Journal of AL-Qadisiyah for computer science and mathematics Vol.11 No.2 Year 2019
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

CHOOSING APPROPRIATE IMPUTATION METHODS FOR MISSING DATA: A DECISION ALGORITHM ON METHODS FOR MISSING DATA

[1] Wisam A. Mahmood, [2] Mohammed S. Rashid, [3] Teaba wala Aldeen

[1] University of Technology-Iraq / Department of Computer Sciences
Email: 10860@uotechnology.edu.iq
[2] University of Technology-Iraq / ITC Information technology Center
Email: 11643@uotechnology.edu.iq
[3] University of Technology-Iraq / Department of Computer Sciences
Email: 110053@uotechnology.edu.iq

Recived : 11\3\2019 Revised : 20\3\2019 Accepted : 17\4\2019

Available online : 6 /5/2019

Abstract:

Missing values commonly happen in the realm of medical research, which is regarded creating a lot of bias in case it is neglected with poor handling. However, while dealing with such challenges, some standard statistical methods have been already developed and available, yet no credible method is available so far to infer credible estimates. The existing data size gets lowered, apart from a decrease in efficiency happens when missing values is found in a dataset. A number of imputation methods have addressed such challenges in early scholarly works for handling missing values. Some of the regular methods include complete case method, mean imputation method, Last Observation Carried Forward (LOCF) method, Expectation-Maximization (EM) algorithm, and Markov Chain Monte Carlo (MCMC), Mean Imputation (Mean), Hot Deck (HOT), Regression Imputation (Regress), K-nearest neighbor (KNN),K-Mean Clustering, Fuzzy K-Mean Clustering, Support Vector Machine, and Multiple Imputation (MI) method. In the present paper, a simulation study is attempted for carrying out an investigative exploration into the efficacy of the above mentioned archetypal imputation methods along with longitudinal data setting under missing completely at random (MCAR). We took out missingness from three cases in a block having low missingness of 5% as well as higher levels at 30% and 50%. With this simulation study, we concluded LOCF method having more bias than the other methods in most of the situations after carrying out a comparison through simulation study.

Keywords:
Missing data, imputation methods, multiple imputation

1-Introduction

When a large database is analyzed by many users, there is a desire to “clean up” the data, which includes dealing with missing values [1]. The reason is that standard procedures cannot be used when there are missing values and corresponding procedures that adjust for missing values may not be easy to derive. Imputation (estimating the missing values) is one of the most common procedures for handling missing values [2][6]. Single imputation is just as the name suggests, filling in a single value for each missing value. Single imputation is attractive for several reasons. The analysts find it cumbersome to effectively conduct data analysis due to the presence of missing observations. Types of problems that are usually associated with missing values are 1) loss of efficiency; 2) complications in handling and analyzing the data; 3) bias resulting from differences arising involving missing and complete data (bias estimates). The other problems include lower capacity in statistics (inefficient estimates). The missing data pattern as well as the missing data means largely determine the process in which the right methods are chosen to handle missing observations in relation to time series. However, if the observations were more than 60 percent missing, no method was found suitable to cure the data embracing the imputation techniques are commonly used for the treatment of missing data. But, such a technique has encountered some challenges like maximizing the application of existing data for minimizing the mean square error in univariate statistics, besides preserving covariance structures in multivariate data sets[4][5]. The other challenges are related to the application of imputed data in estimating variance of the uncertainty as done in the case of synthetic (unobserved) data. This paper primarily focuses on reviewing the methods of imputation i.e. single and multiple imputations and their limitation. In sharp contrast from earlier works, this paper mostly focuses on effectively applying specific imputation methods for finding solutions to the problems of missingness with a thorough review of the application of single and multiple imputations in many fields, thereby enabling modifications for enhanced prediction.

1.1Missing Data Mechanisms

There are two important types of missing data describe by known as ignorable and nonignorable. The probability of missing items depends on the values of observations in Non-ignorable types. On the other hand, there is no dependence of probability of missing items on the value of observations in case of ignorable missing data. There are three types of missing data mechanisms that integrated with ignorable missing data [11][9]. In a data matrix, the correlation between missingness and the values of variables is represented by application of the missing data mechanism. As given an observed variable $\mathbf{Y}$ as $\mathbf{Y}_{\mathbf{obs}}$ and a missing variable $\mathbf{Y}$ as $\mathbf{Y}_{\mathbf{mis}}$, it can be said that $\mathbf{Y=[Y_{obs}, Y_{mis}]}$. The missing completely at random (MCAR) is quite effective in case of the missingness happens randomly all over the entire data sets. Thus, the probability of missing value is independent of both $\mathbf{Y}_{\mathbf{obs}}$ and $\mathbf{Y}_{\mathbf{mis}}$. The second form is missing at random (MAR)[7][10]. Such a perplexing missing data form emerges when the probability of a record with a missing value for an attribute is not dependent on the value of the missing data itself, but might have dependence on the observed data. This effectively means that the probability of missing value has no dependence upon $\mathbf{Y}_{\mathbf{mis}}$. In general parlance, in the event of the entire missing data being MAR, missing data can be handled by applying simple techniques like carrying out analysis on complete and available cases. The indicator method and overall mean imputation will only provide biased results. Nonetheless, adoption of other sophisticated techniques like single and multiple imputations give unbiased results for MAR form of missing data reported that MAR and MCAR are able to be ignorable because it is possible to adjust for the missingness[12][14]. The next experiment is linked to sampling as it is impossible to harness data from all the constituents in a population. Here, probability sampling has widespread application in obtaining data from the marked population. This form is not considered further. When the probability of a missing datum depends on its value, Non-ignorable sets in, besides occurring when the pattern of missingness does not guarantee any reliable prediction of the missing value of *Y* from other dataset variables. One form of non-ignorable is missing not at random (MNAR). When the probability of a record with a missing value for an attribute is dependent upon the value of the attribute, then the former condition sets in. Critical information vanishes when missing data resembles MNAR, and is particularly worrisome a scenario in the absence of any proven method to effectively cure the missing data at one go.

Fig.1. Classification of missing data

Moreover, in the past relatively no study has been conducted to measure the influence of missing values on the outcome of classification. In this paper, a multi-stage process is suggested for handling missing values, which consists of three stages.

1.2 Discarding method

The two primary principles used to discard data with missing values are Complete Case Analysis and Available Case Analysis.

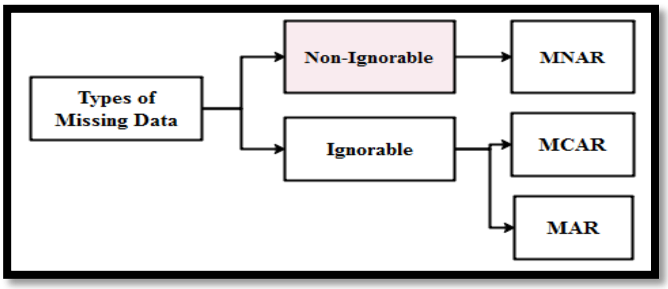
1.2.1 Complete Case Analysis

In such an approach missing data is directly addressed by excluding such records. In essence, analysis is done on the data or records that contain attributes having value. However, there two problems with this approach [4].

- If the units with missing values differ systematically from the completely observed cases, a bias would be incorporated into the dataset. For example, it is possible that the majority of missing attributes could be for one class, and removing these data records could result in ore complication in the analysis that includes bias and the introduction of an imbalance of classes.
- If there are a large number of variables required for the model, it is possible that the number of records available would be far less than is required.

1.2.2 Available Case Analysis

Available-case analysis arises when a variable or set of variables are completely excluded from the analysis because of their percentage of missing data. This method consists of determining the extent of missing data in each instance and attribute, and deletes the instances and/or attributes with high levels of missing data**.** In the context befitting a causal inference, it often leads to omitting a variable for satisfying the required assumptions for the preferred interpretation. Thus, before deleting any attribute, consideration must be given to whether that variable or attribute is necessary to the analysis [4][5]. In some situations, attributes should be retained even in the presence of a high degree of missing values. It is pertinent to note that only if missing data are MCAR, it is highly suggestible to apply methods, complete case analysis and discarding instances and/or attributes as there is a danger that the non-MCAR missing data can generate biased outcomes due to the presence of non-random elements.



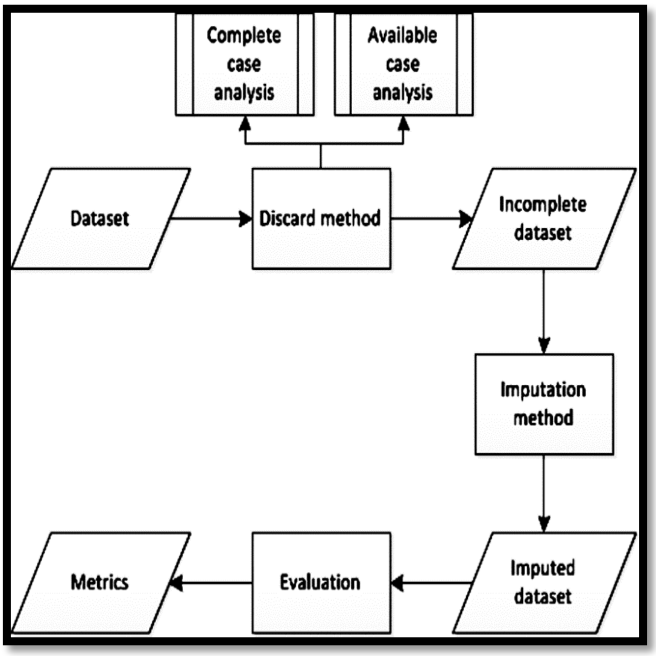


Fig.2. A multi-stage process for handling missing values

2. Imputation Methods

Imputation belongs to a class of universal methods that flexibly handle issues emanating from missing data. They represent the means that function under a predictive distribution of the missing values and often prescribe a predictive distribution to impute data, largely the observed data. Method of creating complete data via filling in missing value can be classified into single imputation and multiple imputation methods [11]. The units with missing values are compensated with the use of Imputation methods, which provided statistical analysis in the case of deficient data. In certain methods, the missing data information belonging to any specific subject is just used, while in other methods, the values of other subjects are used. The classification of Imputation methods is done on the basis of the quantum of imputed values required to replace the missing values in either single or multiple imputations. In cases of single imputation, s single value is applied against each missing value for imputation, whereas multiple imputations substitute the missing values with several values that produce many dissimilar absolute datasets. In this section, the review of these eight imputation methods has been undertaken. This method is different from complete-case and available-case analysis because rather than removing variables or observations with missing data, this approach is to fill in or impute missing values. At the same time, this method retains the full sample size [8] [14].

2.1 Mean Substitution Method (MS)

A subject having missing a value uses the mean of the variable as the most effective estimation for the variable. The missing values are filled in all observations by using the mean value of non-missing ones. In spite of the fact that mean substitution retains the unchanged sample size even after making the decline, it stills poses a few problems. In case of data containing high missingness rate, the application of mean imputation method adversely affects the way variables are distributed. Moreover, the analysis is likely to be complicated when the probable extreme values move to the middle of the distribution, thereby resulting in underestimating the variance that causes large kurtosis. When the mean imputation in the missing subjects contains zero variance, it means the covariance is underestimated as well. Additionally, such methods are like the Complete Case Analysis (CCA) that is dependent on MCAR hypothesis for obtaining neutral and proficient estimation, though may seem extremely limiting[8][9].

2.2 The Hot Deck (HOT) Method

Madow proposed this method that replaces missing values of units with identical responding units of the sample. The similarity criteria are applied to choose the responding unit, which is also done in a random manner. When one or more identical subjects have the same missing value in the sample, it is the closest resembling subject that replaces the missing values from his or her measurements. Moreover, such a method fills the missing value on the basis of the linkage that

Exists between the variable with missing data as well as other variables [14][15]. It acts better in the event of the variable sorting the data accurately predicts the variable having missing values as well as in cases of larger samples for ensuring identification of identical cases. In fact, the sampled values are distributed well in this method, other being simple in concept. Additionally, the similarity criterion employed here conserves certain measurement errors that may accrue when the respondents complete the value. This method offers the standard deviation of the variable with the inserted values as a better estimation of the standard deviation value for the variable with no replacement values. But, it is highly likely to have lower standard deviations in this method [13] [6].

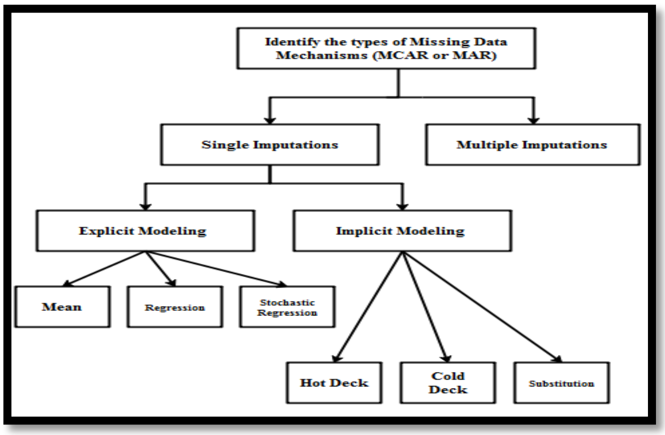


Fig. 3: Steps of managing missing values through Single Imputation.

2.3 K-nearest neighbour imputation (KNNI).

According to this, uses the k-nearest neighbors in order to determine a value from them, this is then imputed. Such a method defines the proximity measures connecting instances. A default or near universal measure is the Euclidean distance.Typically for nominal attributes, the most common value amongst all neighbors is taken, and for numerical values, the average value is used. Method with some advantages (i) more efficient than the mean imputation method (ii) provides asymptotically valid distribution.(iii) Usually, KNN imputation results in making point estimations, though with minor or insignificant bias. However, this method has the best application when MCAR is designated as the missingness method. In case of violations in the hypothesis of MCAR-oriented outcomes accrue. Moreover, the k-nearest neighbour approach is costly in terms of computation [13].

2.4 K-means clustering imputation (KMI)

The KMI method uses the dissimilarity measure within the cluster through the addition of distances among the objects and the centroid of the cluster to which they are assigned. The mean value of the objects in the cluster is represented by a cluster centroid. After the clusters converge, all the non-reference attributes for each unfinished object in the last process are filled on the basis of the cluster data. Data objects found in a single cluster are placed nearer to each other, wherein KMI is applied to a adjoining neighbor algorithm for replacing missing values much like KNNI [3].

2.5Fuzzy *K*-means clustering imputation (FKMI)

In this method, cluster having a cluster centroid does not accept a data object as all data objects belong to all *K* clusters having dissimilar membership degrees. Here, the non-reference attributes for each incomplete data object are linked to the membership degrees and cluster centroid values. The fuzzy clustering contains data objects with membership function based on the degree to which a data object is fixed in specific cluster [3].

2.6 Expectation-maximization imputation (EMI)

EMI iteratively computes the expected values for missing observations by repeatedly updating maximum-likelihood (ML) parameter estimates and imputing updated expected values until convergence is achieved. The expectation-maximization (EM) algorithm is an iterative method for solving the maximum-likelihood estimates for missing values. The algorithm proceeds in two steps (a) The E-Step: The expectation step evaluates the posterior probabilities of the unobserved data and (b) M-Step: The subsequent maximization step updates the model parameters using the posterior distribution of the missing data evaluated in the E-step [15] [10].

2.7 Support vector machine imputation (SVMI)

A support vector machine (SVM) is applied for imputing the missing values in all the attributes in a training set. An SVM uses all specimens even without any missing value. The decision attributes (output or classes) are used as the condition attributes (input attributes) and vice a versa, after which the SVM makes the prediction about the missing conditions and their attribute values. In such a method the original classification value from the dataset are ignored, besides using the value of the attribute imputed as the target value. All other attributes having missing values are ignored to generate fresh training data. In case attributes are continuously imputed, the SVMI uses regression for generating the value. In case of continuous attributes, each SVM models selects the value related to the SVM that classify the specimen as positive. One value is selected randomly in case of SVM generating a positive classification [14].

2.8 The Last Observation Carried Forward Method (LOCF)

The LOCF method handles missing data, specifically in dropout missingness. The unobserved values are imputed by the last observed value for the same subject. In case of dropout missingness, the last observed value is advanced towards the end of the study implying no change for the last observation post dropout. It has applications in longitudinal data for observing subjects at many places, while some subjects fizzle out after follow up or display intermittent missing values, besides being unrealistic in multiple contexts. This method underestimates the real variability of the data, often ignoring valid analyses in case the missingness mechanism not being MCAR. But, it satisfies a robust MCAR hypothesis even it has a bias, giving satisfactory results with the observations in the dataset closely set up by each other. In case of short measurement events, this method proves to be effective [12].

3. Results and Analysis.

3.1. Simulated Data

This simulation on a dataset for subjects with five measurement times evaluates how these eight imputation methods behave in the presence of the missing data methods. A sample size of *n* runs ranging from small to large. The sample sizes *n* =250, *n* = 500, and n= 1000 characterize small, moderate and, large sample sizes in an order on the assumption of having two covariates- the time “TIME” and the treatment group “MeD”. Therefore, the simulation of data is done as per the model given below:

$$\mathbf{Y_{ij}=\beta_0+\beta_1 Time_j+\beta_2 MeD_i+\varepsilon_{ij}}$$

Here, $\mathbf{Time_j}$, is coded 0, 1, 2, 3, 4 for five time points, while $\mathbf{MeD_i}$ represents a dummy variable with 0 value assigned to gesture group and 1 value in case of treatment group, following which applied is a simple linear regression model for the mean profiles of the repeated measurements

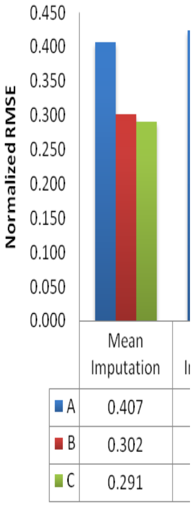
MISSING RATIO=30%			
	A	B	C
Mean Imputation	0.780	0.578	0.558
kNN Imputation	0.812	0.707	0.485
SVD Imputation	0.651	0.523	0.575
SVT Imputation	0.720	0.522	0.585
EM Imputation	0.565	0.509	0.470
LOCF Imputation	0.367	0.331	0.305

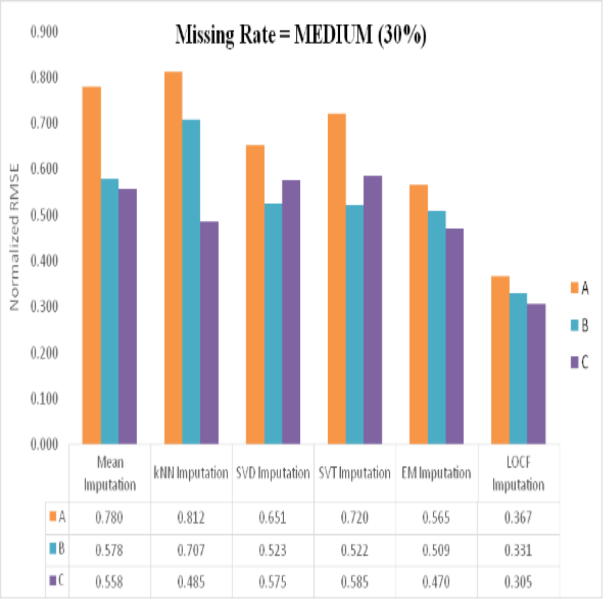
$\mathbf{E(y_{ij})=\mu_j, j=0,1,2,3,4}$, wherein variance-covariance structure is treated as first-order autoregressive **Auto-regression (AR)** (1). It fixes the parameters at $\mathbf{\beta_0=1,\beta_1=0.25\ and\ \beta_2=-1}$, wherein ε_i ’s were obtained from a multivariate normal with zero mean and $\mathbf{V(\varepsilon_{ij})=\sigma^2=1}$, while the correlation coefficient is expressed as $\rho=\mathbf{0.5}$.

Case	Simulations	AR Parameters	Correlation Coefficient
A	250	$\beta_0=1,\beta_1=0.25$ and β_2=-1	$\rho=\mathbf{0.25}$
B	500	$\beta_0=1,\beta_1=1.25$ and β_2=-2	$\rho=\mathbf{1.25}$
C	1000	$\beta_0=1.5,\beta_1=2$ and β_2=-4	$\rho=\mathbf{1.75}$

The comparison between methods depends on Mean Square Error (MSE).The generalized least squares - (**GLS**) method is used for estimating the unknown parameters in the linear regression model. The parameter estimates have been obtained for the selected imputation methods: Mean, KNN, singular value decomposition-SVD, singular value thresholding-SVT and LOCF methods. For the MCAR situation, the data are simulated with dropoutrates 5%, 30%, and 50%. If subject is missing at completely.

MISSING RATIO=10%			
	A	**B**	**C**
Mean Imputation	0.407	0.302	0.291
kNN Imputation	0.424	0.369	0.253
SVD Imputation	0.340	0.273	0.300
SVT Imputation	0.376	0.273	0.306
EM Imputation	0.295	0.266	0.245
LOCF Imputation	0.236	0.213	0.196





MISSING RATIO=50%			
	A	**B**	**C**
Mean Imputation	1.061	0.923	0.863
kNN Imputation	1.164	0.862	0.832
SVD Imputation	1.211	1.055	0.724
SVT Imputation	0.972	0.781	0.858
EM Imputation	1.075	0.779	0.874
LOCF Imputation	0.843	0.759	0.701

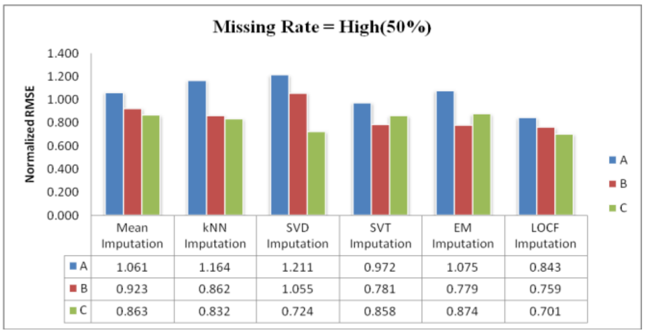
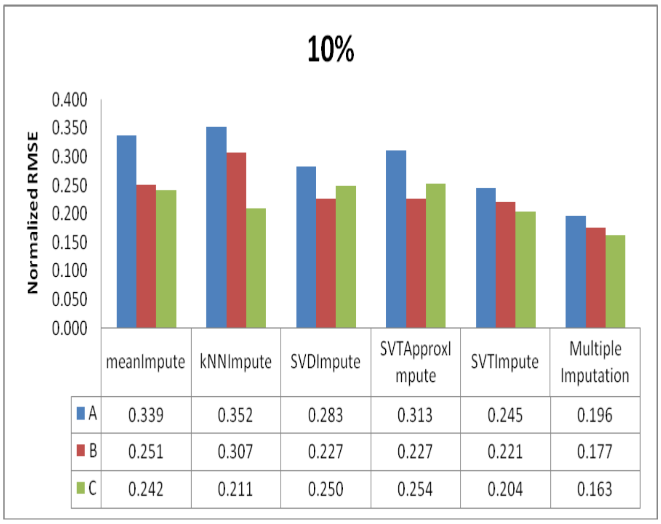


Fig.4. Performance evaluation with Normalized root mean square error (NRMSE) after m imputations (i.e., m=10, 30 and 50) with different sample size MCAR missingness mechanisms.

The results MAR mechanisms is shown below.



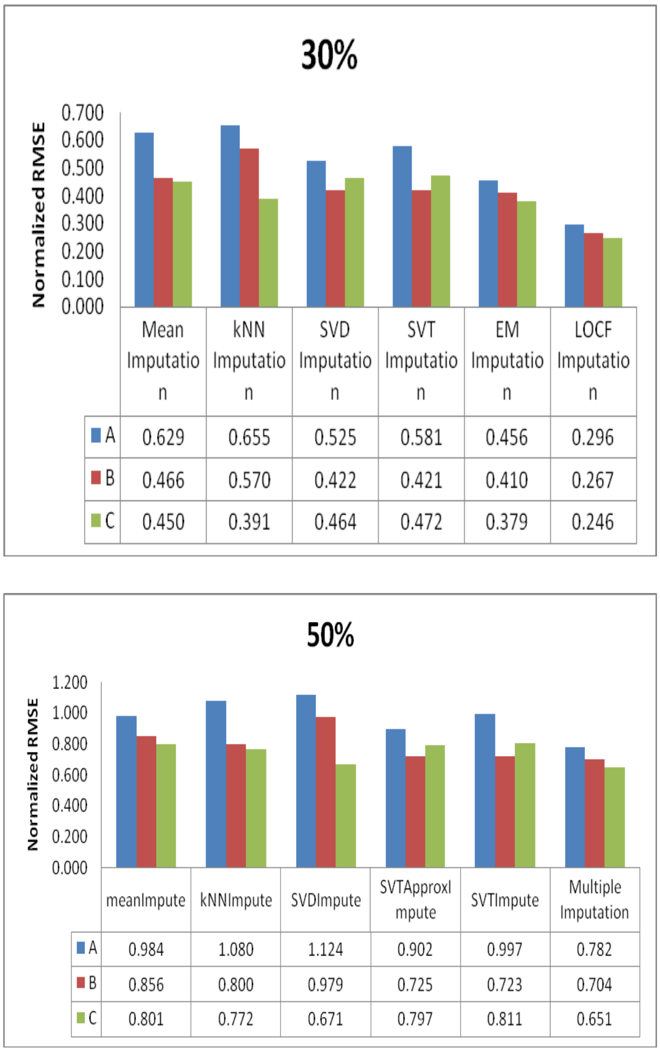


Fig.5. Performance evaluation with NRMSE after m imputations (i.e., m=10, 30 and 50) with different sample size MAR missingness mechanisms

The outcomes justify the occurrence of many standard errors in the approximate use of LOCF method, resulting in the least efficiency. The previous observation comes in the place of the missing value in the LOCF method, implying the failure of the imputed value in accurately predicting the missing value. In fact, LOCF may fail to efficiently impute unless each time point values are proximate to each other.

4. Conclusion

It may be concluded that the Mean imputation method does not offer an effective mechanism for the dropout pattern in alignment with the MCAR assumption. However, its performance is somewhat better in case of MAR assumptions, producing lesser NRMSE in comparison with rest of the methods. In the LOCF, the parameter is estimated better, thereby yielding smaller NRMSE, with the exception of MAR assumption, though large bias is perceptible in certain parameters. When aligned with MAR missingness, the EM method succeeds in countering large biases, which get better in large samples under the MCAR and the MAR. But, when examined under all the three missingness mechanisms, the EM scores a smaller MSE. In fact, it's the KNN that performs reasonably well under MCAR and the MAR mechanisms, yielding better results with bigger sample sizes. This clearly shows its better applicability to larger sample sizes as compared to smaller sample sizes. On the other hand, the EM algorithm fails to make any substantial success in accurately predicting the missing values when the three missing data mechanisms are applied, including the MCAR. Nevertheless, here a smaller NRMSE is produced in comparison to any other method. Moreover, though it is not possible to avoid relatively biased estimates in Multiple Imputation (MI) method, yet the smallest amount of bias is noticed when applied to MCAR mechanism.

5. References

[1] Brick JM, Kalton G. Handling missing data in survey research. Statistical methods in medical research 5(3):215-38, 1996.

[2] Graham JW. Missing Data: Analysis and Design. New York: Springer Science Business Media, 2012.

[3] Gelman, A. and Hill,J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge, 2007.

[4] Little, R.J.A. "Regression with Missing X's: A Review." Journal of the American Statistical Association, 87, pp.1227-1237, 1992.

[5] Little R.J.A. *and* Rubin D. B. Statistical analysis *with* missing data (2nd ed.*)Multiple Imputation for Nonresponse in Surveys*, New York, NY, & John Wiley, 2002.

[6] Little, R.J.A. and Rubin, D.B. Statistical Analysis with Missing Data. John Wiley & Sons, New York, 1987.

[7] Rubin, D.B. "Multiple imputation after 18+ years (with discussion)." Journal of the American Statistical Association, 91, pp.473-489, 1996.

[8] Schafer, J.L. The Analysis of Incomplete Multivariate Data. Chapman & Hall, London, 1997.

[9] S. Thirukumaran and A. Sumathi, "Missing value imputation techniques depth survey and an imputation Algorithm to improve the efficiency of imputation," Advanced Computing (ICoAC), 2012 Fourth International Conference on, Chennai, 2012.

[10] Soullier, N, Rochebrochard , E., and Bouyer , J. "Multiple Imputation for Estimation of an Occurrence Rate in Cohorts with Attrition and Discrete Follow-up Time Points: A Simulation Study." BMC Medical Research Methodology, 10, 79, pp.1-7, 2010.

[11] Schafer, J. L. and Olsen, M. K. "Multiple Imputation for Multivariate Missing-Data Problems: A Data

Analyst’s Perspective.” Multivariate Behavioral Research, 33, 4, pp.545-571.1998.

[12] Tibshirani, R. “Regression Shrinkage and Selection via the Lasso.” Journal of the Royal Statistical Society. 58, 1, pp.267-288, 1996.

[13] Van Buuren S, Oudshoorn K. Multivariate Imputation by Chained Equations: MICE V1.0 User's manual, volume PG/VGZ/00.038. TNO Prevention and Health, Leiden, 2000.

[14] Wang, K. and Jiang, W. “High-Dimensional Process Monitoring and Fault Isolation via Variable Selection.” Journal of Quality Technology, 41, 3, pp.247-258, 2009.

[15] Zhu, M., Chipman, H.A. “Darwinian Evolution in Parallel Universes: A Parallel Genetic Algorithm for Variable Selection.” Technometrics, 48, 4, pp. 491-502, 2006.

Journal of AL-Qadisiyah for computer science and mathematics Vol.11 No.2 Year 2019
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

اختيار طرق احتساب مناسبة للبيانات المفقودة بتحديد طرق لوغارتمية للبيانات

وسام علي محمود [1] محمد صباح رشيد [2] طيبة ولاء الدين [3]

قسم علوم الحاسوب / الجامعة التكنولوجية/ بغداد/ العراق

المستخلص:

القيم المفقودة التي تحدث في مجال البحث الطبي ، والتي تنشأ الكثير من حالة عدم الاستقرار في حالة إهمالها مع سوء التعامل. ومع ذلك ، عند التعامل مع مثل هذه التحديات ، تم بالفعل تطوير بعض الأساليب الإحصائية القياسية المتاحة ، ولكن حتى الآن لا توجد طريقة موثوقة متاحة لاستخلاص تلك التقديرات. حيث يتم تقليل حجم البيانات الموجودة ،بعيد عن انخفاض الكفاءة عندما يتم العثور على القيم المفقودة في مجموعة بيانات. بعض الطرق العادية التي تشمل طريقة الحالة الكاملة ، طريقة خصم الوسيط الحسابي ، طريقة ترحيل المراقبة الأخيرة (LOCF) ، خوارزمية زيادة التوقعات (EM) ، و سلسلة ماركوف مونتي كارلو (MCMC) ، خوارزمية زيادة التوقعات ،Hot Deck (HOT) , تراجع الانحدار (تراجع) ,أقرب جار (KNN) ، متوسط التجميع K،K - متوسط العنقودية الضبابي، دعم آلة المتجهات ، طريقة إسناد متعددة (MI) في هذه البحث ، تم محاولة إجراء دراسة محاكاة لإجراء استكشاف تحقيقي في فعالية أساليب الاستناد التكراري المذكورة أعلاه جنبا إلى جنب مع إعداد البيانات الطولية تحت المفقودين تماما على نحو عشوائي (MCAR). اخذنا حالة من فقدان ثلاث حالات في كتلة انخفاض نسبة فقدانها 5 ٪ وكذلك مستويات أعلى عند 30 ٪ و 50 ٪. مع دراسة المحاكاة هذه ،استنتجنا أن طريقة LOCF لها تحيز أكثر من الطرق الأخرى في معظم الحالات بعد إجراء المقارنة .