Egyptian Computer Science Journal (ISSN: 1110 – 2586)
Volume 40 – Issue 03, September 2016

Handling of Missing Data

Maha A. Hana
Canadian International College, Engineering Institute (deputed from Department of Information Systems, Faculty of Computers and Information, Helwan University)
maha_attia@cic-cairo.com, maha_hana_eg@yahoo.com

Abstract

One of the success factors for any scientific research is acquisition of real high quality data. This research aims to examine different imputation methods within a data cleaning process and identifies the best performance one.The research proposes a Best Imputation System "BIS" that has four phases. Data preprocessing phase results in complete, validate, non-duplicated and normalized data. Data preparation phase induces missing variable values randomly. The imputation phase tests nine different methods and the performance phase evaluates Mean Square Error and CPU processing time. BIS is tested by different data subsets extracted from five data sets with up to two missing variables and five levels of missing value percent using 5-fold cross validation. The results indicate that the bestperformed imputation methods are similarity measures and the worst one is Mode.

Keywords: *Data Mining and Knowledge Discovery, Data Cleaning,Data Imputation.*

1. Introduction

Real data is essential in studying any phenomenon especially in knowledge discovery, data mining and machine learning. Collecting real data is a scientific process which is featured by being practical, applicable, reliable and accurate. The acquisition of real data is usually encountered by difficulties and challenges as the collection process design,the selection of a reliable accurate measurement device, the handling of noise and the choice of analytic technique. So, the term data cleaning or data cleansing is introduced as the process of correcting anomalies in a data source or multiple data source resulting in high quality data[1, 2]. Causes for data cleaning are abbreviation misuse, embedded control data, mistakes in data entry, missing values, duplicate records, spelling variations, different formats, different measurement units, legacy data, inaccuracy in data measurement, measurement device limitation, data transformation and data discrizartion errors [3]. Typical data cleaning steps are noise effect reduction, elimination of outliers, removal of unnecessary duplicates, data standardization, data unification and data normalization. Clean data is accessible, accurate, correct, current, valid, complete, and interpreted [2]. Extra features of clean data are meaningfulness, availability, integrity and appropriateness for processing. Also, it must be normalized, unified, standardized, well organized, fully described,easily administrated and secured. [4] Investigated the weakest condition that enables the presences of missing data without affecting correct results and reported that a process can be ignored if two simple conditions exist. First, both missing and observed data are missed and observed at random. Second, the parameters of the distribution for the missing data and missing data process are distinct from each other. [4]Concluded that the process that causes missing data should be studied extensively. The idea of this research is to find an imputation method(s) that result in

predicted data values that are as close as possible to the original one without constrains or assumptions on data distribution. This requires imposing missing variable values to compare it with original ones. The proposed system is called Best Imputation system, "BIS". BIS examinesnine different imputation methodson several data sets with different missing variable(s) values.Section two in this paper reviews research done on handling missing data values, section three describes the proposed system, section four describes the conducted experiments, section five demonstrates the results and section six concludes the research.

2. Related Work

Since [4] investigated the causes of missingness, it started to be regarded as a probabilistic phenomenon.The probabilistic distribution indicates the type and the rate of patterns of missing values.It is classified into missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR occurs if the probability of missing data is the same for all cases, or does not depend on any observed data. MAR or (Ignorable nonresponse) occurs if the probability of missing data depends on some of the observed data. MAR is more general and more realistic than MCAR. MNAR or (Nonignorable) occurs if theprobability of missing data depends on some of the missed data values.There are different approaches to handle missed data that are Complete-Case Analysis [5, 6, 7, 8, 9, 10, 11, 12], Weighted Complete-Case Analysis [6], Pair-wise deletion (Available-Case Analysis) [6, 7, 9, 11, 12,13], Imputation [7, 8, 9, 10, 13, 14], Model-based method [8, 10, 11, 12] and Similar response pattern imputation [10, 13]. Complete-Case Analysis (Casewise Deletion, Listwise Deletion) excludes all observations/records with missed values. Weighted Complete-Case Analysis weights complete and incomplete observations/records differently. Pair-wise deletion creates multiple complete data sets by excluding specific missing values variables from the analysis.Imputation is either single [5, 7, 8, 9, 10, 13, 14, 15, 16]or multiple [9, 13] imputation. Single imputation replaces missed data value(s) by an approximate value(s). Single imputation includes several methods among them mean, median, regression, hot deck and cold deck, last observation carried forward, worst case imputation and Missing Indicator method. Regression Imputation uses linear or nonlinear, parametric or non-parametric and stochastic regression in order to estimate the missing values[5, 7, 9, 17]. Hot deck and Cold deck search for similar observations differently either from existing data set or from external source, respectively. The last observation carried forward "LOCF" assumes that all missed values have the same value as the last one [7]. This method is used in special situation where the condition of data values continuation exists. Worst case imputation replaces data value with its maybe worst value and is used to indicate that imputing does not affect the study outcomes [8]. Missing Indicator method recalculates the value of the variable to be the value of the indicator if missing or by the observed value if present [14]. In Multiple Imputation, a number of data sets are generated from the incomplete original set and imputation is performed for each data set. Some researchers recommend Model-Based method [8, 10, 13, 18] in which neither missed values are imputed nor handling of incomplete observations occurs. The techniques within this category are maximum likelihood [6, 8, 10], Neural Networks [19], Bayesian network [20], Linear discriminate analysis [8, 13] and K-nearest neighbor [8, 13, 15, 16]. A typical technique in Model-Based method is the full information maximum likelihood method which is able to analyze incomplete data sets from a multivariate normal distribution in order to predicate estimates. Expectation Maximization algorithm "EM" uses observed data to find the maximum likelihood parameter of a statistical model [15, 16, 18]. EM algorithm assumes that there are unobserved data and the statistical model parameters are unknown. It uses two steps

iteratively to build the model. The expectation step predicts the estimate using the maximum likelihood parameter and the second step maximizes the expected log-likelihood. Similar Response Pattern Imputation identifies the most similar complete records and imputes the missed value by the corresponding variable value in the complete record [10]. [10] Used four techniques on ERP data set which are full information maximum likelihood method, listwise deletion, mean imputation and similar response pattern imputation. Then, the resultant data sets are used to calibrate a regression-based prediction models. They reported that only full information maximum likelihood is appropriate when the data are not missing completely at random. While prediction models constructed on listwise deletion, mean imputation, and similar response pattern imputation data sets are biased unless the data are MCAR. [13] Examined the effect of four methods to deal with missing values on the accuracy of two classifiers; Linear discriminate analysis and K-nearest neighbor. The used imputation methods are case deletion, mean imputation, median imputation and K-nearest neighbor. They experimented with twelve data sets from the UCI Machine Learning Database Repository. They reported that there was no difference between case deletion and the other three methods in case of small percent of missing values. They concluded that KNN classifier performed better with KNN imputation than with mean imputation or median imputation.[19] Studied the use of dynamic programming concepts namely, an auto-associative neural networks and genetic algorithm to impute missing values using South African antenatal seroprevalence survey of 2001. They reported that the dynamic programming approach adds many advantages to the Base Model. [20] Proposed two methods based on Bayesian networks which are BN-K2IX^2 and 1BN-K2IX^2. Both methods used chi-squared test as a heuristics to determine the conditional relation among different variables in data set using four data sets from UCI Machine Learning repository. The two proposed methods were compared with Expectation-Maximization, Data Augmentation, Decision Trees and Mean/Mode methods in case of prediction. It was reported that estimated bias for two proposed Bayesian network was better for only two data sets.In case of classification, 1BN-K2IX^2 was better than the other methods. 1BN-K2IX^2 was better than BN-K2IX^2 in imputation with reasonable computational time.[21] Used Group Method of Data Handling "GMDH" which is an inductive hierarchal modeling for the input-output data structure. They proposed a method for imputation called Robust Imputation Based on GMDH "RIBG". RIBG was compared with four other methods using nine benchmarks from UCI Machine Learning Repository. The four methods are regression imputation, EM Imputation, Grey-based neighbor imputation and multiple imputation. The performance measure was the normalized mean absolute error. The results indicate that RIBG has the lowest error rate, greatest robustness and highest processing time.[22] Handled missing value imputation using three classifications methods categories; rule induction learning category, approximate models category and lazy learning category. The performance is evaluated using the Wilcoxon Signed Rank test. They used fourteen imputation method and twenty-three classification methods on twenty-one data sets from UCI Machine Learning Repository. For the rule induction learning category, the best method is Fuzzy K-means Clustering.For the approximate methods, the best method is Event Covering.For the lazy learning category, the best method is Global Most Common Attribute Value or Symbolic Attributes and Global Average Value for Numerical Attributes.[23] Proposed an automated approach to CiteSeerx metadata cleaning that uses DBLP an external data source to build a scholarly big data set. The record linkage algorithm starts by building an inverted index for entries in DBLP, then queries each entry in CiteSeerx against DBLP and similarity is measured using Jaccard and cosine similarity. They found that Jaccard performs better than the cosine similarity.

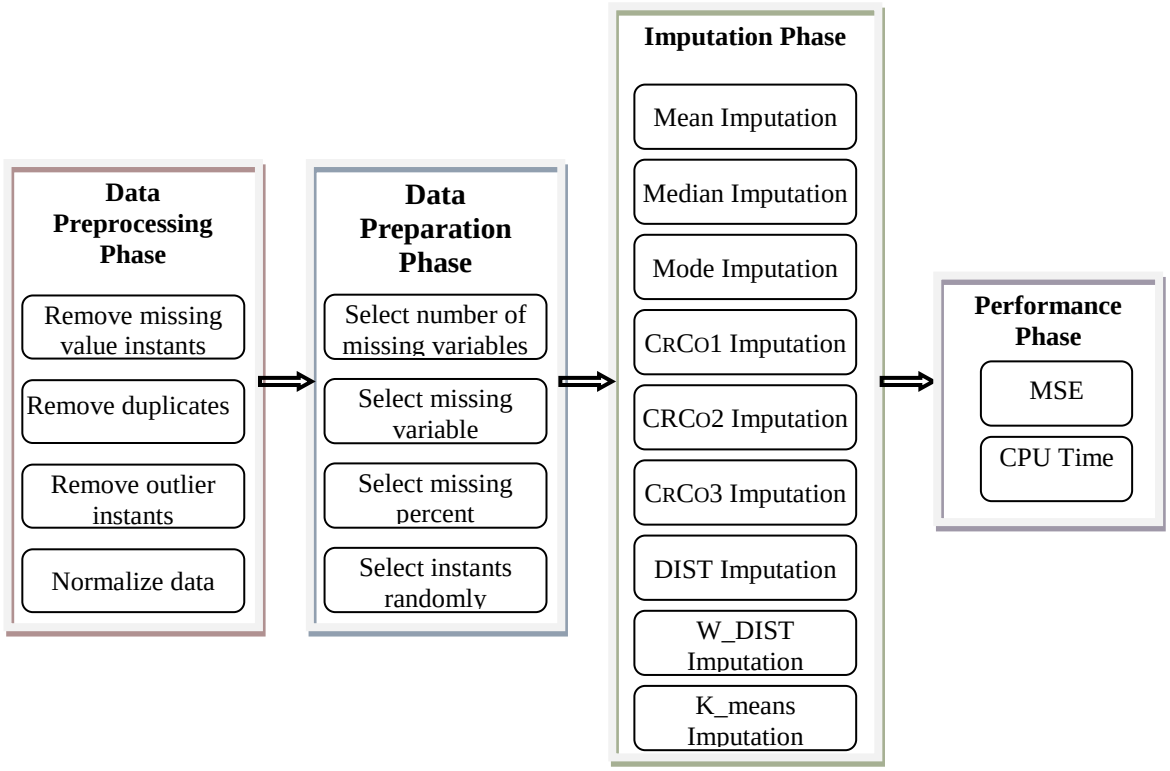
3. Best Imputation System "BIS"

BIS is a proposed system to find the best imputation method among several ones. BIS consists of four phases which are Data Preprocessing phase, Data Preparation phase, Imputationphase and Performance phase, Figure 1.

Figure 1. BIS Phases

3.1 Data preprocessing phase

This phase consists of foursteps, Figure 1. The first step removesall missing value observations/records from data set and the second $[\mu-n\sigma, \mu+n\sigma]$ step removes duplicate records from the data set. The third step is to detect and to remove outliers.An outlier is detected, if it satisfies one of the two conditions.The first condition is if the variable takes



discrete values, then the permissible variable value is determined by its all valid possible values. In the second condition, an outlierrecordfor a continuous variable is removed if the variable value lies outside a certain permissible rangewhere μ, σ and n are mean, standard deviation and scalar respectively. In the fourth step, data is normalized by rescaling any variable values to be in the range [-1, 1].

3.2 Data Preparation Phase

This phase generates multiple incomplete data sets from the original data set using induction of missing value process, Figure 1. The process is specified by two user selected parameters which are the number of variable(s) with missing values and the percent of missing values. The induction of missing value process chooses those variables randomly and forces missing values in randomly chosen records.

3.3 Imputation phase

In this phase, nine different methods are used to estimate the missing value of variable(s) which are:

1. Median imputation: it replaces missing values by median value.
2. Mode imputation: it replaces missing values by the most frequent value.
3. Mean imputation: it replaces missing values by average.
4. CrCo1 Imputation: it uses correlation coefficients to find the variable "Xh" that is highly correlated with missing variable "Xi" value. ScinceXi(k) value is missing at record k, then Xh(k) is used to impute the missing value. CrCo1 looks for records that have the value of Xh(k) and uses these recordsto calculate the mean value of variable Xi. The mean value replaces the missed one. In other words, CrCo1 assigns the missed value by the mean of Xifor all records that have the same value of Xh(k).
5. CrCo2 Imputation: It is the same as CrCo1, but uses the correlation coefficients to find the most similar record. CrCo2 replaces the missed value of variable Xi(k) by the corresponding value of the most similar record. The most similar record is determined as the record that has the minimum Euclidian distance. If more than one record have the same minimum distance, then the average value of Xi is imputed.
6. CrCo3 Imputation: It is the same as CrCo2, but uses a weighted Euclidian distance. The weight vector is selected to be the normalized correlation coefficients.
7. Distance Imputation "DIST": It calculates the similarity between each of missing value record and all given records by using Euclidian distance. The record(s) with the minimum distance is used to impute the missing variable value.
8. Weighted Distance Imputation "W-DIST": It is similar to DIST imputation, but uses a weighted Euclidian distance. The weight vector is selected to be the normalized correlation coefficients.
9. K_Means: It is an unsupervised clustering technique. It clusters data sets into a user specified number of clusters. A description about K-means is found in [24]. The cluster is totally specified by its centroid. Each record with a missing value is classified to one of the clusters and the missing variable value is substituted with the corresponding centroid variable value.

3.4 Performancephase

The performance of each imputation method is determined using two criteria. The first criterion is the most common criterion which is mean square error "MSE". MSE is the mean squareof difference between original values and the imputed values. This criterion measures the accuracy of the used imputed methods. The second criterion is comparing among imputation methods using the required CPU processing time. Also, it is important in case of imputing large datasets.

4. Experiment

Five data sets are used to evaluate the nine imputation methods, Table 1. Shuttle data set, Skin Segmentation and Abaloneare available from UCI learning machine repository [25]. The fourth data set is a twelve year students' data from Cairo Higher Institute for Engineering, Computer Science, and Management "CHI"[26]. The last data set is US January Temperatures Data file from StatLib website [27]. The first three data sets are divided in different subsets.That justification of using different data subsets is to study the relation between the imputation method performance and the size of the used data. Experiments used 5-fold cross

validation with one and two missing variable and missing value percent 1%, 5%, 10%, 15% and 20%. The used platform is HP Pavlion DV2000 Notebook PC.CPU is Intel(R) Core(TM) with 2GHz and RAM is 4GB. The code is written in Matlab.

Table 1. Data sets description,data subsets and experiments parameters

Data set	Area	Attribute characteristic	No. of Attributes	No. of Instances	Missing Values	Data Subset Size	No. of Missing Variables	Missing Value Percent
Shuttle	Physical	Integer	9	58000	No	15000,25000, 58000	1,2	1%, 5%, 10%, 15%, 20%
Skin Segmentation	Computer	Real	4	245057	No	100000, 245057		
Abalone	Life	Categorical, Integer, Real	8	4177	No	2000, 4177		
Students	Education	Categorical, Integer, Real	8	1826	Yes	1826		
US January Temperature	Metrology	Categorical, Integer, Real	3	56	Yes	56		

5. Results

MSE for each data subset for the different number of missing variables values is displayed in Table 2 and 3.The range of MSE from the all runs is from [0, 0.3646]. Also,8% of MSE equals to zero when MSE is approximated to thousandth. The majority of minimum values result from three imputation methods; CrCo3, DIST and W_DIST.There are several methodsamong them Median, Mode, CrCo2 andCrCo3that results in high MSE values. Table 4displays the MSE of the best and worst imputation methods for each experiment for each data subsets. Table 5 and Figure 2 show the MSE of each of the best and the worst imputation methods according to the number of missing variables. For one missing variable value, the best imputation methods are W_DIST, DIST and CrCo3 while the three worst methods are Mode, Median andMean. For two missing variables values, the best imputation methods are DIST, W_DIST, and CrCo3 while the three worst methods are Mode, CrCo3 and CrCo2. Therefore, it can be inferred that MSE of DIST and W_DIST methods arestable enough to be thebest imputation methods. It can be inferred that the Mode results in the worst imputation method. Also, CrCo2, Medianand Meanmethods results are inconsistent as sometimes they are within the best methods other times they are within the worst imputation methods.

The overall performance is determined by the frequency that the imputation method is performed as best or worst, Table 6 and Figure 3.The frequency for both DIST and W_DIST as the best imputation methods is 56 in value which equates to 62.22%. The frequency for both DIST and W_DIST as the worst imputation methods is 1 which equates to 1.11%. The third isCrCo3 with 37.78%.DIST and W_DIST converge to the best results more when the data subsets size increases. The worst imputation methods are MODE with 88.89% of the time, thenCrCo3 with 6.67%, the third is CrCo2 with 5.56%. It is reasonable to find that the Mode is the worst method as it doesn't take into accountthe infrequentor the full range ofdata values. The results of CrCo2, Medianand Mean yields swing between low and high MSE, therefore they are characterized as unstable methods. The result of Mean is ranked in the worst imputation method more than in the best ones. Table 7 and Figure 4 show the CPU

processing time for all imputation methods for full data sets. The rank of imputation methods according to CPU time is median and mode, mean, K-Means, CrCo1, CrCo3, CrCo2, DIST and W_DIST. CPU time for median, mode and mean is almost zero regardless of the data set size. Also, in generalW_DISTneeds 33% extra processing time than DIST in all data sets.

Table 2. MSE for imputation methods for all data subsets for 1 missing variable

Data subset	%	Median	Mode	Mean	CRCO1	CRCO2	CRCO3	DIST	W_DIST	K_Means
Shuttle data set (15000 instants, 1 missing variable)	1%	0.0243	*0.0266*	0.0226	0.0090	0.0054	**0.0001**	**0.0001**	**0.0001**	0.0039
	5%	0.0275	*0.0317*	0.0250	0.0098	0.0057	**0.0001**	**0.0001**	**0.0001**	0.0042
	10%	0.0281	*0.0310*	0.0259	0.0097	0.0059	**0.0001**	**0.0001**	**0.0001**	0.0042
	15%	0.0285	*0.0323*	0.0266	0.0105	0.0062	**0.0001**	**0.0001**	**0.0001**	0.0045
	20%	0.0297	*0.0344*	0.0276	0.0106	0.0063	**0.0001**	**0.0001**	**0.0001**	0.0044
Shuttle data set (25000 instants, 1 missing variable)	1%	0.0288	*0.0329*	0.0267	0.0107	0.0056	**0.0000**	**0.0000**	**0.0000**	0.0047
	5%	0.0268	*0.0312*	0.0252	0.0112	0.0058	**0.0000**	**0.0000**	**0.0000**	0.0047
	10%	0.0277	*0.0314*	0.0262	0.0116	0.0059	**0.0000**	**0.0000**	**0.0000**	0.0053
	15%	0.0278	*0.0336*	0.0263	0.0116	0.0058	**0.0000**	**0.0000**	**0.0000**	0.0042
	20%	0.0281	*0.0334*	0.0266	0.0120	0.0062	**0.0000**	**0.0000**	**0.0000**	0.0045
Shuttle data set (58000 instants, 1 missing variable)	1%	0.0270	*0.0340*	0.0256	0.0109	0.0057	**0.0000**	**0.0000**	**0.0000**	0.0050
	5%	0.0254	*0.0302*	0.0241	0.0111	0.0058	**0.0000**	**0.0000**	**0.0000**	0.0047
	10%	0.0271	*0.0336*	0.0257	0.0116	0.0059	**0.0000**	**0.0000**	**0.0000**	0.0046
	15%	0.0274	*0.0332*	0.0260	0.0119	0.0060	**0.0000**	**0.0000**	**0.0000**	0.0043
	20%	0.0284	*0.0344*	0.0269	0.0122	0.0062	**0.0000**	**0.0000**	**0.0000**	0.0048
Skin Segmentation data set (100000 instants, 1 missing variable)	1%	0.0511	*0.0831*	0.0505	0.0210	0.0211	0.0156	**0.0144**	**0.0144**	0.0230
	5%	0.0538	*0.0887*	0.0529	0.0211	0.0208	0.0146	**0.0135**	0.0136	0.0233
	10%	0.0536	*0.0871*	0.0530	0.0216	0.0211	0.0164	**0.0149**	**0.0149**	0.0225
	15%	0.0553	*0.0901*	0.0546	0.0225	0.0222	0.0173	**0.0157**	**0.0157**	0.0235
	20%	0.0570	*0.0937*	0.0562	0.0233	0.0229	0.0178	**0.0160**	**0.0160**	0.0241
Skin Segmentation data set (245057 instants, 1 missing variable)	1%	0.0519	*0.0875*	0.0515	0.0236	0.0235	0.0170	**0.0168**	**0.0168**	0.0254
	5%	0.0537	*0.0915*	0.0531	0.0243	0.0241	0.0199	**0.0192**	**0.0192**	0.0243
	10%	0.0544	*0.0919*	0.0539	0.0239	0.0237	0.0181	**0.0175**	**0.0175**	0.0263
	15%	0.0561	*0.0945*	0.0557	0.0248	0.0245	0.0187	**0.0180**	**0.0180**	0.0261
	20%	0.0576	*0.0972*	0.0569	0.0253	0.0251	0.0194	0.0187	**0.0186**	0.0278
Students data set (1826 instants, 1 missing variable)	1%	*0.0056*	*0.0056*	0.0055	0.0025	0.0030	0.0033	**0.0021**	**0.0021**	0.0036
	5%	*0.0069*	*0.0069*	*0.0069*	0.0036	**0.0026**	0.0039	0.0042	0.0039	0.0039
	10%	*0.0076*	*0.0076*	*0.0076*	0.0033	**0.0029**	0.0040	0.0042	0.0040	0.0042
	15%	*0.0066*	*0.0066*	*0.0066*	0.0034	**0.0029**	0.0038	0.0042	0.0041	0.0037
	20%	*0.0072*	*0.0072*	*0.0072*	0.0038	**0.0031**	0.0045	0.0050	0.0049	0.0041
Abalone data set (2000 instants, 1 missing variable)	1%	0.0512	*0.0630*	0.0512	0.0076	0.0076	0.0075	**0.0022**	**0.0022**	0.0068
	5%	0.0493	*0.0609*	0.0492	0.0075	0.0075	0.0076	**0.0054**	**0.0054**	0.0113
	10%	0.0479	*0.0576*	0.0477	0.0080	0.0080	0.0079	0.0038	**0.0037**	0.0108
	15%	0.0501	*0.0654*	0.0501	0.0085	0.0085	0.0085	**0.0048**	**0.0048**	0.0109
	20%	0.0522	*0.0623*	0.0523	0.0091	0.0091	0.0092	**0.0050**	**0.0050**	0.0119
Abalone data set(4177 instants, 1 missing variable)	1%	0.0466	*0.0527*	0.0466	0.0083	0.0083	0.0088	**0.0045**	**0.0045**	0.0115
	5%	0.0454	*0.0511*	0.0454	0.0084	0.0083	0.0083	**0.0046**	**0.0046**	0.0113
	10%	0.0489	*0.0636*	0.0489	0.0097	0.0096	0.0094	**0.0047**	**0.0047**	0.0133
	15%	0.0493	*0.0607*	0.0493	0.0092	0.0092	0.0093	**0.0055**	**0.0055**	0.0121
	20%	0.0538	*0.0598*	0.0537	0.0100	0.0099	0.0097	0.0053	**0.0052**	0.0125
January	1%	0.0175	0.0203	0.0160	**0.0120**	0.0150	0.0150	*0.0322*	*0.0322*	0.0293

Temperatures (56 instants, 1 missing variable)	5%	0.0199	0.0222	0.0190	0.0266	0.0293	*0.0328*	**0.0165**	**0.0165**	0.0170
	10%	**0.0166**	0.0179	0.0167	0.0202	*0.0234*	0.0225	0.0171	0.0171	0.0226
	15%	0.0148	0.0159	**0.0144**	0.0250	0.0250	*0.0266*	0.0207	0.0207	0.0183
	20%	0.0124	0.0172	**0.0118**	0.0152	*0.0186*	0.0182	0.0155	0.0155	0.0172

Table 3. MSE for imputation methods for all data subsets for 1 missing variable

Data subset	%	Median	Mode	Mean	CRCO1	CRCO2	CRCO3	DIST	W_DIST	K_Means
Shuttle data set (15000 instants, 2 missing variables)	1%	0.0134	*0.0147*	0.0121	0.0046	0.0029	**0.0003**	**0.0003**	**0.0003**	0.0025
	5%	0.0155	*0.0170*	0.0139	0.0055	0.0037	**0.0007**	**0.0007**	**0.0007**	0.0031
	10%	0.0148	*0.0171*	0.0137	0.0057	0.0036	**0.0006**	0.0007	0.0007	0.0030
	15%	0.0159	*0.0174*	0.0147	0.0060	0.0038	**0.0007**	**0.0007**	**0.0007**	0.0030
	20%	0.0164	*0.0188*	0.0155	0.0062	0.0039	**0.0007**	**0.0007**	**0.0007**	0.0029
Shuttle data set (25000 instants, 2 missing variables)	1%	0.0145	*0.0171*	0.0137	0.0069	0.0040	**0.0007**	**0.0007**	**0.0007**	0.0031
	5%	0.0140	*0.0164*	0.0133	0.0061	0.0033	**0.0002**	**0.0002**	**0.0002**	0.0027
	10%	0.0146	*0.0165*	0.0138	0.0063	0.0035	**0.0003**	**0.0003**	**0.0003**	0.0029
	15%	0.0161	*0.0188*	0.0149	0.0069	0.0037	**0.0004**	**0.0004**	**0.0004**	0.0028
	20%	0.0162	*0.0193*	0.0152	0.0072	0.0039	**0.0004**	**0.0004**	**0.0004**	0.0031
Shuttle data set (58000 instants, 2 missing variables)	1%	0.0145	*0.0182*	0.0136	0.0058	0.0030	**0.0009**	**0.0009**	**0.0009**	0.0027
	5%	0.0140	*0.0172*	0.0134	0.0067	0.0036	**0.0005**	**0.0005**	**0.0005**	0.0031
	10%	0.0147	*0.0180*	0.0139	0.0067	0.0037	**0.0004**	**0.0004**	**0.0004**	0.0030
	15%	0.0157	*0.0185*	0.0150	0.0070	0.0039	**0.0005**	**0.0005**	**0.0005**	0.0031
	20%	0.0169	*0.0212*	0.0160	0.0075	0.0042	**0.0006**	**0.0006**	**0.0006**	0.0035
Skin Segmentation data set (100000 instants, 2 miss. variables)	1%	0.0734	*0.1976*	0.0732	0.0423	0.0411	0.0296	0.0276	**0.0270**	0.0589
	5%	0.0774	*0.2108*	0.0769	0.0484	0.0424	**0.0320**	0.0321	0.0326	0.0557
	10%	0.0827	*0.2227*	0.0822	0.0573	0.0448	**0.0349**	0.0363	0.0365	0.0590
	15%	0.0879	*0.2362*	0.0874	0.0679	0.0486	**0.0394**	0.0419	0.0424	0.0611
	20%	0.0928	*0.2505*	0.0922	0.0765	0.0511	**0.0420**	0.0467	0.0465	0.0653
Skin Segmentation data set (245057 instants, 2 miss. variables)	1%	0.0723	*0.1984*	0.0720	0.0418	0.0406	0.0336	**0.0320**	**0.0320**	0.0561
	5%	0.0770	*0.2122*	0.0765	0.0491	0.0499	0.0392	**0.0344**	0.0345	0.0563
	10%	0.0816	*0.2241*	0.0812	0.0582	0.0628	0.0534	**0.0403**	0.0404	0.0614
	15%	0.0866	*0.2356*	0.0861	0.0683	0.0738	0.0652	**0.0455**	0.0456	0.0642
	20%	0.0920	*0.2489*	0.0917	0.0773	0.0781	0.0697	**0.0502**	0.0504	0.0613
Students data set (1826 instants, 2 missing variables)	1%	0.0121	*0.0172*	0.0120	0.0087	**0.0083**	0.0094	0.0106	0.0105	0.0091
	5%	0.0128	*0.0178*	0.0127	0.0101	**0.0099**	0.0149	0.0157	0.0156	0.0105
	10%	0.0141	*0.0195*	0.0139	0.0111	**0.0105**	0.0162	0.0161	0.0157	0.0109
	15%	0.0158	*0.0220*	0.0156	0.0128	**0.0125**	0.0183	0.0191	0.0190	**0.0125**
	20%	0.0156	*0.0219*	0.0154	0.0125	**0.0119**	0.0187	0.0177	0.0176	0.0122
Abalone data set (2000 instants, 2 missing variables)	1%	**0.1029**	*0.2343*	0.2340	0.1161	0.1156	0.1357	0.1711	0.1679	0.1793
	5%	**0.1184**	*0.2694*	0.2630	0.1434	0.1429	0.1758	0.1831	0.1842	0.1870
	10%	**0.1299**	*0.2810*	0.2766	0.1653	0.1650	0.1957	0.1741	0.1749	0.1729
	15%	**0.1369**	*0.2787*	0.2769	0.1447	0.1447	0.1986	0.2097	0.2104	0.1813
	20%	**0.1509**	*0.3143*	0.3052	0.1896	0.1894	0.2221	0.2091	0.2078	0.2159
Abalone data set(4177 instants, 2 missing variables)	1%	**0.1040**	*0.2737*	0.2192	0.1101	0.1118	0.1469	0.1526	0.1526	0.1459
	5%	**0.1259**	*0.3011*	0.2466	0.1290	0.1289	0.1822	0.1647	0.1646	0.1934
	10%	0.1280	*0.3203*	0.2613	**0.1108**	**0.1108**	0.1831	0.1698	0.1696	0.1604
	15%	0.1391	*0.3446*	0.2865	**0.1195**	**0.1195**	0.2018	0.1780	0.1786	0.1891
	20%	**0.1539**	*0.3646*	0.3133	0.1576	0.1576	0.1962	0.1902	0.1902	0.2040
January	1%	0.0201	0.0275	0.0201	0.0205	*0.0286*	*0.0286*	0.0050	**0.0037**	0.0268

Temperatures (56 instants, 2 missing variables)	5%	0.0127	0.0150	0.0120	0.0158	0.0291	*0.0314*	**0.0088**	**0.0088**	0.0126
	10%	0.0114	0.0120	0.0117	0.0119	*0.0323*	*0.0323*	**0.0048**	**0.0048**	0.0079
	15%	0.0150	0.0178	0.0149	0.0165	*0.0246*	0.0241	**0.0089**	0.0090	0.0118
	20%	0.0168	0.0270	0.0161	0.0252	0.0386	*0.0389*	0.0115	**0.0112**	0.0125

Table 4. Best and worst imputation methods for each data subset and each experiment

Data subset	%	1 Missing Variable Value		2 Missing Variable Values	
		Best Imputation Method	**Worst Imputation Method**	**Best Imputation Method**	**Worst Imputation Method**
Shuttle (15000 instants, 1 missing variable)	1%	CrCo3, DIST, W_DIST	Mode	CrCo3, DIST, W_DIST	Mode
	5%				
	10%			CrCo3	
	15%			CrCo3, DIST, W_DIST	
	20%				
Shuttle (25000 instants, 1 missing variable)	1%	CrCo3, DIST, W_DIST	Mode	CrCo3, DIST, W_DIST	Mode
	5%				
	10%				
	15%				
	20%				
Shuttle (58000 instants, 1 missing variable)	1%	CrCo3, DIST, W_DIST	Mode	CrCo3, DIST, W_DIST	Mode
	5%				
	10%				
	15%				
	20%				
Skin Segmentation (100000 instants, 1 missing variable)	1%	DIST, W_DIST	Mode	W_DIST	Mode
	5%	DIST		CrCo3	
	10%	DIST, W_DIST			
	15%				
	20%				
Skin Segmentation (245057 instants, 1 missing variable)	1%	DIST, W_DIST	Mode	DIST, W_DIST	Mode
	5%			DIST	
	10%				
	15%				
	20%	W_DIST			
Students (1826 instants, 1 missing variable)	1%	DIST, W_DIST	Median, Mode	CrCo2	Mode
	5%	CrCo2	Median, Mode, Mean		
	10%				
	15%			CrCo2, K_Means	
	20%			CrCo2	
Abalone (2000 instants, 1 missing variable)	1%	DIST, W_DIST	Mode	Median	Mode
	5%				
	10%	W_DIST			
	15%	DIST, W_DIST			
	20%				
Abalone (4177 instants, 1 missing variable)	1%	DIST, W_DIST	Mode	Median	Mode
	5%				
	10%			CrCo1, CrCo2	
	15%				
	20%	W_DIST		Median	
January	1%	CrCo1	DIST, W_DIST	W_DIST	CrCo2, CrCo3

Tempratures (56 instants, 1 missing variable)	5%	DIST, W_DIST	CrCo3	DIST, W_DIST	CrCo3
	10%	Median	CrCo2		CrCo2, CrCo3
	15%	Mean	CrCo3	DIST	CrCo2
	20%		CrCo2	W_DIST	CrCo3

Table 5. Performance ofbest and worst imputation methods

Imputation Method	1 Missing Variable Value		2 Missing Variable Values	
	Freq. of Best Results	Freq. of Worst Results	Freq. of Best Results	Freq. of Worst Results
Median	1	5	8	0
Mode	0	40	0	40
Mean	2	4	0	0
CRCO1	1	0	2	0
CRCO2	4	2	7	3
CRCO3	15	2	19	4
DIST	34	1	22	0
W_DIST	36	1	20	0
K_Means	0	0	1	0

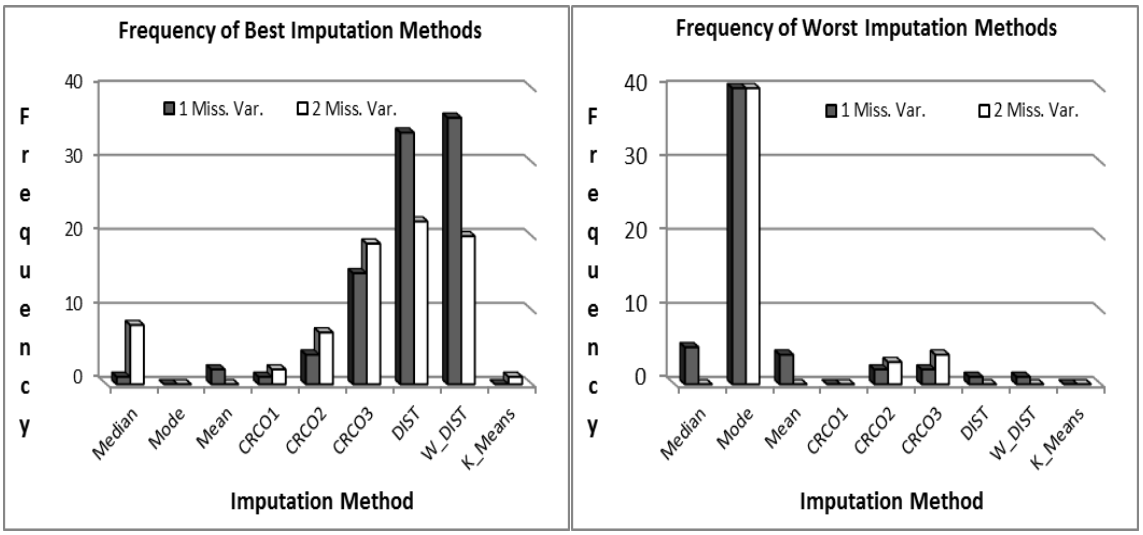


Figure 2. Performance of best and worst imputation methods

Table 6. Overall best and worst imputation methods

Imputation Method	% Best Results	% Worst Results
Median	10.00%	5.56%
Mode	0.00%	88.89%
Mean	2.22%	4.44%
CRCO1	3.33%	0.00%
CRCO2	12.22%	5.56%
CRCO3	37.78%	6.67%
DIST	62.22%	1.11%
W_DIST	62.22%	1.11%
K_Means	1.11%	0.00%

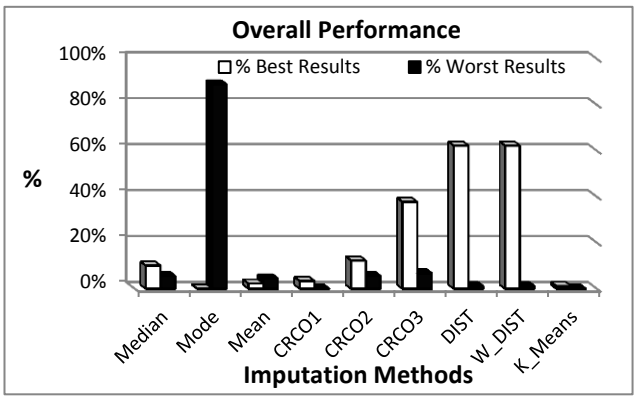


Figure 3. Overall best and worst imputation methods

Table 7. CPU processing time for all imputation methods

Data set	Median	Mode	Mean	CRCO1	CRCO2	CRCO3	DIST	W_DIST	K_Means
Shuttle data set (58000 instants, 2 missing variables)	0.000	0.000	0.003	345.558	3661.574	2727.241	2368.148	3146.022	5.547
Skin Segmentation data set (245057 instants, 2 missing variables)	0.000	0.000	0.000	20.299	111.812	89.988	3481.462	4531.632	7.625
Students data set (1826 instants, 2 missing variables)	0.000	0.000	0.000	0.262	1.788	1.123	8.290	11.060	0.555
Abalone data set (4177 instants, 2 missing variables)	0.000	0.000	0.001	122.040	1258.391	939.451	1952.633	2562.905	4.576
January Temperatures (56 instants, 2 missing variables)	0.000	0.000	0.000	0.003	0.006	0.012	0.016	0.006	0.078

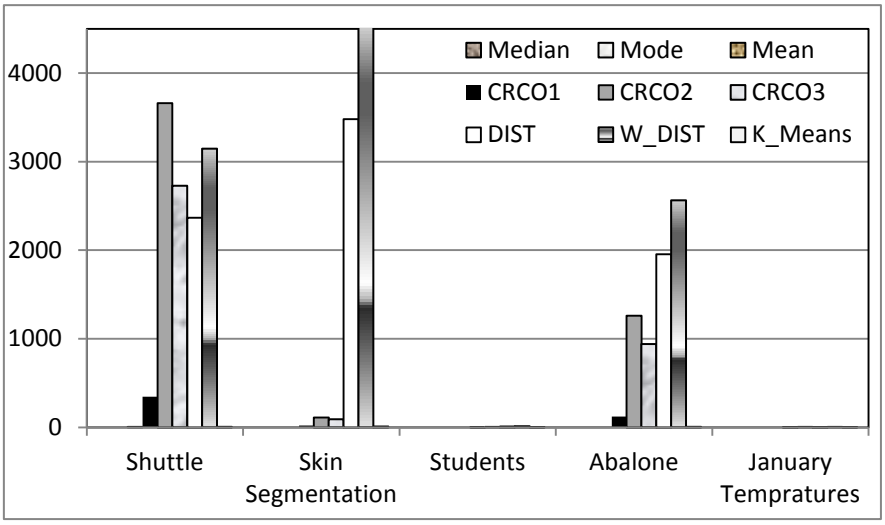


Figure 4. CPU processing time

6. Conclusion

Data cleaning is an essential in studying any research phenomena. The better the quality of data, the more realistic and trustful are the results. Since missing data is unavoidable and processing of Big data is gaining more attention these days, then the focus of simple ways for imputation methods are recommended. This research proposes an imputation system to compare among nine different imputation methods within the process of data cleaning. MCAR missingness is imposed on five data sets with different number of missing variables values and different missing percent. The chosen imputation methods are chosen either because of their popularity, simplicity or imposing no constrains on data distribution. The results show that DIST and W_DIST are the best imputation methods above 62% of the time. Thus, it can be concluded that DIST is more preferable than W_DIST for its simplicity and less computation time. Also, the results show that Mode imputation results in high MSE about89% of the time.

7. Acknowledgement

The author would like to thank University of California, School of Information and Computer Science, Center of Machine Learning and Intelligent System for providing free data through UCI Machine Learning Repository that helps researchers to investigate scientific problems. Also, the author is grateful to Cairo Higher Institute for Engineering, Computer Science, and Management "CHI" to provide real data for research.The author appreciate Data, Software and News from the Statistics Community for providing DASL an online Library of free data files and stories to help scientist to work with through their website StatLib.

References

[1]. L. Ciszak, "Application of clustering and association methods in data cleaning," Proceedings of the International Multiconference on Computer Science and Information Technology, vol. 3, 2008, pp. 97 - 103.

[2]. M. Weis, and I. Manolescu, "Declarative XML data cleaning with XClean," Advanced Information Systems Engineering,:Springer Berlin, Heidelberg, 2007.

[3]. K. Vilas, "Big data mining," International Journal of Computer Science and Managment Research,vol.1 (1), 2012, pp.12 - 17.9

[4]. D. Rubin, "Inference and missing data," Biometrika, vol. 63(3), 1976, pp.581-592.

[5]. Y. Qin,S. Zhang, X. Zhu, J. Zhang, and C. Zhang,"Semi-parametric optimization for missing data imputation," Applied Intelligence, vol. 27(1), 2007, pp. 79-88.

[6]. J. Schafer, and J. Graham, "Missing data: our view of the state of the art," Psychological Methods, vol. 7(2), 2002, pp. 147-177.

[7]. S. Van Buuren,"Flexible imputation of missing data," CRC Press, 2012.

[8]. Baraldi, and C. Enders, "An introduction to modern missing data analyses," Journal of School Psychology vol.48(1), 2010, pp. 5-37.

[9]. J. Haukoos and C. Newgard, "Advanced statistics: missing data in clinical research—part 1: an introduction and conceptual framework," Academic Emergency Medicine, vol. 14(7), 2007, pp 662-668.

[10]. Myrtveit, E. Stensrud andU. Olsson, "Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods," IEEE Transactions on Software Engineering, vol.27(11), 2001,pp. 999-1013.

[11]. J. Kaiser and C. Republic, "Dealing with missing values in data," Journal of Systems Integration,vol.5(1), 2014, pp. 42-50.

[12]. L.Silva and L. Zárate, "A brief review of the main approaches for treatment of missing data,"Intelligent Data Analysis,vol.18(6),2014, pp.1177-1198.

[13]. E. Acuna and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," Classification, Clustering, and Data Mining Applications: Springer Berlin Heidelberg, 2004, pp.639-647.

[14]. Donders,G.Van der Heijden, T. Stijnen and K. Moons, "Review: a gentle introduction to imputation of missing values," Journal of Clinical Epidemiology, vol. 59(10),2006, pp. 1087-1091.

[15]. P. García-Laencina,P. Abreu, M. Abreu and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values,"Computers in Biology and Medicine, vol.59, 2015, pp. 125-133.

[16]. P. Schmitt, J. Mandel, andM. Guedj, "A comparison of six methods for missing data imputation," Journal of Biometrics and Biostatistics, vol. 6(1), 2015, pp. 1 - 6.

[17]. R. Johnson and D.Wichern. "Applied Multivariate Statistical Analysis." Prentice Hall, Inc, 1992.

[18]. Dempster,N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B (Methodological), vol. 39(1), 1977, pp. 1-38.

[19]. F. Nelwamondo, D. Golding, andT. Marwala, "A dynamic programming approach to missing data estimation using neural networks,"Information Sciences,vol.237,2013, pp. 49-58.

[20]. E. Hruschka, E. Hruschka and N. Ebecken, "Bayesian networks for imputation in classification problems," Journal of Intelligent Information Systems,vol. 29, 2007, pp. 231-252.

[21]. B. Zhu, C.He, and P. Liatsis, "A robust missing value imputation method for noisy data," Applied Intelligence, vol. 36(1), 2012, pp. 61-74.

[22]. J. Luengo, S.García andF. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," Knowledge and Information Systems,vol.32(1),2012, pp. 77-108.

[23]. C. Caragea, J. Wu, A. Ciobanu, K. Williams, J. Fernandez-Ramrez, H. Chen, Z. Wu, and L. Giles, "CiteSeer x: A scholarly big dataset," Advances in Information Retrieval: Springer International Publishing, 2014, pp. 311-322.

[24]. C. Therrien, "Decisionestimation and classification: an introduction tp pattern recognition and related topics,"John Wiley& Sons,Inc. 1989.

[25]. K. Bache, and M. Lichman,"UCI Machine Learning Repository", [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science,2013.

[26]. W. Aly, O. Hegazy and H. Rashad, "Automated atudent advisory using machine learning," International Journal of Computer Applications (0975 – 8887), vol. 81(19), 2013, pp.19-24.

[27]. US Temperatures Story, http://lib.stat.cmu.edu/DASL/Stories/USTemperatures.html