ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2018-62788

Performance Comparison of Some Imputation Methods Used in Missing Value(s) Analysis: A Simulation Study

Kayıp Değer Atamasında Kullanılan Bazı Yöntemlerin Atama Performanslarının Karşılaştırılması: Bir Benzetim Çalışması

Ahmet Kadir ARSLAN,[a]
Zeynep TUNÇ,[a]
Emek GÜLDOĞAN,[a]
Cemil ÇOLAK[a]

[a]İnönü University Faculty of Medicine,
Department of Biostatistics and
Medical Informatics,
Malatya, TURKEY

Received: 07.09.2018
Received in revised form: 02.10.2018
Accepted: 02.10.2018
Available Online: 19.04.2019

Correspondence:
Zeynep TUNÇ
İnönü University Faculty of Medicine,
Department of Biostatistics and
Medical Informatics, Malatya,
TURKEY/TÜRKİYE
zeynep.tunc@inonu.edu.tr

This study was presented as "Oral Presentation" at 11. International Statistics Days Conference, 03-07 October 2018, Muğla,Turkey.

ABSTRACT Objective: In a research, it is not desirable that the dataset to be used contains missing value (s) and researchers try to cope with this situation. The main purpose of this research is to develop new user-friendly web-based software that uses various techniques to handle missing value(s). **Material and Methods:** In this study, to assess the performance of the software, various scenarios were tested: 5 variables were normally distributed, different sample sizes (n=1000, 1500, 2000 and 2500), high (r <-0.70 or r> 0.70) and low correlations (-0.30 <r <0.30) among between variables, different number of missing value in variables (5%, 10% and 20% missing data). The missing values were imputed by the developed web software and the results were compared. Thus, the performance of the software under different conditions was evaluated. Shiny, an open source R package was used to develop the web tool. In the developed software, linear regression (LR), random forest (RF), classification and regression trees (CART) and predictive mean matching (PMM) methods were used to impute missing values. In order to achieve more unbiased and reliable results, the 'number of repetitions' and 'number of multiple imputations' sections were used in the software. The normalized root mean squared error (NRMSE) metric was used to assess performance of imputation techniques. The developed web-based application can be accessed free of charge at http://biostatapps.inonu.edu.tr/KDAY/. **Results:** According to the outputs of the developed web-based application, better results were obtained by LR and PMM models for missing value imputation in datasets with high correlation. For missing value imputation in low-correlated data sets, the models showed similar imputation performances. **Conclusion:** For the datasets used in this study, when the correlation between the variables is high, the best imputation performance is obtained with the DR and PMM models regardless of the size of the dataset and the percentage of missing values.

Keywords: Assignment methods; missing value(s) analysis; Shiny; simulation; web based software

ÖZET Amaç: Bir araştırmada kullanılacak veri setinin kayıp değer(ler) içermesi istenmeyen bir durum olup, araştırıcılar kayıp veri ile ilgili sorunları gidermeye çalışırlar. Bu araştırmanın temel amacı kayıp veri analizini ele almak için çeşitli teknikleri kullanan, yeni kullanıcı dostu bir web yazılımı geliştirmektir. **Gereç ve Yöntemler:** Bu çalışmada, yazılımın performansını değerlendirmek için çeşitli senaryolar test edilmiştir: 5 değişkenin normal olarak dağılması, Farklı örneklem büyüklüklerinin (n = 1000, 1500, 2000 ve 2500) olması, Değişkenler arasında yüksek (r <-0.70 veya r> 0.70) ve düşük korelasyonların (-0.30 <r <0.30) olması, Değişkenlerde farklı sayıda eksik değerlerin (% 5,% 10 ve% 20 eksik veri) olması Bu kayıp veriler geliştirilen web yazılımı ile doldurularak çıkan sonuçlar karşılaştırılmıştır. Böylece yazılımın farklı koşullardaki çalışma performansları değerlendirilmiştir. Açık kaynaklı bir R paketi olan Shiny, web aracını geliştirmek için kullanıldı. Yazılımımızda eksik değerlere atama yapmak için doğrusal regresyon (DR), rastgele orman (RF), sınıflandırma ve regresyon ağaçları (CART) ve tahmini ortalama eşleme (PMM) ele alındı. Kayıp veri atamalarından daha iyi sonuçlar alabilmek için yazılımda 'Tekrar sayısı' ve 'Çoklu Atama Sayısı' kısımları kullanıldı. Atama tekniklerinin performansını değerlendirmek için normalleştirilmiş hata kareler ortalamasının karekökü (NRMSE) metriği kullanılmıştır. Geliştirilen web tabanlı uygulamaya http://biostatapps.inonu.edu.tr/KDAY/ adresinden ücretsiz olarak erişilebilir. **Bulgular:** Geliştirilen web tabanlı uygulamanın çıktılarına göre yüksek korelasyona sahip veri setlerinde kayıp değer atama işlem için DR ve PMM modelleri ile daha iyi sonuçlar elde edilmiştir. Düşük korelasyona sahip veri setlerinde kayıp değer atama işlem için ise yazılımda yer verilen dört kayıp değer atama yönteminin hiçbirinin üstünlük sağlayamadığı görülmüştür. **Sonuç:** Bu çalışmada kullanılan veri kümeleri için, değişkenler arasındaki korelasyon yüksek olduğunda, verisetinin büyüklüğüne ve kayıp değerlerin yüzdesine bakılmaksızın DR ve PMM modelleri ile en iyi atama performansı elde edilmektedir.

Anahtar Kelimeler: Atama yöntemleri; benzetim; kayıp veri analizi; Shiny; web tabanlı yazılım

Missing values are considered as a common problem in most scientific research areas. Before doing the missing value analysis, it is necessary to investigate how these values emerge. Some studies have reported that missing values may occur due to 3 different situations. Data are missing completely at random (MCAR) when the probability of an instance (case) having a missing value for a variable does not depend on either the known values or the missing data. Missing at random (MAR) is defined when the probability of an instance having a missing value for a variable may depend on the known values but not on the value of the missing data itself. Finally, data are missing not at random (MNAR) when the probability of an instance having a missing value for a variable could depend on the value of that variable.[1,3] Missing value(s) can be caused by different reasons. For example, it may be due to individuals participating in the study leave some questions consciously or unconsciously, or fail to reach certain items within a specified period of time. In addition, it can be arisen from conditions such as inadequacy of data collection technique and unfavorable conditions of application. Missing value(s) is a major problem for almost all of the statistical methods to be used in the analysis phase. Since all these methods are developed under the condition that the data set is complete.[4,6] Researchers should complete the missing data to obtain accurate results from the analysis being conducted by them. Therefore, researchers use some methods. These methods include adding new observations to the database, extracting missing observations from the dataset, and using approximate values obtained by making estimates of missing data instead of missing data. There are many methods for assigning the approximate values instead of missing data. Methods such as mean substitution, median of nearby points, linear interpolation are called simple assignment methods. In addition, there are also methods such as expectation maximization algorithm, propensity score matching and Markov Chain Monte Carlo which are known as more advanced methods. Researchers can cope with the problems created by missing values with one of these methods.[7] The process of adding new observations leads to an increase in parameters such as time and effort. Extraction of observations including missing values from the dataset (list wise deletion) may lead to a substantial decrease in the number of observations. This causes the sample size to decrease and biased results. In addition, the power of statistical analysis to be used also decreases.[8,9] For this reason, assignment methods are used to deal with missing values.

As a result, open-sources software using machine learning methods for data processing tasks are needed more. Therefore the main purpose of this research is to develop a new user-friendly web tool that implements various techniques for assigning missing value.

MATERIAL AND METHODS

DATASET

Full data sets of different size and correlation with normal distribution will be simulated in order to evaluate the performance of the software in this study. Then, new data sets containing lost data at different ratios will be obtained with the aid of the data sets created. The obtained data sets will be filled in with the missing data assignment methods in our software and the results for each data set will be compared with each other so that our software will evaluate the performance of the study in different conditions. The studied data sets are simulated to provide the following conditions:

- Multivariate normal distribution
- Data sets with 5 various variables of different sample sizes (n=1000, 1500, 2000 and 2500)
- Data sets with high (r <-0.70 or r> 0.70) or low correlations (-0.30 <r <0.30) according to the correlation between variables (8)
- Data sets with different number of missing value in variables (5%, 10% and 20% missing data)

As a result, 8 different data sets with high and low correlations are simulated by model tab (simulation) of IBM SPSS Statistics version 25.0.[10] In addition, 5%, 10%, 20% of these data sets were reduced and 24 new data sets were established in total.

MISSING VALUE ANALYSIS

Researchers have to deal with missing value(s) in order to get accurate results from the analysis that they are applying on the data. So then, researchers use some methods. Accordingly, there is a need for missing value assignment methods to get better results from research.[1,2] We use linear regression (LR), Random Forest (RF), classification and regression trees (CART) and predictive mean matching (PMM) which are known as missing value assignment methods in our software. The algorithms of the assignment methods used in the software are as follows.

Linear Regression (LR): In this technique missing values are estimated according to a regression equation established by using other variables that do not contain missing data. The regression equation (model) is established such that the variable containing the missing data is predicted and the other variables without the missing data are the predictors. Observations that are missing in the predicted variable are estimated by substitution of the values of other variables in this equation and a complete data set is achieved.[11,12]

Random Forest (RF): Instead of extracting observations with missing value(s) from the dataset, RF aims to assign the most appropriate value(s) instead of these missing value(s) thanks to the its advanced algorithm. This is done by calculating the proximity measure between observation pairs. The distance between the two observations is equal to the ending rates at the same leaf node. This rate is calculated on the trees in the forest. The missing value assignment algorithm is as follows. In this method, the missing value assignment procedure is executed as follows.

The missing value(s) in the data set is detected. If the variable including the missing value(s) is continuous, the missing data is assigned by finding the median value of the complete data of this variable. If the variable containing the missing value(s) is categorical, the assignment is made to the category with the highest frequency value from the complete data. A Random Forest model is established from the completed data set. A distance matrix is obtained through this model. The distances in this matrix are used as the weighting measure. For the missing value(s) of a continuous variable, the weighted average is calculated using the distance measures of the complete data. The resulting value is assigned to the missing data. For categorical missing value(s), the category value of the one with the highest distance from the completed data is assigned. After the new assignment processes are completed, a Random Forest model is constructed on the imputed data set again. Afterwards, a new distance matrix is achieved. With the same rules, different assignments are made to the missing values. This process, in which the missing value(s) is/are assigned using the distance matrix, is repeated the determined number to determine a consistent result. Since this process is some kind of distance based-method for closest neighborhood, it will be valid in cases where the missing data are random.[13,15]

Classification and regression trees (CART): CART can be used in data sets containing missing data. CART has provided a solution to the missing data problem with an algorithm in its structure. This algorithm is based on surrogate variables. Surrogate variables are calculated according to a specific association score. At the node where the division occurs; the left or right lower node is placed according to the surrogate variable of the estimator variable providing the discrimination. If there is a missing data on the first surrogate variable for the same observation, the second surrogate variable is used for the discrimination. If all surrogate variables contain missing data, this observation value is placed on the most crowded left and right bottom node.[16]

Predictive Mean Matching (PMM): The predictive mean matching is a method that can be used in the presence of continuous variables. The general structure of this method is similar to that of assigning missing value(s) by linear regression; but the only difference is that for each missing value, a random value is assigned from a set of values that are observed to have the predicted values closest to the predicted values.[17] The predictive mean matching method ensures that imputed values are plausible and might be more appropriate than the regression method if the normality assumption is violated.[18]

THE DEVELOPED WEB-BASED SOFTWARE

To construct the web-based application Shiny version 1.0.5, allowing to design interactive web-based apps on the basis of the R programming language, was used. The main and sub menus in the software are detailed as below.

FILE UPLOAD MENU

In the first stage of developing this web-based application, the file containing the dataset is uploaded (Figure 1). The most commonly used file types with different extensions in data analysis can be uploaded to the software. These different extensions are MS Excel (.xls/.xlsx), SPSS (.sav) and text (.csv/.txt) file types.

VARIABLE AND DISTRIBUTION TYPE DETERMINATION

On this menu of our web based application, we need to determine variable types as continuous numerical variables and discrete numerical variables (Figure 2).

Thereafter, the distribution types menu is used to determine the distribution types of variables for which variable types are specified (Figure 3). In the distribution type determination step, the rows containing

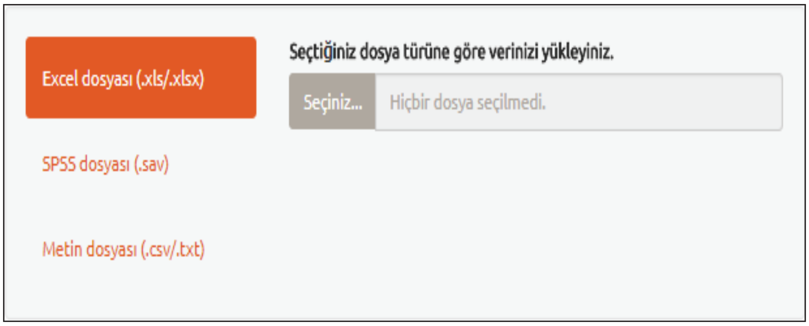

Excel dosyası (.xls/.xlsx)
SPSS dosyası (.sav)
Metin dosyası (.csv/.txt)
Seçtiğiniz dosya türüne göre verinizi yükleyiniz.
Seçiniz...
Hiçbir dosya seçilmedi.

FIGURE 1: File Upload Menu.

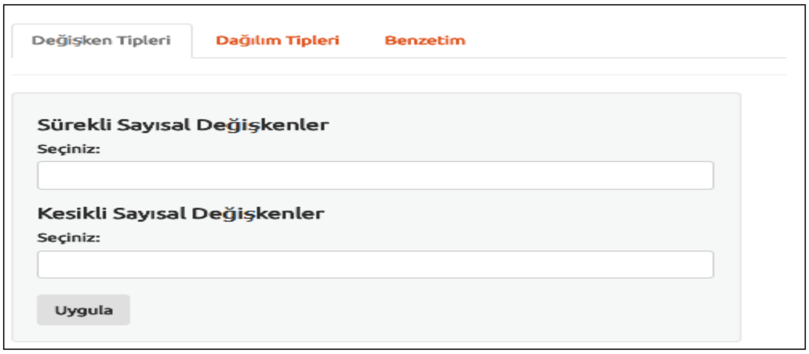


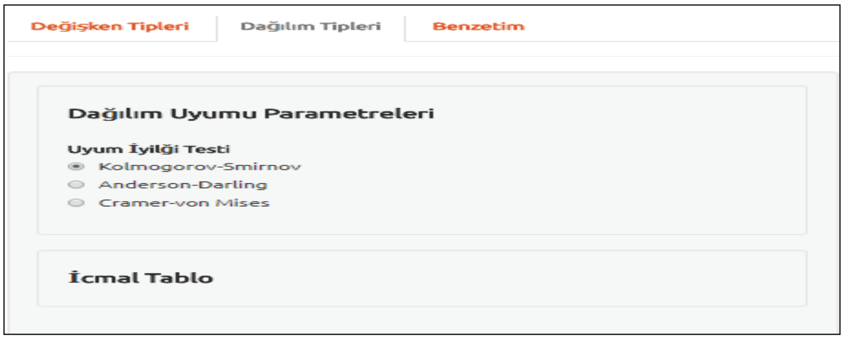
FIGURE 2: The variable types menu.

the missing values are subtracted from the data set to determine which theoretical statistical distribution type is based on the variable values of the remaining data. The Kolmogorov-Smirnov (K-S) test is used to test compliance. After the K-S test, if the distribution of the relevant variable has been determined to be more than one of several statistical distributions, the statistical distribution is better determined based on the Akaike Information Criteria.

Finally, in this section a new complete data set is created in which the distribution types are similar to the specified variables. In the output, descriptive statistics and correlation tables are given for the real data and the derived data set. In addition, the derived data set is indicated by the 'Show derived data set' button (Figure 4).

EVALUATION OF ASSIGNMENT PERFORMANCE

The derived data set obtained in this part of the software is processed. First of all, 'loss value assignment methods' section is used to select lost value assignment models used in software. Here the 'repeat number' part is used to generate in the simulated data a randomized version of the missing value of the first uploaded data. The software allows to create 1 to up to 50 different lost data. This ensures that our results are not bias when we get missing values. In addition, with 'Number of Multiple Assignments' part is used to assignments are made to the missing values. This section will help you to assign 1 to up to 20 different assignments and choose the most appropriate assignment. When choosing the most appropriate assignment from different assignments, we take advantage of the NRMSE value given in the software. The smallest NRMSE value is considered to be the appropriate assignment. Finally, we use 'comparative model performance graph' which is output in software in order to determine which of the missing value assignment models we use is most appropriate. And the model that gives the best result in assigning the missing value is selected (Figure 5).


Değişken Tipleri | Dağılım Tipleri | Benzetim

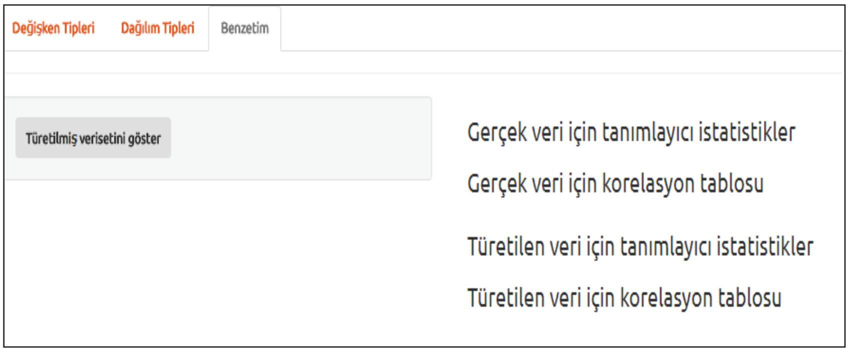
Dağılım Uyumu Parametreleri

Uyum İyilği Testi

● Kolmogorov-Smirnov
○ Anderson-Darling
○ Cramer-von Mises

İcmal Tablo

FIGURE 3: The distribution types menu


Değişken Tipleri | Dağılım Tipleri | Benzetim

Türetilmiş verisetini göster

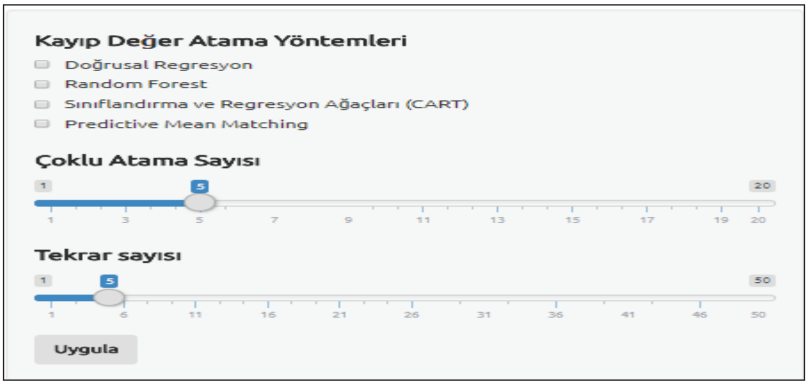
Gerçek veri için tanımlayıcı istatistikler

Gerçek veri için korelasyon tablosu

Türetilen veri için tanımlayıcı istatistikler

Türetilen veri için korelasyon tablosu

FIGURE 4: Show derived data set



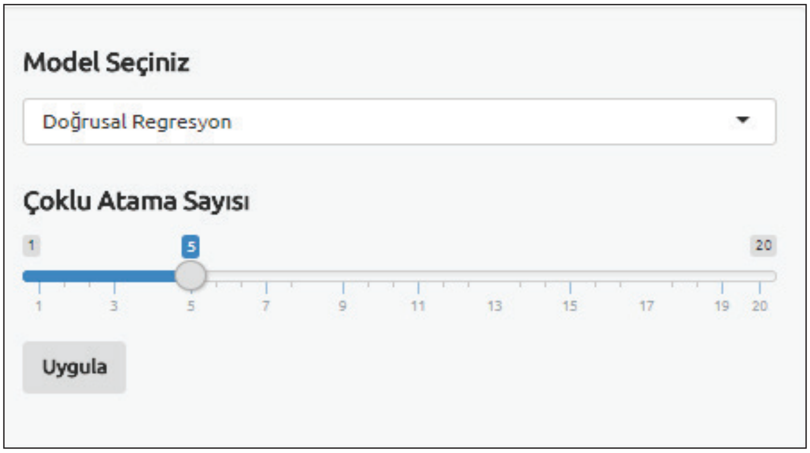
Kayıp Değer Atama Yöntemleri

□ Doğrusal Regresyon
□ Random Forest
□ Sınıflandırma ve Regresyon Ağaçları (CART)
□ Predictive Mean Matching

Çoklu Atama Sayısı: 5

Tekrar sayısı: 5

FIGURE 5: Evaluation of Assignment Performance



Model Seçiniz

Doğrusal Regresyon

Çoklu Atama Sayısı: 5

FIGURE 6: Outputs Based on Real Data

OUTPUTS BASED ON REAL DATA

The last part of the software, the ‘application’ is used to we select the method that we decided is best in the previous tab. In addition, the number of multiple assignments is selected and the missing values are assigned. The complete data set obtained is given as the output of the software. In addition, the ‘data download’ button is used to save the complete data set in MS Excel format (Figure 6).

ACCESSIBILITY OF THE DEVELOPED INTERACTIVE WEB APPLICATION

The developed interactive web application is freely accessible throughhttp://biostatapps.inonu.edu.tr/KDAY/.This web tool will be updated upon the updated R packages, shiny[19], shinyBS[20], shinythemes[21], shinydashboard[22], ShinySky[23], goftest[24], missForest[25], e1071[26], nortest[27], devtools[28], mice[29]

RESULTS

24 data sets which each variables with normal distribution, different observation numbers (1000, 1500, 2000, 2500), different correlations between variables (low and high correlation) and different missing data containment rates (5%, 10%, 20%) analyzes were done. The number of multiple assignments and repetitions in the software must be determined before the analysis starts. These values will ensure that our results are more reliable and accurate.[29] Different numbers of data sets containing the same number of missing

value as the data set originally loaded from the dataset that does not contain the missing value created by the simulated method will be generated, similar to the original dataset, with the repetition portion. These data sets are completed as different as the number of multiple assignments. In the application to be performed, the number of multiple assignments and the number of repetitions were taken as 5 and 10 respectively. The results of the comparative model performance graph generated by 4 assignment methods of one of 24 data sets in terms of brevity after the analyzes made in the developed software are given (Figure 7).

Table 1 shows which of the missing value assignment methods we use in our software for all datasets gives the best results.

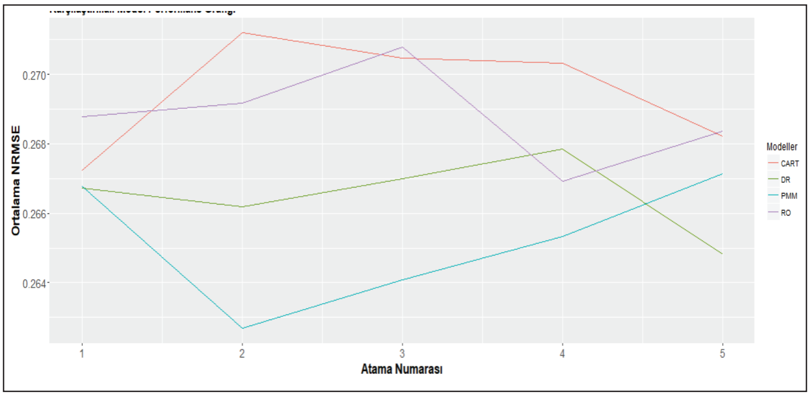


FIGURE 7: Comparative Model Performance Graph

TABLE 1: Method table showing the best assignment performance for missing data

Correlation	N	Percent of Missing Values (%)	The Method with the Best Performance
HIGH	1000	5	PMM
		10	LR
		20	LR
	1500	5	LR
		10	PMM
		20	PMM
	2000	5	PMM
		10	LR
		20	LR
	2500	5	PMM
		10	PMM
		20	LR
LOW	1.000	5	RF
		10	RF
		20	CART
	1.500	5	LR
		10	RF
		20	LR
	2.000	5	RF
		10	RF
		20	PMM
	2.500	5	RF
		10	RF
		20	RF

According to these results, DR and PMM methods in data sets with high correlation between variables showed better performance than other methods by looking at NRMSE values.

DISCUSSION

In many scientific investigations, missing / incomplete data are problematic for researchers because of the incorrect filling of case report forms, faulty measuring device, not entered or updated data, non-response from subjects. Because, if the observation(s) for the at least one variable in the data in which the statistical methods are applied is missing, it causes the value representing that observation to be empty. This means that there is a problem that will affect the analysis process.[2] Missing values in a study can lead to reduced information collected, depending on the amount and structure of missing values, affecting the structure of the data matrix and impairing the quality of the data. In this case, the reliability and validity of the measurement tools are expected to decrease. Many analysis and assignment methods for missing values are proposed to avoid bias and unintended consequences.[30] In the presence of missing data, researchers can find solutions to possible problems using one of the methods of adding new observations, extracting missing data from the dataset, making estimates of missing values with different models, and assigning approximate values to missing values. In this study, a software that includes web based user friendly and model based missing value assignment methods was developed with Shiny, an open source R package, to enable researchers to deal with problems related to missing value. Developed Software differently from other software that makes missing value assignments (IBM SPSS Statistics[10], Minitab[31], etc.) has been proposed to remove the bias effects by making better assignments to missing values by using “Number of repetitions” and “Number of Multiple Assignments” options.[29] Within the scope of this study, applications were made on data set containing 24 lost data with different conditions derived to introduce the developed web based software to users. For the data sets with high correlation according to the findings, DR and PMM methods of the model based assignment methods in the software gave better results. As is known, the PMM method has a structure similar to the method of assigning missing value with DR in general.[18] For this reason, both are expected to perform better in highly correlated data sets. It has been found that the four methods proposed in the data sets with low correlation, different sample sizes, and different missing data containment rates are not superior to each other.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Ahmet Kadir ARSLAN, Zeynep TUNÇ, Emek GÜLDOĞAN, Cemil ÇOLAK*; ***Design:*** *Ahmet Kadir ARSLAN, Zeynep TUNÇ, Emek GÜLDOĞAN, Cemil ÇOLAK*; ***Control/Supervision:*** *Ahmet Kadir ARSLAN, Zeynep TUNÇ, Emek GÜLDOĞAN, Cemil ÇOLAK*; ***Data Collection And/Or Processing:*** *Ahmet Kadir ARSLAN, Zeynep TUNÇ*; ***Analysis And/Or Interpretation:*** *Ahmet Kadir ARSLAN, Zeynep TUNÇ, Emek GÜLDOĞAN, Cemil ÇOLAK;* ***Literature Review:*** *Ahmet Kadir ARSLAN, Zeynep TUNÇ*

REFERENCES

1. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92. [Crossref]
2. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. New Jersey: Wiley-Interscience; 2002. p.408. [Crossref]
3. Çüm S, Demir EK, Gelbal S, Kışla T. [A comparison of advanced methods used for missing data imputation under different conditions]. Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi. 2018;(45):230-49.
4. Pigott TD. A review of methods for missing data. Educational Resarch and Evaluation. 2001;7(1):353-83. [Crossref]
5. Allison PD. Missing data techniques for structural equation modeling. J Abnorm Psychol. 2003;4(1):545-57. [Crossref]
6. Osborne JW. Best Practices in Data Cleaning: A Complete Guide to Everything You Need to do Before and After Collecting Your Data. 1st ed. California: Sage Publication Inc; 2013. p.275. [Crossref]
7. Çüm S, Gelbal S. [The effects of different methods used for value imputation instead of missing values on model data fit statistics]. Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi. 2015;1(35):87-111.
8. Roth PL. Missing data: a conceptual review for applied psychologists Personnel Psychology. 1994;47(3):537-60. [Crossref]
9. Alpar R. Çok Değişkenli İstatistiksel Yöntemler. 5. Baskı. Ankara: Detay Yayıncılık; 2017. p.840.
10. IBM Corp. IBM SPSS Statistics for Windows. Version 25.0. Armonk, Ny: IBM Corp; 2017.
11. Enders CK. Applied Missing Data Analysis. 1st ed. New York: The Guilford Publications; 2010. p.377.
12. Akman M, Genç Y, Ankarali H. [Random forests methods and an application in health science]. Turkiye Klinikleri J Biostat. 2011;3(1):36-48.
13. Cutler A, Cutler DR, Stevens JR. Ensemble Machine Learning. 1st ed. New York: Springer; 2012. p.329.
14. Cutler A, Cutler DR, Stevens JR. Tree-based methods. High-Dimensional Data Analysis in Cancer Research. 1st ed. New York: Springer; 2009. p.1-19. [Crossref] [PMC]
15. Bertsimas D, Pawlowski C, Zhuo YD. From predictive methods to missing data imputation: an optimization approach. The Journal of Machine Learning Research. 2017;18(1):7133-71.
16. Zhang S, Qin Z, Ling CX, Sheng S. “Missing is useful”: missing values in cost-sensitive decision trees. IEEE Trans Knowl Data Eng. 2005;17(12):1689-93. [Crossref]
17. Schenker N, Taylor JM. Partially parametric techniques for multiple imputation. Computational Statistics & Data Analysis. 1996;22(4):425-46. [Crossref]
18. Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. The American Statistician. 2003;57(4):229-32. [Crossref]
19. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. Shiny: web application framework for R. R Package version 0.13. 2016;2.
20. Bailey E. shinyBS: Twitter bootstrap components for shiny. R package version 0.61; 2015.
21. Chang W. Shinythemes: themes for shiny. R package version 1.0. 2015;1.
22. Chang W. Ribeiro BB. Shinydashboard: create dashboards with Shiny. R package version 0.6. 2017;1.
23. AnalytixWare, shinysky: A set of Shiny UI components/widest R package version 0.1.2.
24. Faraway J, Marsaglia G, Marsaglia J, Baddeley A. Goftest: Classical goodness-of-fit tests for univariate distributions; 2014.
25. Stekhoven DJ. missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1;2012.
26. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. The e1071 package; 2005.
27. Gross J, Ligges U. Nortest: Tests for Normality, R package version 1.0-2; 2012.
28. Wickham H, Chang W. Devtools: tools to make developing R packages easier. R package version 1.12. 0. 2016; 2017.
29. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. Ann Transl Med. 2016;4(2):30. [PMC]
30. Demir E, Parlak B. [Missing value problem in educational research in Turkey]. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi. 2012;3(1):230-41.
31. Minitab, I. N. C. MINITAB statistical software, Minitab Release, 13;2000.