Journal of Mahani Mathematical Research Center
Print ISSN: 2251-7952 Online ISSN: 2645-4505

ON HIERARCHICAL MULTIPLE IMPUTATION METHOD FOR HANDLING MISSING DATA

A. Sheikhi*, A. Arabpour, M. Khosravi, M. Mashinchi, R. Pourmousa, M. Rezapour, M.J. Rostami, A. Abbdollah Nejad, and A. Badakhshan

Dedicated to sincere professor Mehdi Radjabalipour on turning 75

Article type: Research Article

(Received: 16 June 2021, Revised: 01 August 2021, Accepted: 31 August 2021)

(Communicated by N. Shajareh Poursalavati)

Abstract. TIn this work we carry out a multiple imputation technique for handling missing observations. We propose an algorithm, which performs a hierarchical multiple imputation using edition rules to impute missing values. We assess our algorithm using a simulation study and a numerical application of our algorithm in dataset of Kerman Chamber of Commerce, Industries, Mines and Agriculture is presented for more illustration.

Keywords: Missing Data, Multiple Imputation, Editing Rules, Data Cleaning.
2020 MSC: Primary 62Dxx, 62D10.

1. Introduction

One of the first and foremost steps in data mining and knowledge discovery is data preparation and one of the data preparation steps is data cleaning, which provided a variety of tools and approaches to achieve a best data preparation results [10, 13]. Data cleansing is a process used to address noisy, inaccurate and incomplete information, and suggests some methods include correcting unusable data, duplicates, and omissions to improve quality of data preparation [14].

The data cleaning process includes evaluating the data, checking the format, identify faults or other errors, and uses some standard tasks and rules before validation to ensures compliance. In this process, two most important approaches among them is outlier detecting and dealing with missing values. There are four techniques in the literature, which are frequently used to handling outliers, namely, Numeric Outlier [2], Z-Score [34], Density-Based Spatial Clustering of Application with Noise algorithm (DBSCAN) [3] and Isolation Forest [9] methods. We refer to Rousseeuw and Hubert (2011) [27] Aggarwal and Sathe (2017) [1]), Gravesteijn et al. (2021) [15] and Grund et al. (2021) [16] for more details.

*Corresponding author, ORCID: 0000-0002-3731-6012
E-mail: sheikhy.a@uk.ac.ir
DOI: 10.22103/jmmrc.2021.17749.1153

How to cite: A. Sheikhi et al., *Hierarchical Multiple Imputation Method for Handling Missing Data*, J. Mahani Math. Res. Cent. 2021; 10(2): 103-114.

Concerning with missing values, comprehensive works have been done by researchers. Josse and Hussen [17] have handled missing values in exploratory multivariate data analysis. Chen et al. (2014) [8] have developed multiple regression analysis in the presence of missing data. Pratama et al. (2017) [25] have presented a review of missing values detecting methods on time series data. Kwak et al. (2017) [21] have discusses the effect of the presence of missing values in the dataset on the reliability of statistical analysis. Recently, Ramosaj and Pauly (2019) [26] have predicted missing values based on some non-parametric approaches for imputation.

Although, sometimes, data deletion which is omitting entire records for variables is done especially when there is a substantial number of missing data, but there are some useful techniques to avoid variable deleting. Multiple imputation (MI) (Rubin, 1996, 2004 [28, 29]) has been shown that makes appropriate results in the presence of a high number of missing data or a small sample size of observations [23]. MI as originally conceived proceeds in two stages: A data disseminator creates a small number of completed datasets by filling in the missing values with samples from an imputation model. Analysts compute their estimates in each completed dataset and combine them using simple rules to get pooled estimates and standard errors that incorporate the additional variability due to the missing data. [24] Little and Rubin (2109) [22] have pointed out that an apparent advantage of this approach is its ability to make standard complete-data methods applicable to incomplete data. It also turns out to be robust when the data violate the normality assumptions [18]. The most important application if MI can be found in multivariate data, see e.g, Schafer (1997) [31], Schafer and Olsen (1998) [32], Shobha and Nickolas (2019) [35] and Buuren and Groothuis (2010) [6]. The latter introduces a most popular imputation method which is called Multivariate Imputation by Chained Equations (MICE).

This work concerns with the implementation of statistical techniques to multiple imputing the missing values in a hierarchical fashion. So, the organization of this work is as follows. Section 2 is devoted to some preliminaries about editing rules and introduces our hierarchical multiple imputation algorithm. A simulation of study as well as a data cleaning in a real data set is presented in section 3. Also, we apply our algorithm to clean the dataset of Kerman Chamber of Commerce, Industries, Mines and Agriculture. Finally, section 4 presents some concluding remarks.

2. **Methodology**

Generally, missingness in a dataset can be classified into three groups based on their occurrence: missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR). MAR, means there is a systematic relationship between the propensity of missing values and the observed data, but not the missing data, while in MCAR, the probability of an observation being missing does not depend on observed or unobserved measurements [4]. In contrast, MNAR, means there is a relationship between the propensity of a value to be missing and its values.

It should be pointed out that, MAR and MCAR are both considered ignorable because researchers do not have any information about the missing data when they deal with the missing data. Besides, MNAR occurs when the missingness depends on the value of the variable and is called non-ignorable because the missing data mechanism itself has to be modeled as deals with the missing data. See for more information Seaman et al. (2013) [33] and Bhaskaran and Smeeth (2014) [12]. In this work we are encountered with missing (completely) at random and we can use multiple imputation to dealing with their missing values [7, 12]. A recent book of Little and Rubin (2019) [22] and references therein give a comprehensive information about this topic.

Edit rules or editing rules are constraints on the variables of dataset that assure the validity of the variables in each record. Each of records must satisfy all edit rules in order to qualify as an admissible record. Some simple types of edit rules may be seen as: range restrictions ($V_1 \leq a$), ratio constraints ($V_1 \leq bV_2$), and balance constraints ($V_1 + V_2 = V_3$). When a record fails a set of edits, agencies typically select some fields to replace with imputed values so that all constraints are satisfied (Fellegi and Holt 1976 [11] and Kim et al. 2015 [20]).

Consider we have an $n \times k$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1\boldsymbol{x}_2...\boldsymbol{x}_k)$, including k variables with n observations. Some of $\boldsymbol{x}_i$, $i = 1, 2, .., k$, contain missing values. Also, we donote V_1, $V_2, ..., V_k$ are respectively represents variable's rule of $\boldsymbol{x}_1$, $\boldsymbol{x}_2, ..., \boldsymbol{x}_k$. Following Kim et al. 2015 [20]), we consider set of editing rules:

$$\mathbb{S} = \begin{cases} V_i \leq a & \forall i = 1, 2, ...k \\ V_i \leq bV_j & \forall i, j = 1, 2, ..., k \\ \sum_{i=1}^{k} a_i V_i = 0 \end{cases}, \tag{1}$$

where a, b and a_i, $i = 1, 2, ...k$, are constants, which always determined by experts.

In this work we do multiple imputations using hierarchical edit rules to replace missing values with the most valid values. A pseudo algorithm of this multiple imputations hierarchical edit rules is presented in Algorithm **??**.

It is worth noting that, this algorithm is a extension of other traditional multiple Imputation algorithms, see for an instance Khan et al. (2020) [19], because this algorithm updates the editing rules after imputing the missing values of each variables.

3. Numerical analysis

3.1. **Simulated data.** In this subsection, we carry out some numerical analyses to assess our algorithm in imputation of missing values in two scenarios:

Algorithm 1: Multiple Imputation using Hierarchical Editing Rules

Data: Data matrix $\boldsymbol{X}$, set of editing rule $\mathbb{S}$

Result: Cleaned data matrix $\boldsymbol{X}^*$

Initialization : Label editing rules $V_1,\ V_2, ..., V_k$ for variables $\boldsymbol{x}_1,\ \boldsymbol{x}_2, ..., \boldsymbol{x}_k$

for $i = 1, 2, ..., k$ **do**

  if $\boldsymbol{x}_i$ *has missing values* **then**

    Impute missed values using editing rules $\mathbb{S}$ and construct imputed $\boldsymbol{x}_i^*$;

  end

  else

    $\boldsymbol{x}_i^* = \boldsymbol{x}_i$

  end

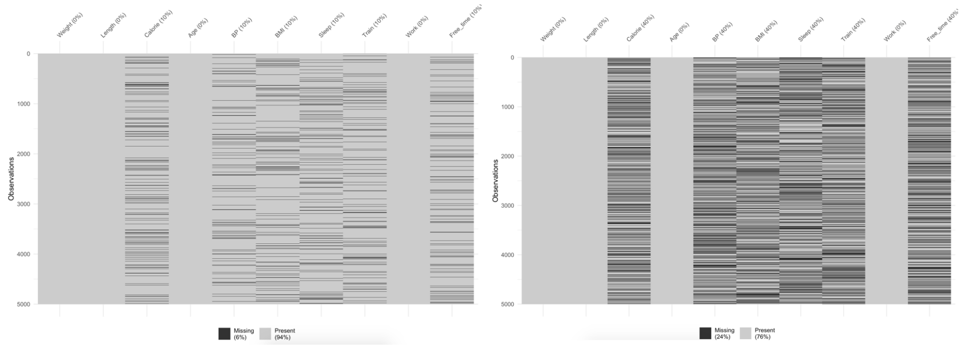
  Update if needs $\mathbb{S}$, based on expert's advise ;

end

few number and many number of missing values. As we mentioned before, since the MICE approach is the most popular technique in multiple invitations, we compared our results with the results of MICE. In this regard, inspiring Sasaki et al. (2020) [30], we simulated 5000 observations from obesity associated variables such as "age, weight, length, BMI, Blood Pressure(BP), Calorie, Train, Work, Sleep and Free_time" [1]. In order to use their correlations and regression we generated observations of "age, weight, length, Blood Pressure(BP), Calorie, Train, Work, Sleep" from multivariate a normal variables with reasonable correlation values between variables. Also, in order to use the edition rules of section 3, we obtained values of BMI variable based on relation $BMI = \frac{Weight}{Length^2}$ and the so called Free_time of this observations using $Free\_time = 24 - (Train + Work + Sleep)$. In the first scenario, we set 10 percent of observations of variables "Calorie, Blood Pressure(BP), BMI, Train, Sleep and Free_time" to missing values (NA) and for the second scenario we replaced 30% of the observations of these variables with missed values.

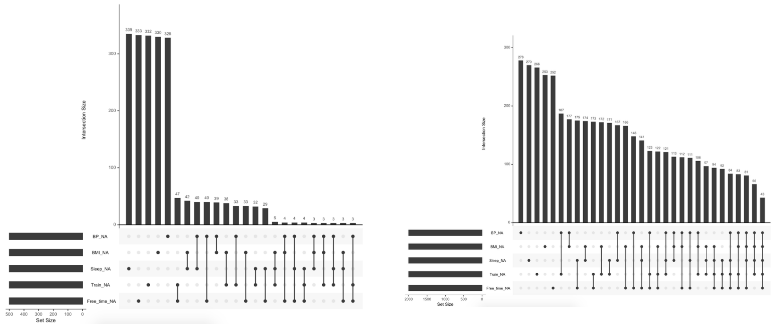
Figures 1-a and 1-b are describing the map of missing value in two cases of few/many number of the missing values. See also Figures 2-a and 2-b for frequencies and intersections of missing values in these two cases. Using Algorithm 1 we hierarchically imputed these missed observations. First, based on the second equation of 1, the relation $BMI = \frac{Weight}{Length^2}$ yielded imputed (and exact) values of the missed observations of the BMI variable. Similarly, applying the the third equation of 1 we used, $Free\_time = 24 - (Train + Work + Sleep)$ for importing missed values of the Free_time variable. Also, imputing the missed observations of the other remaining variable were done hierarchically using procedure of algorithm 1, especially using their correlation [6]. The minimum difference between original values and the imputed values was desired. So, the absolute R.Bias and root mean squared errors (RMSE) for the imputed values were our criterion to compare our approach with the traditional MICE. These

[1] We use the R software and all codes are available upon request.



a b

FIGURE 1. Map of missing values: a) many number of missingness b) few number of missingness



a b

FIGURE 2. Frequencies and intersections of missing values: a) many number of missingness b) few number of missingness

two indices were calculated as follows:

$$R.Bias = \frac{1}{5000}\sum_{i=1}^{5000} \mid \frac{\hat{y}_i - y_i}{y_i} \mid \quad RMSE = \sqrt{\frac{1}{5000}\sum_{i=1}^{5000}(\hat{y}_i - y_i)^2}$$

where $\hat{y}_i$ is the imputed value of y_i in the ith case. The results are summarized in Table 1, for two scenarios few/many number of missing values. From this table we can deduce in contrast of MICE approach, our hierarchical method can improve the accuracy of imputation in both cases of law and high number of missing observations.

3.2. **Real data.** In this subsection we use our proposed algorithm to impute missed observation in data of pistachio in the data set of Kerman chamber

Frequency of missing values	Criterion	MICE	Hierarchical MI
A few	R.Bias	78.24	65.46
	RMSE	45073.67	41911.67
Many	R.Bias	231.20	193.43
	RMSE	135765.91	125547.05

TABLE 1. R.Bias and RMSE of two methods few/many numbers os missing values

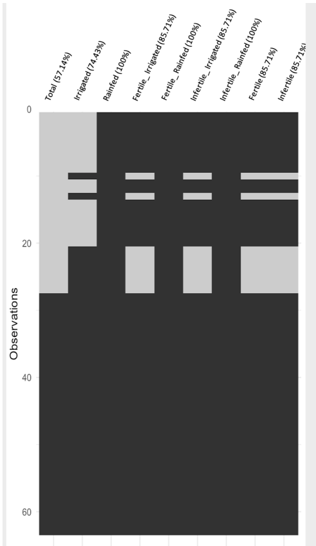


FIGURE 3. Percentage of missing observation separated by cultivated type and fertility

of commerce, industries ,mines and agriculture. This data set, which is collected by Organization of agriculture-Jihad-Kerman, contains some information about pistachio product in three providences: North Kerman, Orzueiyeh and Bam between years 2011 t0 2017[2]. Also, this dataset contains some attributes like: cultivated area, cultivated type (rain-fed or irrigated), fertility (fertile or infertile), year, yield rate per hectare, harvest by ton and etc.

[2]This data set are available at http://www.data.kccstat.com which is the Statistics and Information Bank of Kerman Chamber of Commerce, Industries, Mines and Agriculture. This data set has been set up since 2016 with the aim of quickly accessing the economic activities of the province to the data required for planning, as well as integration and transparency. The coding system of this statistical bank is designed for statistical items based on the statistical coding of the United Nations Statistical Bank, which is a combination of ISIC and CPC codes.

There exist a lot of missing values in this data set in which traditional imputation techniques such as neighborhood data or regression type imputation methods cannot help us to deal with these missing values. In some cases more than 80% of observations are missed. Figure 3 shows percentage of missing of variables separated by cultivated type and fertility. More information about missingness of these data based on other attributes are depicted in Figure 4.

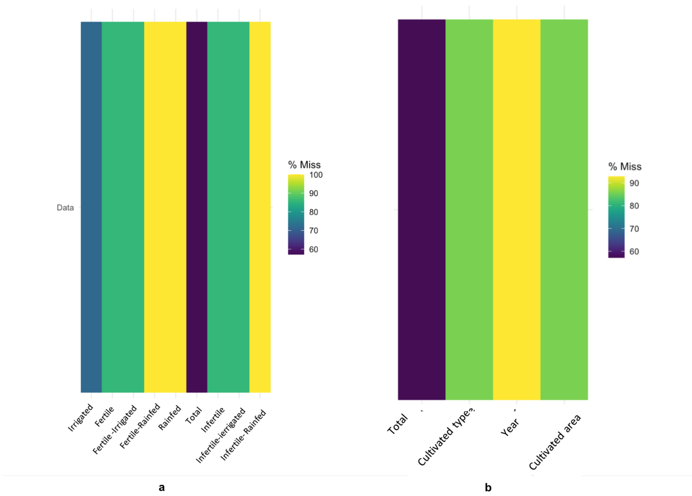


FIGURE 4. Percentage of missingness separated by irrigated type and fertility and their interactions

Since occurrence of these missing values does not depends on other observations, we use our algorithm to impute them. First, based on the experts advise, we note that some of missed values actually had zero values, so we can replace these type of missed observations to zero values. For example, in this data set, experts tell us that cultivate of pistachio in Bam province has begun from 2017; so all of missing values before 2017 in that province will change form “NA” to “0”. This trick reduces the proportion of missingness from 87% to 50%, see Figure 5-a, noting that we have filtered the rain-fed data set in this figure only for a better visualization. Then, in order to impute other missed observations, we have to perform the steps of algorithm 1 for all attributes respectively. We have used editing rules that are determined by experts of this field.

Considering the following attributes,

V=Total
V_1=Fertile
V_{11}=Irrigated Fertile
V_{12}=Rain-fed Fertile
V_2=Infertile
V_{21}=Irrigated Infertile
V_{22}=Rain-fed Infertile

V_3=Irrigated
V_4=Rain-fed
V_5=Cultivated area
V_6=Yield rate per hectare
V_7=Harvest by ton
V_8=Year ,

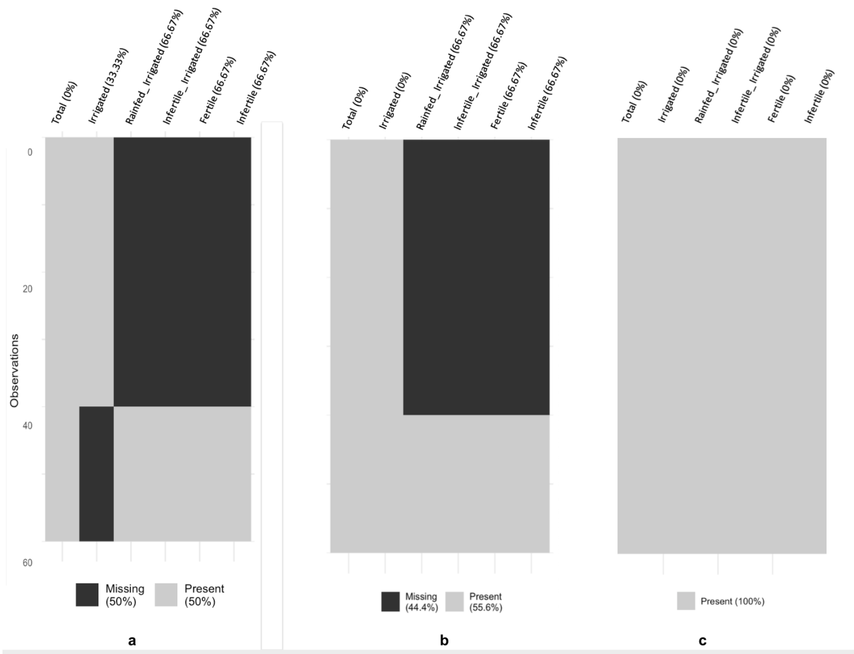


FIGURE 5. Percentage of missingness via the hierarchical imputation algotithm

Experts tell us that the following editing rules are hold between these attributes:

$$(2)\qquad \mathbb{S} = \begin{cases} V \geq 0 \ \text{ and } \ V_i \geq 0 & \forall i = 1, 2, ...5 \\ V = V_1 + V_2 = V_3 + V_4 & \\ V_i = V_{i1} + V_{i2} & \forall i = 1,\ 2 \\ V_3 = V_{11} + V_{21} \text{ and } \ V_4 = V_{21} + V_{22} & \\ V_5 = 2.2031 \times V_7 \ \text{ and } \ V_6 \times V_5 = V_7 \times 1000 & \end{cases} .$$

We perform this rules via Algorithm 1 in which impute all missing values. For example, from the forth equation in (2) we have Irrigated=Fertile_Irrigated+ Infertile_Irrigated ($V_3 = V_{11} + V_{21}$), so we can use this rule to impute missed cases of the Irrigated attribute (see Figure 5-b). Continuing this process and using the last relation of (2) yields 0% of missingness (Figure 5-c).

4. **Conclusion**

The algorithm 1 which uses a hierarchical multiple imputation is employed to deal with some completely at random missing values. A set of editing rules, which are advised by experts have helped us in this imputation.

The idea of this paper could be extended in some manner. In this work we consider only crisp constraints, however some fuzzy constraints may be of interest. Also, increasing the number of the constraints may increases the accuracy of the work, in which, some of these constraints can be used to validate others.

Recently, neural networks attracted much attentions and their improvements appear in many topics such as deep neural network, convolutional neural networks, recurrent neural networks and so on. Our approach can be used in these machine learning subjects especially when we are dealing with big data. In this regards, the Hot Deck multiple imputation would be of interest, as can be seen in the recent work by Butera (2019) [5].

Acknowledgment

The work on this paper was supported by the grants provided by Kerman Chamber of Commerce, Industries, Mines and Agriculture, Kerman, Iran, Aug. 15, 2020.

References

[1] Charu C Aggarwal and Saket Sathe. *Outlier ensembles: An introduction.* Springer, 2017.

[2] Malik Agyemang, Ken Barker, and Rada Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis*, 10(6):521–538, 2006.

[3] Zohreh Akbari and Rainer Unland. Automated determination of the input parameter of dbscan based on outlier detection. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 280–291. Springer, 2016.

[4] Krishnan Bhaskaran and Liam Smeeth. What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology*, 43(4):1336–1339, 2014.

[5] Nicole M Butera, Siying Li, Kelly R Evenson, Chongzhi Di, David M Buchner, Michael J LaMonte, Andrea Z LaCroix, and Amy Herring. Hot deck multiple imputation for handling missing accelerometer data. *Statistics in Biosciences*, 11(2):422–448, 2019.

[6] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[7] James R Carpenter, Michael G Kenward, and Ian R White. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical methods in medical research*, 16(3):259–275, 2007.

[8] Ya Chen, Yongjun Li, Huaqing Wu, and Liang Liang. Data envelopment analysis with missing data: A multiple linear regression analysis approach. *International Journal of Information Technology & Decision Making*, 13(01):137–153, 2014.

[9] Zhangyu Cheng, Chengming Zou, and Jianwei Dong. Outlier detection using isolation forest and local outlier factor. In *Proceedings of the conference on research in adaptive and convergent systems*, pages 161–168, 2019.

[10] Tamraparni Dasu and Theodore Johnson. *Exploratory data mining and data cleaning.* John Wiley & Sons, 2003.

[11] Ivan P Fellegi and David Holt. A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71(353):17–35, 1976.

[12] Gary Fraser and Ru Yan. Guided multiple imputation of missing data: using a subsample to strengthen the missing-at-random assumption. *Epidemiology*, pages 246–252, 2007.

[13] Alex A Freitas. *Data mining and knowledge discovery with evolutionary algorithms*. Springer Science & Business Media, 2013.

[14] Salvador García, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining*. Springer, 2015.

[15] Benjamin Yaël Gravesteijn, Charlie Aletta Sewalt, Esmee Venema, Daan Nieboer, Ewout W Steyerberg, and CENTER-TBI Collaborators. Missing data in prediction research: A five-step approach for multiple imputation, illustrated in the center-tbi study. *Journal of neurotrauma*, 38(13):1842–1857, 2021.

[16] Simon Grund, Oliver Lüdtke, and Alexander Robitzsch. Multiple imputation of missing data in multilevel models with the r package mdmb: a flexible sequential modeling approach. *Behavior Research Methods*, pages 1–19, 2021.

[17] Julie Josse and François Husson. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153(2):79–99, 2012.

[18] Hyun Kang. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402, 2013.

[19] Shahidul Islam Khan and Abu Sayed Md Latiful Hoque. Sice: an improved missing data imputation technique. *Journal of Big Data*, 7(1):1–21, 2020.

[20] Hang J Kim, Alan F Karr, and Jerome P Reiter. Statistical disclosure limitation in the presence of edit rules. *Journal of Official Statistics*, 31(1):121–138, 2015.

[21] Sang Kyu Kwak and Jong Hae Kim. Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, 70(4):407, 2017.

[22] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[23] Daniel McNeish. Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics*, 44(1):24–39, 2017.

[24] Jared S Murray et al. Multiple imputation: a review of practical and theoretical findings. *Statistical Science*, 33(2):142–159, 2018.

[25] Irfan Pratama, Adhistya Erna Permanasari, Igi Ardiyanto, and Rini Indrayani. A review of missing values handling methods on time-series data. In *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 1–6. IEEE, 2016.

[26] Burim Ramosaj and Markus Pauly. Predicting missing values: a comparative study on non-parametric approaches for imputation. *Computational Statistics*, 34(4):1741–1764, 2019.

[27] Peter J Rousseeuw and Mia Hubert. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79, 2011.

[28] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

[29] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[30] Akiyo Sasaki-Otomaru, Kotaro Yamasue, Osamu Tochikubo, Kyoko Saito, and Masahiko Inamori. Association of home blood pressure with sleep and physical and mental activity, assessed via a wristwatch-type pulsimeter with accelerometer in adults. *Clinical and Experimental Hypertension*, 42(2):131–138, 2020.

[31] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[32] Joseph L Schafer and Maren K Olsen. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4):545–571, 1998.

[33] Shaun Seaman, John Galati, Dan Jackson, and John Carlin. What is meant by "missing at random"? *Statistical Science*, 1:257–268, 2013.

[34] Ronald E Shiffler. Maximum z scores and outliers. *The American Statistician*, 42(1):79–80, 1988.

[35] K Shobha and S Nickolas. Imputation of multivariate attribute values in big data. In *Smart intelligent computing and applications*, pages 53–60. Springer, 2019.

Ayyub Sheikhi
Orcid number: 0000-0002-3731-6012
Department of Statistics
Faculty of Mathematics and Computer
Shahid Bahonar University of Kerman
Kerman, Iran
E-mail address: sheikhy.a@uk.ac.ir

Alireza Arabpour
Orcid number: 0000-0001-5287-2263
Department of Statistics
Faculty of Mathematics and Computer
Shahid Bahonar University of Kerman
Kerman, Iran
E-mail address: arabpour@uk.ac.ir

Mohsen Khosravi
Orcid number: 0000-0001-6862-7983
Department of Statistics
Faculty of Mathematics and Computer
Shahid Bahonar University of Kerman
Kerman, Iran
E-mail address: khosravimm@yahoo.com

Mashallah Mashinchi
Orcid number: 0000-0001-5848-8027
Department of Statistics
Faculty of Mathematics and Computer
Shahid Bahonar University of Kerman
Kerman, Iran
E-mail address: mashinchi@uk.ac.ir

Reza Pourmousa
Orcid number: 0000-0001-6412-7260
Department of Statistics
Faculty of Mathematics and Computer
Shahid Bahonar University of Kerman
Kerman, Iran
E-mail address: pourm@uk.ac.ir

Mohsen Rezapour
Orcid number: 0000-0002-4008-0100
Department of Statistics
Faculty of Mathematics and Computer
Shahid Bahonar University of Kerman
Kerman, Iran
E-mail address: mohsenrzp@uk.ac.ir

Mohammad Javad Rostami
Orcid number: 0000-0003-1226-345X
Department of Computer Engineering
Shahid Bahonar University of Kerman
Kerman, Iran
and
Kerman Chamber of Commerce
Industries, Mines and Agriculture
Kerman, Iran
E-mail address: rostami@uk.ac.ir

Amin Abbdollah Nejad
Orcid number: 0000-0002-7062-2484
Kerman Chamber of Commerce
Industries, Mines and Agriculture
Kerman, Iran
E-mail address: aminabdolahnejad@yahoo.com

Abid Badakhshan
Department of Statistics
Faculty of Mathematics and Computer
Shahid Bahonar University of Kerman
Kerman, Iran
E-mail address: abidbadakhshan@gmail.com