating the regression models here, not all parameters, but the parameter values with the highest correlation with the parameter with the missing observation value were used. The success of the results obtained from this approach was compared with the median and mean value assignment methods which are commonly used methods in practice when the number of observations is not much.

2. The Proposed Method and Algorithm

The proposed method is an application of simple linear regression-based imputation. In this study, different from the simple linear regression imputation, a criterion for determining the independent variable in the regression model is added. Pearson's pairwise correlation matrix was used to determine the dependent and independent variables of the regression model. This method is called Correlation Based Regression Imputation (CBRI for short) in this paper.

The flowchart for the algorithm is shown in Figure 1. When fitting a simple linear regression model, the 'parameter i' (contain a missing value or multiple missing values) is the dependent variable, the 'parameter j' (According to Pearson's pairwise correlation matrix, the parameter with the highest correlation with 'parameter i') is the independent variable. Also, outliers in these parameters are subtracted from the training data used to create the regression model. Unlike multiple regression, only the parameter with the highest correlation is considered for the missing value prediction in the CBRI method. Because it's expected that independent variables (or features) are uncorrelated. Therefore, As a matter of fact, including all indepent variables except the independent variable with missing data into the model will increase the standard error of the prediction.

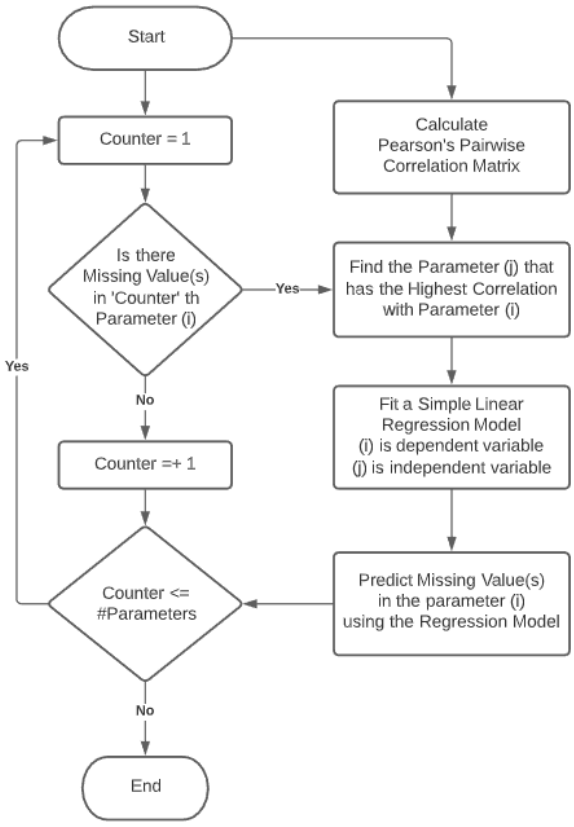

Figure 1. The flowchart of the CBRI algorithm

Besides, separate models are created for each missing value to decrease the standard error of the prediction. Therefore, fitting a simple linear regression model is repeated for each parameter that contains missing value or values. Thus, the high pairwise correlation relationship requirement for the regression imputation is ensured and that the parameters that are unrelated to each other do not affect the prediction of missing data. The difference between the proposed method and imputation using multiple regression model is that only the most correlated is considered to predict a parameter's missing values. Therefore, irrelevant features are not considered. Besides, multiple regression models require many observations per number of features. For regression models, there must be 10 observations per predictor variable at least [10]. In some cases, it is not always possible to collect 10 observations per predictor. If this is the case, fitting multiple regression models to predict missing values is likely

to yield high standard errors. However, the CBRI method predicts missing values by using just one predictor so that this method allows us to make good predictions even with a small number of observations.

3. Experimental study and results

To test the proposed method (CBRI) a real dataset was used. The data is called vehicle body dimensional data that includes how much the measured some assembly points on the vehicle body deviate from the design specifications. Vehicle body dimensional data is created in two steps. First, certain points on a vehicle body are measured with a device called CMM (Cartesian Measuring Modeling) just after the production of the vehicle body. Second, the deviations between measurement data and design specifications are calculated. These two processes are repeated on a large number of vehicle bodies to create data called vehicle body dimensional data.

The measurement points that are close to each other on the vehicle body have correlations due to their physical relationships. For example, there are some measurement points on the roof of the vehicle body and since most of the roof sections are continuous, close points are likely to have a high correlation for the bodies produced on same production day. Because the measurement pace is lower than the production pace, not all previously defined points can be measured. However, estimating unmeasured points with a reasonable standard error is especially important for the quality process. Therefore, this data is appropriate to use for the proposed method (CBRI) in this study.

3.1. Descriptive statistics of the dataset

The dimensional dataset consists of 66 parameters (measurement points) and 382 observations. All parameters contain float values. The descriptive statistics of some (since there are so many) parameters of the data are given in Table 1. In the dataset, some measurement points (parameters) that are close to each other show a high correlation relationship due to their physical and mechanical relationships.

Table 1. The descriptive statistics of some parameters in the data

Parameter	Min	1st Qua.	Median	Mean	3rd Qua.	Max
1	-1,85	-0,7875	-0,27	-0,1874	0,2775	2,07
5	-2,46	-0,7975	-0,245	-0,2466	0,2775	2,03
10	-5,47	-0,9375	0,155	-0,01976	0,905	2,66
15	-1,64	-0,3975	0,24	0,2161	0,6975	2,29
20	-2,31	-0,0575	0,600	0,5674	1,2775	3,49
25	-3,49	-0,76	-0,185	-0,351	0,01	3,61
30	-2,28	-0,63	-0,055	-0,2572	0,06	1,54
35	-2,00	-0,5975	0,295	0,4522	1,4225	3,61
40	-2,96	-1,24	-0,835	-0,709	0,045	1,48
45	-0,9	0,04	0,23	0,2385	0,44	1,21
50	-1,63	-0,1675	0,27	0,3429	0,7675	2,8
55	-3,61	-1,1475	-0,68	-0,7883	-0,33	0,8
60	-0,94	-0,3575	-0,1	-0,07872	0,1975	2,04
65	-3,1	-0,595	0,235	0,1897	1,0075	3,09

The descriptive statistics for all variables (66) in the data used in this study are shown in Figure 2. The minimum, maximum, median and mean values are given (the distributions of Quarter1 and Quarter3 values are not added to the figure to make the figure easily readable).

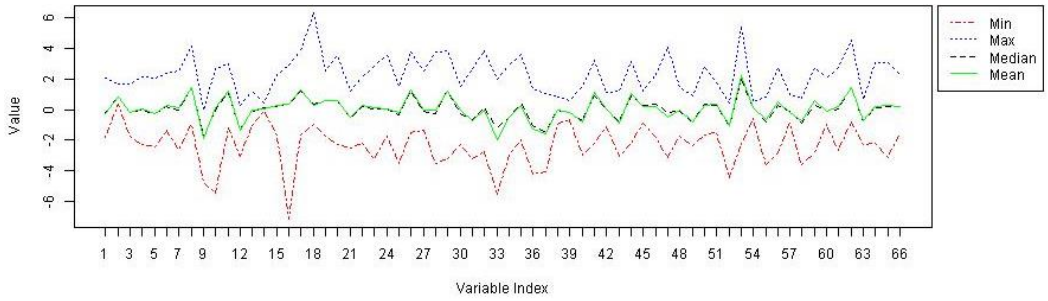

Figure 2. Some descriptive statistics values for all variables in the data

Considering, and the number of parameters, it can be said that this dataset contains a small number of observations. Therefore it's sutiable for the proposed method (CBRI) to predict missing values in the dataset. In the next section, the results are presented.

3.2. Results and Discussion

In this study, 20 random values (using sample.int() method in R programming language) are removed. The missing (removed) values are predicted Using the proposed method (CBRI),. These values are also predicted using the median assignment and mean assignment methods. Then standard errors are calculated. Random value removal was performed 5 times (5 iterations) to ensure randomness. The standard errors of the predictions for 5 iterations are shown in Figure 3 and Figure 4.

Standard Errors for 20 Random Missing Values - Iteration #1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Median Assignment	1,06	0,48	0,02	1,19	0,67	0,37	2,39	0,4	0	0	1,44	0,31	0,14	1,76	1,22	0,09	0,5	1,18	0,92	0,61
Mean Assignment	0,94	0,54	0,03	1,17	0,83	0,6	2,33	0,38	0,03	0,06	1,36	0,52	0,14	1,66	1,05	0,05	0,57	1,13	0,92	0,52
CBRI Method	0,31	0,08	0,23	0,47	0,55	0,05	0,14	0,16	0,08	0,07	0,07	0,02	0,38	0,64	0,22	0,08	0,31	0,08	0,22	0,46

Standard Errors for 20 Random Missing Values - Iteration #2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Median Assignment	1,32	0,78	0,71	2,04	0,34	1,6	0,27	0,89	2,06	1,47	0,15	0,66	0,35	1,78	0,4	0	0,45	1,35	0,21	0,64
Mean Assignment	1,41	0,99	0,89	1,92	0,29	1,6	0,28	0,79	1,72	1,7	0,14	0,58	0,21	1,57	0,48	0,01	0,43	1,51	0,25	0,67
CBRI Method	0	0,15	3,73	1,65	0,45	0,2	0,16	1,2	3,46	0,01	0,43	1,66	0,56	1,58	0,05	0,23	0,57	1,42	0,64	0,09

Standard Errors for 20 Random Missing Values - Iteration #3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Median Assignment	0,57	0,36	0,01	0,45	0,39	0,17	1,47	0,09	0,8	1,17	1,14	0,77	0,48	0,29	1,32	0,34	0,43	1,07	0,81	0,45
Mean Assignment	0,48	0,46	0,13	0,45	0,51	0,18	1,4	0,07	0,56	1,23	1,2	0,78	0,48	0,45	1,27	0,76	0,19	1,08	0,88	0,59
CBRI Method	0,06	0,03	0	0,33	0,04	0,04	0,12	0,08	0,06	0,5	0,1	0,05	0,43	0,01	0,09	0,32	0,02	0,01	0,19	0,21

Standard Errors for 20 Random Missing Values - Iteration #4

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Median Assignment	0,17	0,04	1,69	0,29	0,19	1,85	0,28	0,5	1,68	0,16	0,62	0,23	1,96	1,39	0,53	0,72	1,42	0,97	1,16	0,13
Mean Assignment	0,03	0,04	1,58	0,33	0,37	1,82	0,25	0,48	1,68	0,16	0,87	0,28	1,96	1,5	0,69	0,75	1,41	1,02	1,11	0,16
CBRI Method	0,12	0,45	1,36	0,1	0,05	0,48	0,33	0,13	1,36	0,17	0,02	0,13	1,72	0,15	0,12	0,4	0,06	0,06	0,23	0,06

Standard Errors for 20 Random Missing Values - Iteration #5

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Median Assignment	0,13	0,19	2,1	0,3	0,44	0,26	0,3	0,49	0,09	0,29	5,22	0,01	0,79	1,76	0,4	0,07	1,19	0,91	0,95	0,11
Mean Assignment	0,05	0,05	1,89	0,3	0,37	0,27	0,2	0,73	0,09	0,16	5,13	0,11	0,83	1,56	0,53	0,19	1,25	0,93	1	0,28
CBRI Method	0,05	0,17	0,79	0,17	0,11	0,09	0,18	0,05	0,02	0,13	3,55	0,21	0,32	0,39	0,18	0,12	0,28	0,18	0,35	0,03

Figure 3. Standard errors of the predictions for the 20 missing values
(red is the maximum standard error, yellow is the moderate standard error, green is the lowest standard error)
a)Iteration #1 b)Iteration #2 c)Iteration #3 d)Iteration #4 e)Iteration#5

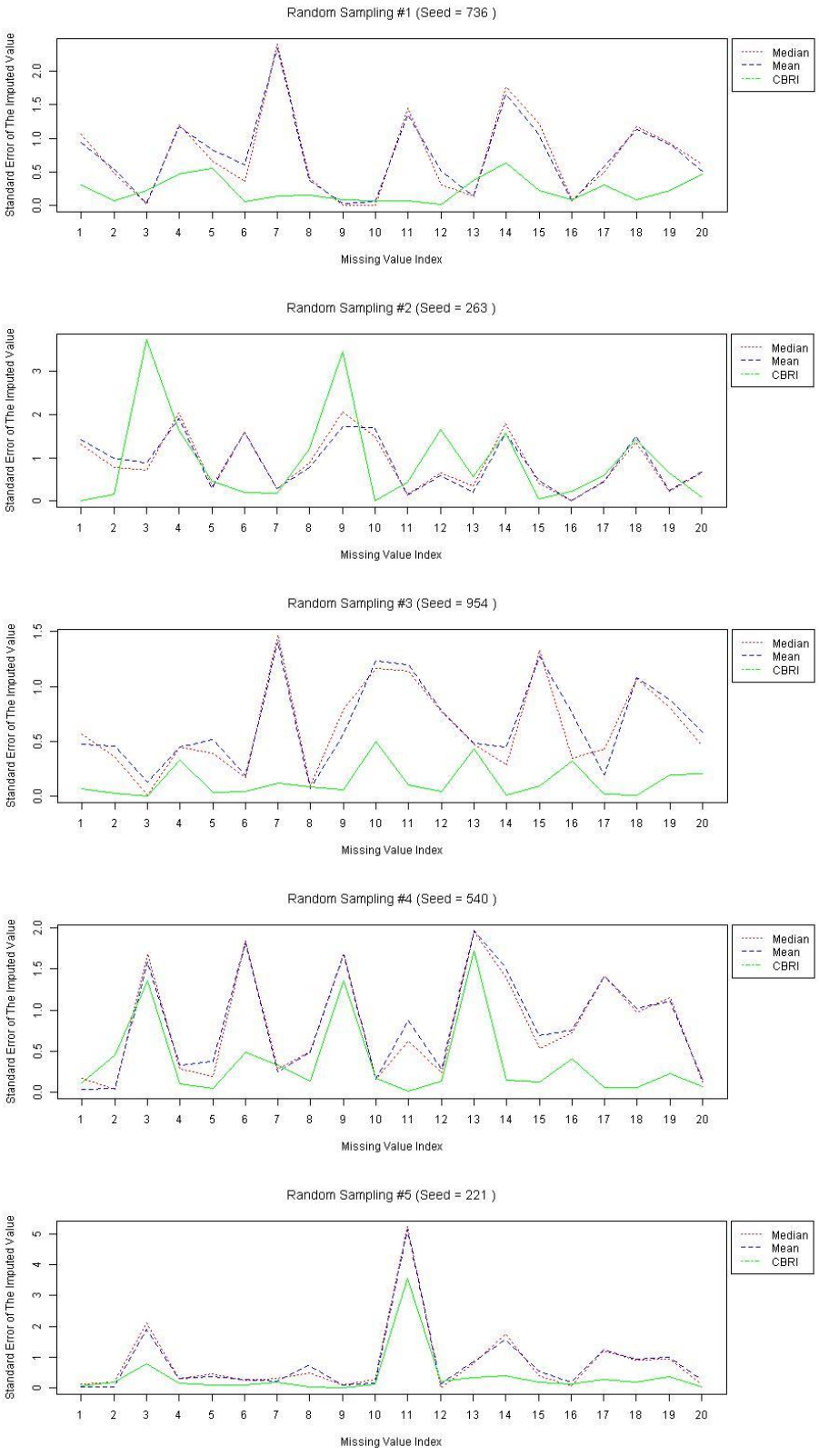

Figure 4. Comparison of the standard errors
a)Iteration #1 b)Iteration #2 c)Iteration #3 d)Iteration #4 e)Iteration#5

As shown in Figure 3 and Figure 4, assigning the median value and assigning the mean value methods yielded similar standard errors in the iterations of missing value prediction. In most of the iterations, the proposed method (CBRI) yielded a much less standard error. In the 5 iterations, a total of 100 randomly removed values were predicted using CBRI, mean assignment and median assignment methods. In 74% of the predictions, the CBRI method yielded the best prediction in terms of the minimum standard error. Since the volume of the data is very small (only 382 observations for the 66 parameters), it is very likely that a value that is randomly removed from the data is equivalent or very close to the mean and median value.

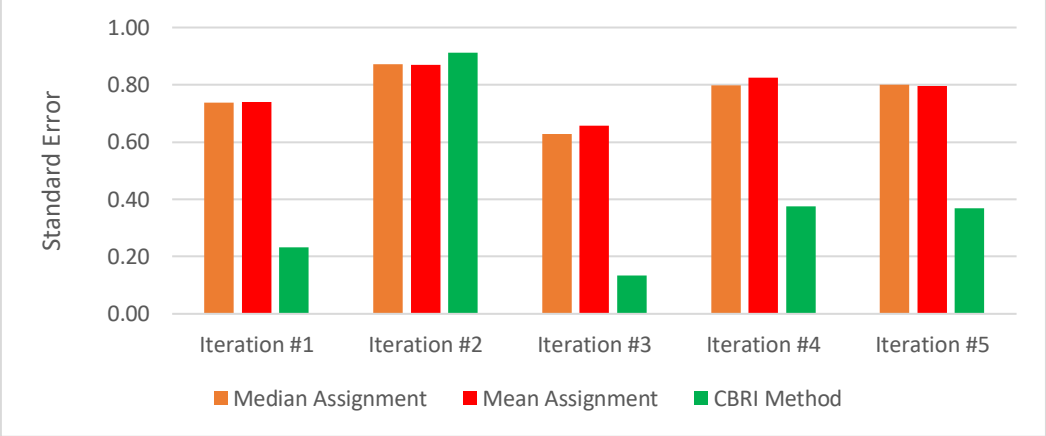

Figure 5. Comparison of the average standard errors in the iterations

In Figure 5, the average of the standard errors for the predictions in the iterations are shown. Considering the difference between the standard errors, CBRI is much better than median assignment and mean assignment methods in 4 of the 5 iterations. In iteration #2, although the CBRI method is not the best in terms of the average standard error, the difference between the average standard errors are significantly low. Thus, CBRI yielded better results in terms of both standard errors of the particular predictions and average standard errors of the predictions in the iterations.

5. Conclusion

In this study, a missing value assignment (imputation) method (CBRI) is proposed that takes into account the Pearson's pairwise correlations between parameters in a dataset. The proposed method has been tested using a real and relatively small data. The results are compared with the median assignment and mean assignment methods. 20 randomly chosen measurement values were removed from the data and these values were predicted by 3 methods and the standard errors of the predictions were compared. In conclusion, the proposed method (CBRI) was found more successful than median assignment and mean assignment methods.

As long as the distribution of data is not uniform, as the number of observations in the data increases, the probability that a missing value will differ from the mean or median value increases. Thus, the CBRI method is quite likely to perform much better in relatively larger data compared to mean and median assignment. In future, researchers may test the CBRI method's performance in bigger datasets and compare with more complex methods instead of mean and median assignment methods.

References

[1] Zain A.M., Ali N.A., Sallehuddin R. A Review On Missing Value Estimation Using Imputation Algorithm. Journal of Physics Conference Series 892(1):012004. 2017.

[2] Alkan B., Alkan N. Investigation of the Multiple Imputation Method in Different Missing Ratios and Sample Sizes. Sakarya University Journal of Science. 605-609. 2019.

[3] Kang H. The prevention and handling of the missing data. Korean J.Anesthesiol 64 (5): 402–406, 2013.

[4] Pelckmans K., Brabante, J.D., Suykens J.A.K., Moor B.D. Handling missing values in support vector machine classifiers. Neural Networks, 684-692. 2005.

[5] Jerez, J.M., Molina I., Subirats J.L., Franco L., Missing data imputation in breast cancer prognosis. Processing of the fourth IASTED International Conference Biomedical Engineering. 2006.

[6] Mohamed S. & Marwala T., Neural Network Based Techniques for Estimating Missing Data in Databases. 16th Annual Symposium of the Pattern Recognition Association of South Africa. 2005.
[7] Alpar, R., Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. 135-157. Nobel Kitabevi. 2010.
[8] Bal C., Özdamar K. Eksik Gözlem Sorununun Türetilmiş Veri Setleri Yardımıyla Çözümlenmesi. Osmangazi Üniversitesi Tıp Fakültesi Dergisi, 26 (2): 67-76. 2004.
[9] Menengiç Y. The Comparison of Missing Value Analysis Methods. MSc, Ondokuz Mayıs University, Samsun, Turkey, 2015.
[10] VanVoorhis C.R, Morgan B.L. Understanding Power and Rules of Thumb for Determining Sample Sizes. Tutorials in Quantitative Methods for Psychology vol. 3 (2), 43-50. 2007.